Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid Graph Learning Reconstructs Global Ocean Oxygen Spatiotemporal Changes

Xinbing Wang

xwang8@sjtu.edu.cn

Shanghai Jiao Tong University

Bin Lu

Shanghai Jiao Tong University https://orcid.org/0000-0001-6452-7029

Luyu Han

University of California San Diego, Shanghai Jiao Tong University

Ze Zhao

Shanghai Jiao Tong University

Haonan Qi

Shanghai Jiao Tong University

Xiaoying Gan

Shanghai Jiao Tong University

Lixin Qu

Shanghai Jiao Tong University

Yuntao Zhou

Shanghai Jiao Tong University

Lei Zhou

Shanghai Jiao Tong University https://orcid.org/0000-0002-0433-3991

Luoyi Fu

Shanghai Jiao Tong University

Meng Jin

Shanghai Jiao Tong University

Chenghu Zhou

Institute of Geographical Science and Natural Resources, Chinese Academy of Sciences

Jing Zhang

East China Normal University

Article

Keywords:

Posted Date: January 14th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-5587966/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Hybrid Graph Learning Reconstructs Global Ocean Oxygen Spatiotemporal Changes

Bin Lu[1], Luyu Han[2,7], Ze Zhao[1], Haonan Qi[1], Xiaoying Gan[1,*], Lixin Qu[2], Yuntao Zhou[2], Lei Zhou[2], Luoyi Fu[3,*], Meng Jin[4], Xinbing Wang[1,*], Chenghu Zhou[5], and Jing Zhang[2,6]

[1]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]School of Oceanography, Shanghai Jiao Tong University, Shanghai, China
[3]Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China
[4]John Hopcroft Center for Computer Science, Shanghai Jiao Tong University, Shanghai, China
[5]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[6]State Key Laboratory of Estuarine and Coastal Research, East China Normal University, Shanghai, China
[7]Department of Computer Science and Engineering, University of California, San Diego, CA, USA
[*]Corresponding authors: Luoyi Fu (yiluofu@sjtu.edu.cn), Xiaoying Gan (ganxiaoying@sjtu.edu.cn), Xinbing Wang (xwang8@sjtu.edu.cn)

ABSTRACT

Climate change and anthropogenic activities have exacerbated hypoxic conditions in the global ocean, threatening marine ecosystems and earth sustainability. Investigating the history of ocean deoxygenation is crucial for developing coping strategies. However, due to the severe sparse in-situ observations and spatiotemporal complexity of deoxygenation mechanism, precise modelling of the oxygen trends, especially in large scale, remains a long-standing problem. Here we propose Jingwei, a hybrid graph learning framework that integrates observation data and ocean climate model in a complementary manner, accurately reconstructing global oxygen levels over the past six decades. A knowledge retrieval module is designed to learn the hidden laws that dominate the complex mechanism of oxygen changes from model-based simulations. Constrained by those laws, a knowledge propagation module spatiotemporally connects the sparse data, capturing dynamic and spatial heterogeneous oxygen loss. With a joint consideration of invariant laws and variant factors, Jingwei significantly reduces reconstruction error, achieving equal accuracy with the dense cruise surveys. It can quantitatively trace the expansion process of oxygen minimum zones (OMZ), revealing potential drivers, effectively supporting on-target strategies. We foresee that the implications of such a graph-based hybrid framework extend beyond ocean deoxygenation and pave way for next-generation earth system modelling.

Introduction

Oxygen is essential for all life, particularly in marine ecosystems, where it exists as dissolved oxygen (DO). Fluctuations in DO concentrations are vital for fisheries and the whole marine ecosystem. Unfortunately, increasing research indicates that the oceans have been steadily experiencing deoxygenation over the past few decades[1–4]. The rapid decline in oxygen levels, compounded by global warming and ocean acidification, poses serious threats to marine environments, leading to unpredictable and potentially irreversible ecological and economic consequences[5,6]. This trend also disrupts global biogeochemical cycles, impairing the ocean's ability to sequester carbon and regulate climate[7]. Moreover, regional variations due to climate change and human activities further complicate the spatial and temporal patterns of ocean deoxygenation[8–10]. Therefore, a comprehensive understanding of global ocean oxygen changes is urgently required to inform effective ecological protection strategies and government policies.

However, historical DO observations are severely sparse. According to statistics, World Ocean Database (WOD)[11], the largest ocean observation database, only accounts for 5.499% of global DO distribution since 1960 after grid conversion[12,13]. Existing approaches to this problem can be broadly categorized into two groups:

(1) *Statistical interpolation*, which reconstructs data for the unobserved time/spatial points with observations in a sliding temporal window/spatial region. For example, World Ocean Atlas (WOA)[14,15] provides 1991-2020 climatological average of DO by interpolating oceanic observations over a 30-year time window. Some other methods[1,16] reconstruct global oxygen data by performing 2D spatial interpolation. They stratify vertical depth and perform spatial interpolation on annual data in each horizontal layer with a large aggregation grid. However, one problem with all the above methods is that, without considering underlying deoxygenation mechanism, they underestimate the variability of oxygen concentrations with region, depth, and time, resulting in overly smooth and suboptimal reconstructions, particularly in data-sparse areas. As a result, statistical interpolation methods achieve high accuracy only in specific regions and timeframes with relatively dense observation data[17–20], which

hinders large-scale analysis of oxygen changes.

(2) *Numerical simulations*. An alternative is to simulate oxygen data through ocean system modelling, where all aspects of the ocean system (including the physical, chemical and biological processes) are jointly considered to provide a global and long-term distributions of ocean oxygen. For example, Coupled Model Intercomparison Project Phase 6 (CMIP6)[21] includes several DO simulation results, which has supported the analysis of deoxygenation trends in many studies[22–25]. However, such a complex modelling system, although it shows high descriptive ability, are vulnerable to biased modelling assumptions and imperfect physical and biogeochemical parameterizations. Even small errors can be continuously accumulated and amplified in the complex and iterative calculations of climate models due to the lack of observation correction. As prior studies[26–28] have shown that there exists substantial discrepancies between simulated results and historical observations, leading to conflicting understandings of oxygen trends.

The sparse in-situ observations and the complex deoxygenation mechanisms pose a conflict between the accuracy and the scale of DO reconstruction. A promising approach to break this conflict lies in the integration of observation data and model based simulations, referred to as hybrid ocean climate modeling[29,30]. This innovative strategy seeks to harness the capabilities of deep learning to identify nonlinear correlations within observational data while maintaining consistency with oceanographic principles in climate modeling. However, effectively combining these methods within a cohesive computational framework for such a complex and large-scale spatiotemporal process poses significant challenges, remaining at the conceptual stage.

Here, we propose a novel hybrid graph learning framework, named Jingwei (See Methods for an explanation of the name 'Jingwei'), that integrates the strengths of theory-simulation-based and data-driven techniques in a complementary manner. Jingwei utilizes *graph-based modeling* to represent the reconstruction process, where the vertical profiles of DO are denoted as nodes, and the relationships between DO profiles across different spatial locations are denoted as edges. Deep learning on graphs, i.e., graph neural networks[31–33], can not only learn the attributes of individual nodes, but also propagate the attributes by perceiving hidden correlations across edges. Our framework is designed with two core modules: *knowledge retrieval* and *knowledge propagation*. First, despite the complexities inherent in ocean climate models, we show that there exists governing patterns in model-based simulations describing DO variation behaviors. The knowledge retrieval module learns and encodes these patterns, discovering the invariant hidden laws for different node. Second, informed by these governing patterns, the knowledge propagation module spatiotemporally connects the sparse data through edges, capturing the dynamic and spatial heterogeneous distribution of oxygen levels.

Trained on mulit-source observation data (more than 566 million DO records) and numerical simulations from CMIP6, Jingwei reconstructs the global ocean oxygen level from 1960 to 2023 with an annual resolution. It covers the global ocean with $1^\circ \times 1^\circ$ (latitude $\times$ longitude) grids, and extends from the ocean surface to 5500-meter depth (33 depth levels). Our multifaceted evaluations demonstrate that the hybrid graph learning model Jingwei significantly enhances the DO reconstruction performance, with particularly notable enhancements in intermediate depths and the deep ocean. We demonstrate that Jingwei effectively identifies geographic variability and adaptively partitions ocean zones purely from data, accounting for factors such as land boundaries, ocean currents and gyres, resulting in interpretable reconstructions. Furthermore, Jingwei facilitates a fine-grained understanding of the drivers of ocean deoxygenation, involving tracking both horizontal and vertical expansion processes of oxygen minimum zone (OMZ) in different regions.

Results

Overview of Jingwei

Jingwei presents a hybrid graph learning model to unify ocean climate models and in-situ observations for historical DO level reconstruction (Fig 1a). On one side, ocean climate models encompass simulations of complicated physical, chemical, and biological processes that collectively reflect the ocean deoxygenation mechanisms. Jingwei learns and encodes the governing patterns via simulation-based hidden law discovery, hereby predicting unobserved profiles. On the other side, discrete in-situ observations reveal actual spatiotemporal variations of dissolved oxygen. Jingwei establishes abundant connections to perceive the dynamics and spatially heterogeneous oxygen changes on graph topology, further improving reconstruction results under hidden law constraints.

Jingwei initializes a graph model to flexibly depict the correlations between profiles both locally and remotely. The gridded oceanic observations consist of spatiotemporal data with 4D coordinates (latitude, longitude, depth and time), while constrained by land plates and seabed topography, resulting in an irregular data cube. Jingwei models the ocean oxygen reconstruction with graphs $\mathbf{G}(\mathbf{V},\mathbf{E})$, where each node $v \in \mathbf{V}$ denotes a DO vertical profile and the relationships between profiles form the edges $e \in \mathbf{E}$. To expand the receptive field of nodes, we define two relationships of nodes including proximity neighbors and information hubs, hereby capturing both local and long-range correlations. Proximity neighbors perceive local changes in accordance with the First Law of Geography[34]. Information hubs associate remote but comprehensive observations, while considering the rationality of geographic, such as land division and enclosed seas.

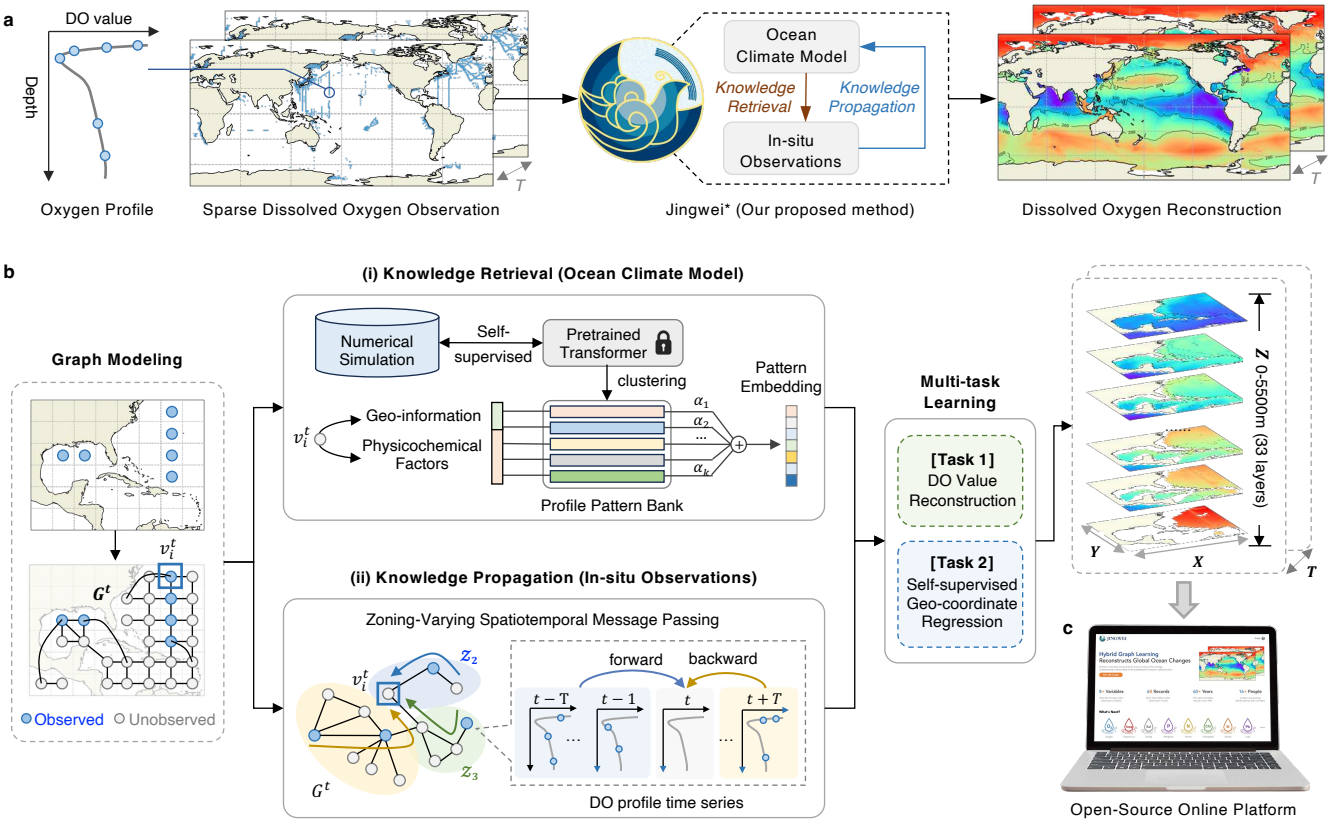

Figure 1. Overview of global ocean oxygen trend reconstruction framework named Jingwei. a, The workflow of Jingwei. We aim to reconstruct sparse dissolved oxygen observation data into a complete 4-dimensional spatiotemporal distribution of ocean dissolved oxygen in a complementary manner. An oxygen profile refers to the curve of dissolved oxygen with depth at a given position. **b**, The detailed architecture design of Jingwei, including graph modelling, knowledge retrieval, knowledge propagation and multi-task learning. Jingwei learns the hidden laws in pattern bank from numerical simulations, while captures inter-profile correlation with zoning-varying graph message-passing among different observations. Dissolved oxygen reconstruction and self-supervised geo-coordinate regression are utilized as two training tasks to jointly guide the model training, and ultimately obtain DO reconstruction results across the four dimensions of longitude, latitude, depth, and time. More details refer to Figure S2. **c**, Alongside, we have released an open-source online platform (https://jingwei.acemap.info/), providing data visualization, resource sharing and ongoing updates to more variables.

Jingwei discovers the hidden laws from model-based simulations for knowledgeable node (DO profile) representation through the *knowledge retrieval* module. Instead of directly employing doubtful simulations, Jingwei extracts effective patterns from numerical simulations, hereby removing the outliers and forming a pattern bank. Each element in the pattern bank is a high-dimensional vector representing the hidden law of oxygen changes from ocean climate model. Then, Jingwei adopts environmental variables to match different nodes with their corresponding dominant patterns. Environmental variables refer to more easily accessible geographic information and physicochemical factors (including longitude, latitude, temperature, salinity, average depth measurement, etc), utilized as queries for knowledge retrieval. Finally, a weighted summation is performed to obtain knowledgeable representations for each node with matching scores.

Jingwei facilitates spatiotemporal correlations among different profiles through the *knowledge propagation* module. The DO characteristics of each node can propagate to each other based on the graph structure, achieving mutual verification and correction. However, the global ocean shows heterogeneous spatiotemporal correlations due to the varying regional impacts of global warming and human activities, making it challenging for a unified model to capture the dynamic changes across various areas. Inspired by the zone partitioning in oceanography[35–37], we propose to carry out differentiated propagation in different zones. Given the unknown zoning results for DO, Jingwei conducts adaptive zone partitioning and non-shared propagation parameter tuning in an end-to-end learning strategy, named zoning-varying spatiotemporal message passing. Among them, the propagation message is the temporal variations of both forward and backward DO profile time series. The propagation parameters are adaptively generated with hypernetwork[38,39], reflecting the differences in spatial zones.

The training process of Jingwei includes two tasks: one is the DO value reconstruction, and the other is the self-supervised geo-coordinate regression. The latter serves as an auxiliary task to enhance supervision for missing values, thereby improving the Jingwei's expressiveness and stability. Jingwei is trained on multi-source gridded observational data (Figure S1, Table S1) after unified data quality control (Table S2), including a total of 566,375,179 records of dissolved oxygen observation data (235,111 grids after 4D gridding). The predictor variables used to train Jingwei is shown in Table S4. After training, we obtain a 4D gridded data product inferenced by Jingwei, covering $1^\circ \times 1^\circ$ global ocean oxygen levels from 1960 to 2023, with depth from 0-5500 meters.

Performance Evaluation of Global Ocean Oxygen Reconstruction

To evaluate the performance of Jingwei and other methods, we reserve a portion of gridded observations ($N_{\text{test}} = 47,153$) for testing. To ensure the heterogeneity of the test dataset, testing samples are selected randomly for each year.

First, we compare Jingwei with statistical interpolation as illustrated in Figure 2a. We reveal that over 60.35% of unobserved nodes are located more than 400 km from the nearest observation, with 15.45% even exceeding 1300 km, highlighting the extreme sparsity of available observations. The root mean square error (RMSE) variation of reconstructions indicates that Jingwei maintains a relatively stable reconstruction performance, while interpolation methods heavily rely on the richness of nearby information. In particular, Jingwei exhibits a notable advantage over interpolation methods for distances beyond 400 km. Gradually deteriorating interpolation methods are not suitable for analysis over large spatiotemporal scales, especially for areas that are poorly understood due to a lack of observations.

Next, we compare Jingwei with three global model-based numerical simulations (CESM2-omip1[12], CESM2-omip2[12], and GFDL-ESM4[13]) from CMIP6, using the same observational test set. These simulations adhere to the same spatiotemporal resolution as Jingwei, ensuring direct and fair comparisons. Figure 2b displays the spatial distribution of RMSE across different methods. We observe that Jingwei significantly improves reconstruction error compared to the three numerical simulation methods, both in coastal and open ocean areas. The average RMSE values for CESM2-omip1, CESM2-omip2, GFDL-ESM4, and Jingwei are 37.47, 38.35, 36.15, and 27.13, respectively. Jingwei outperforms all numerical simulation methods with an average improvement of 27.28%, demonstrating its powerful capability.

Third, we further compare Jingwei with frequently used artificial intelligence (AI) methods. However, we observe that many AI methods (LSTM, XGBoost, MLP, etc.) fall into the overfitting-underfitting dilemma[40], with a large number of glitches in the reconstruction results (Fig 2c). This phenomenon is defined as image artifacts[41,42], which are undesired characteristics that degrade the visual quality of the reconstruction. Image artifacts are especially prevalent in less observed areas, where classical AI methods struggle to obtain sufficient supervised signals because of the scarcity of DO observations. Consequently, these limitations prevent reliable application for further scientific analysis. To provide a quantitative assessment, we adopt Sobel operator[43], where a higher Sobel value indicates more image artifacts. The Sobel value of the XGBoost method is the highest at 81.69, while the values for LSTM and MLP are 43.27 and 40.67 respectively. The Sobel value of Jingwei is the lowest at 20.91, demonstrating our superior generalization performance.

Hence, a more comprehensive performance evaluation should integrate both the reconstruction error on observational test set, and the generalization performance in unobserved areas. For the reconstruction error, we consider widely used metrics such as Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and R-squared (R^2). Although generalization performance is not directly measurable, we assess it indirectly through the visual quality of reconstructions. We apply the

Sobel[43] operator alongside deep learning-based visual quality assessment models, specifically MUSIQ[44] and hyperIQA[45]. Figure 2d illustrates the performance of different baselines in two dimensions. Jingwei, marked by a red square, is positioned in the lower-left quadrant, indicating an effective balance between reconstruction accuracy and visual quality. In contrast, numerical simulation methods (marked by blue circles) cluster in the lower-right quadrant, displaying good visual quality but significant reconstruction errors compared to actual observation data. Most AI methods (marked by yellow triangles) fall within the upper quadrant; while some exhibit statistically better error performance than Jingwei, their visual quality is inferior, indicating that their generalization performance is inadequate for scientific applications. Additionally, Figure 2e presents a heatmap of various metrics, with red indicating rankings in the top four. Our Jingwei model consistently ranks highly across all evaluated metrics, demonstrating its overall effectiveness.

Regional Comparison with Dense Cruise Surveys

To gain a comprehensive understanding of the hydrological characteristics in specific regions, marine scientists conduct dense observations through ocean cruise surveys. To evaluate the effectiveness of Jingwei, we utilize multiple sets of cruise data with dissolved oxygen (DO) observations from the GEOTRACES project (https://www.geotraces.org/) as comparative test sets. Figure 3a illustrates the cruise track of the joint Australia-New Zealand GP13 expedition. This survey involved 61 stations along a transect at approximately 30°S, extending east of Australia into the South Pacific and concluding at 32.5°S, 150°W (south of Tahiti). Figure 3b displays the dissolved oxygen concentrations measured by the CTD sensor, along with the interpolation results generated by Ocean Data View (ODV)[46]. We consider these measurements as the benchmark for DO distribution in this region, which we then compare with the results obtained from Jingwei and three numerical models (Figure 3c, d). Our analysis reveals that Jingwei outperforms other ocean climate models, which have been widely noted for their poor reconstruction results of intermediate and deep water masses (as seen with CESM2-omip1 and CESM2-omip2)[26,47]. Jingwei significantly enhances the accuracy of these crucial water mass estimations. Furthermore, we assess the dissolved oxygen results from additional cruises, including GP15, GApr08, and GPc06 (Figure S4), and consistently find that Jingwei delivers the most accurate reconstructions. The robust performance of Jingwei underscores its potential to enhance future cruise planning and target area selection.

Interpretability of Jingwei's Reconstruction Process

What knowledge does Jingwei retrieve and propagate from numerical simulations and observations? Can we interpret the reconstruction process of Jingwei? To solve these problems, we first investigate the knowledge retrieval module by "opening" the pattern bank to determine the correlations between learned hidden laws and different environmental factors (Figure 4). For knowledge propagation module, we focus on whether Jingwei detect heterogeneous spatial correlations reasonably through zoning-varying spatiotemporal message passing.

In the knowledge retrieval module, we form a pattern bank composed of 264 patterns learned from ocean climate models. Utilizing environmental variables (including longitude, latitude, temperature, salinity, etc) of different nodes as queries, we can identify different relevant patterns through vector matching. Figure 4 displays the correlation analysis between these environmental variables and the selection of patterns. From this analysis, we can draw several key conclusions: (1) There exists some dominant patterns that are repeatedly selected, indicating that there are indeed hidden laws that reflect trends in DO fluctuations within model-based numerical simulations. Numerical variations of environmental variables are sensitive to pattern selection and exhibit selectivity. (2) There are notable differences in pattern selection between the hemispheres (Figure 4b). Specifically, latitude appears to exert a greater influence on pattern selection than longitude. Changing longitude values at the same latitude generally does not significantly affect pattern selections. This finding aligns with the common observation that climate-related variations along the north-south axis are more pronounced than those along the east-west axis at equivalent distances. (3) Pattern selection is particularly responsive to temperature, indicating that different patterns reflect varying hypoxia conditions and demonstrate strong discriminatory power (Figure 4c). The average salinity of seawater is around 35 practical salinity units (psu), and pattern selection remains relatively stable during small-scale fluctuations. However, significant changes in pattern selection occur when salinity levels are notably low, highlighting the model's ability to detect environmental anomalies (Figure 4d).

In the knowledge propagation module, we are pleased to observe that Jingwei can effectively recognize spatial variability, such as ocean currents, gyres and the differences between coastal and open ocean environments, without explicit information. This spatial awareness allows for adaptive feature extraction tailored to different zones. To clarify and interpret the spatial heterogeneity identified by Jingwei, we employ t-SNE[48] to reduce the learned propagation parameters to three dimensions, normalizing them into RGB color combinations. Jingwei utilizes two layers of variable message passing, resulting in two sets of dimensionality-reduced parameters, as depicted in the second and third columns of Figure 5. Each color represents a distinct zone. Figure 5a illustrates the junction of the South Indian Ocean and South Atlantic Ocean, where multiple cold and warm currents converge. Our result indicate that coastal areas share similarities in message propagation, while ocean regions further from land exhibit significant diversity along the east, west, and south sides of Africa. Areas impacted by the West Wind Drift

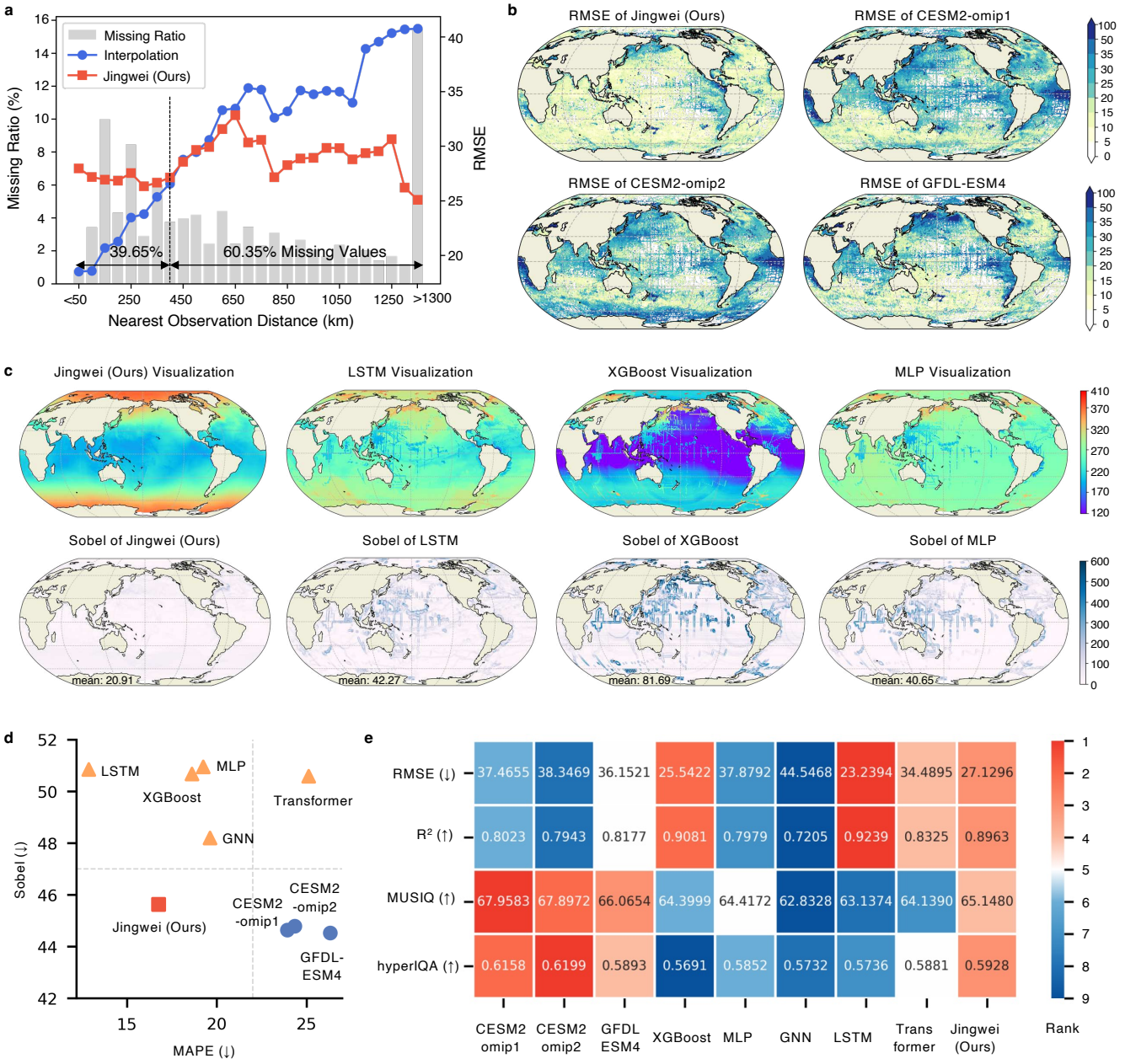

Figure 2. Performance Evaluation of Global Ocean Oxygen Reconstruction. **a**, Performance variation of statistical interpolation and Jingwei with regard to the distance between prediction samples and the nearest observed samples. **b**, Performance comparison of Jingwei and other three numerical simulation methods in spatial distribution. **c**, Visualization cases and corresponding Sobel values of different methods at the same time and depth layer. Directly using machine learning models for reconstruction lead to the overfitting-underfitting dilemma, i.e. overfitting to sparse observed samples but lacking extrapolation ability to unobserved regions. **d**, The two-dimensional joint evaluation of reconstruction error and visual quality. The closer different methods are to the origin of the coordinate quadrant, the better their overall performance. **e**, Further comparison of different methods in terms of reconstruction error (RMSE, R^2) and visual quality (MUSIQ, hyperIQA). Jingwei ranks among the top and demonstrates its superior performance when jointly considering all indicators. The ablation study of each module and the uncertainty of the model please refer to Figure S3 and S5.

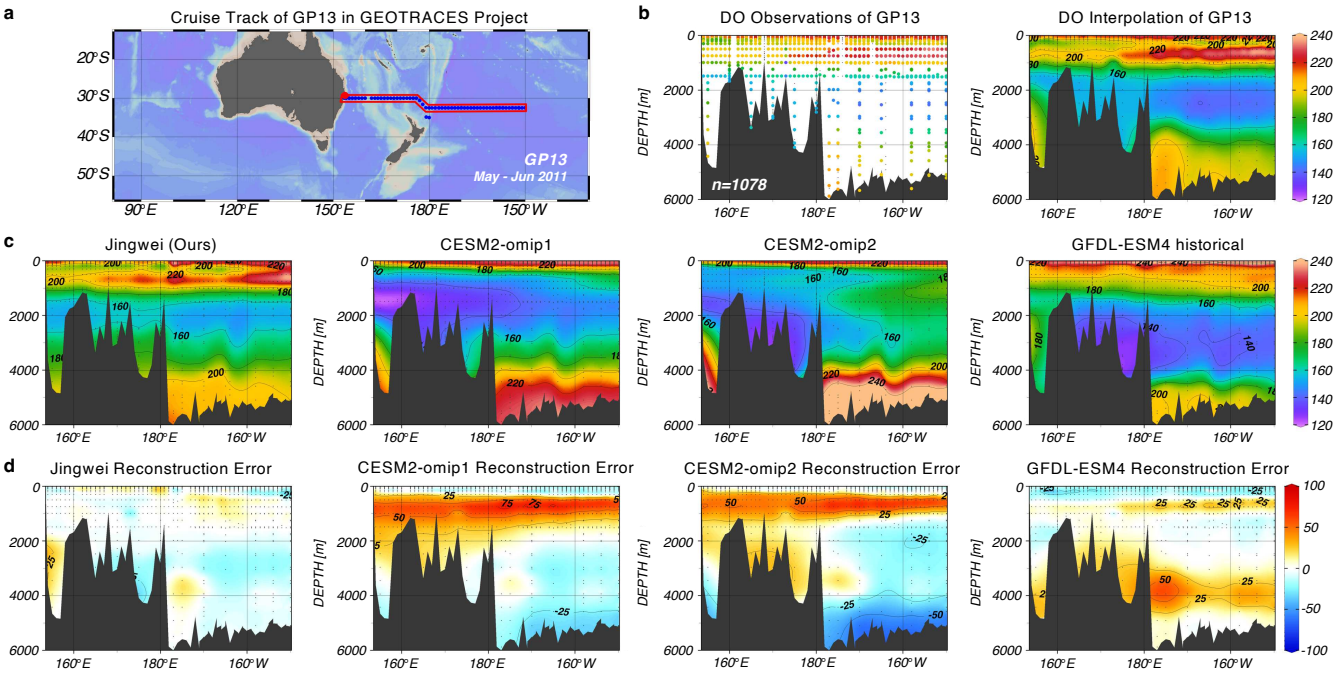

Figure 3. Observations and reconstruction comparison of dissolved oxygen in GP13 cruise. a, GP13 cruise map information. **b**, Dense observation and ground truth interpolation of GP13. **c**, Reconstruction result of Jingwei and three numerical simulations without any observations. **d**, Reconstruction error compare with ground truth interpolation. Jingwei shows significant improvement compared to numerical simulation methods in the subsurface and deep seawater. More comparison of other cruise observations excluded in training can be found in Figures S4. Moreover, we evaluate the reconstruction results of Atlantic Meridional Overturning Circulation (AMOC) in Figure S6.

display consistent features. Figure 5b focuses on the North Indian Ocean. Jingwei significantly differentiates between the Arabian Sea and the Bay of Bengal, highlighting the continuity of coastal features and differences between monsoon currents and the Somali cold current. Meanwhile, Figure 5c addresses the Western Equatorial Pacific, a region featured by numerous islands, complex seabed topography, and intensive human activities. Here, the propagation parameters demonstrate more complex and variable characteristics, underscoring Jingwei's capability to adapt to intricate environments. Finally, Figure 5d presents the Arctic region. Despite limited training data, Jingwei successfully identifies changes in the Beaufort Gyre and the Transpolar Drift. Overall, these examples illustrate that, unlike traditional spatial interpolation methods focused solely on smoothing based on distance, Jingwei can discern the ocean's inherent patterns from data, applying this insight to reconstruct dissolved oxygen levels. Additional cases are discussed in Figure S7.

Ocean Oxygen Spatiotemporal Changes Analysis based on Jingwei

How can comprehensive, accurate, and large-scale spatiotemporal reconstruction data enhance our understanding of ocean deoxygenation? We first provide a brief overview of global and representative regional oxygen levels and their changes, outlining the overall ocean deoxygenation trends over the past 60 years. We then focus on the vertical expansion patterns of global hypoxic regions, with particular emphasis on oxygen minimum zones. This analysis is made possible by the unique design in processing profile data in our study, which distinguishes our findings from previous research.

Utilizing the 4D gridded data products reconstructed with Jingwei, we can quantitatively analyze trends of ocean dissolved oxygen levels. Our findings estimate the average global ocean oxygen content since 1960 at 232.13 ± 0.20 Pmol (where 1 Pmol = 10^{15} mol), aligning closely with the estimate of 227.38 ± 1.10 Pmol reported in the earlier study[1]. Notably, we observe regional variations of oxygen content. In equatorial regions—including the Equatorial Atlantic, Equatorial Pacific, and Equatorial Indian Ocean—Jingwei suggests that prior assessments of total oxygen content may be slightly overestimated. Conversely, oxygen levels in the Arctic Ocean, North Pacific, and Southern Indian Ocean appear to be underestimated (see Table S6 for details). Regarding the temporal trend, the global oxygen in seawater exhibited a slight increase until the mid-1980s, followed by a marked decline thereafter. This turning trend, first observed in the upper ocean from 0 to 1000 meters globally[49,50], is now confirmed across full depths. Between 1960 and 2023, the linear trend for global ocean O_2 inventory is -1098.48 ± 144.49 Tmol per decade (1 Tmol = 10^{12} mol), which is comparable to the -960.4 ± 429.1 Tmol per decade

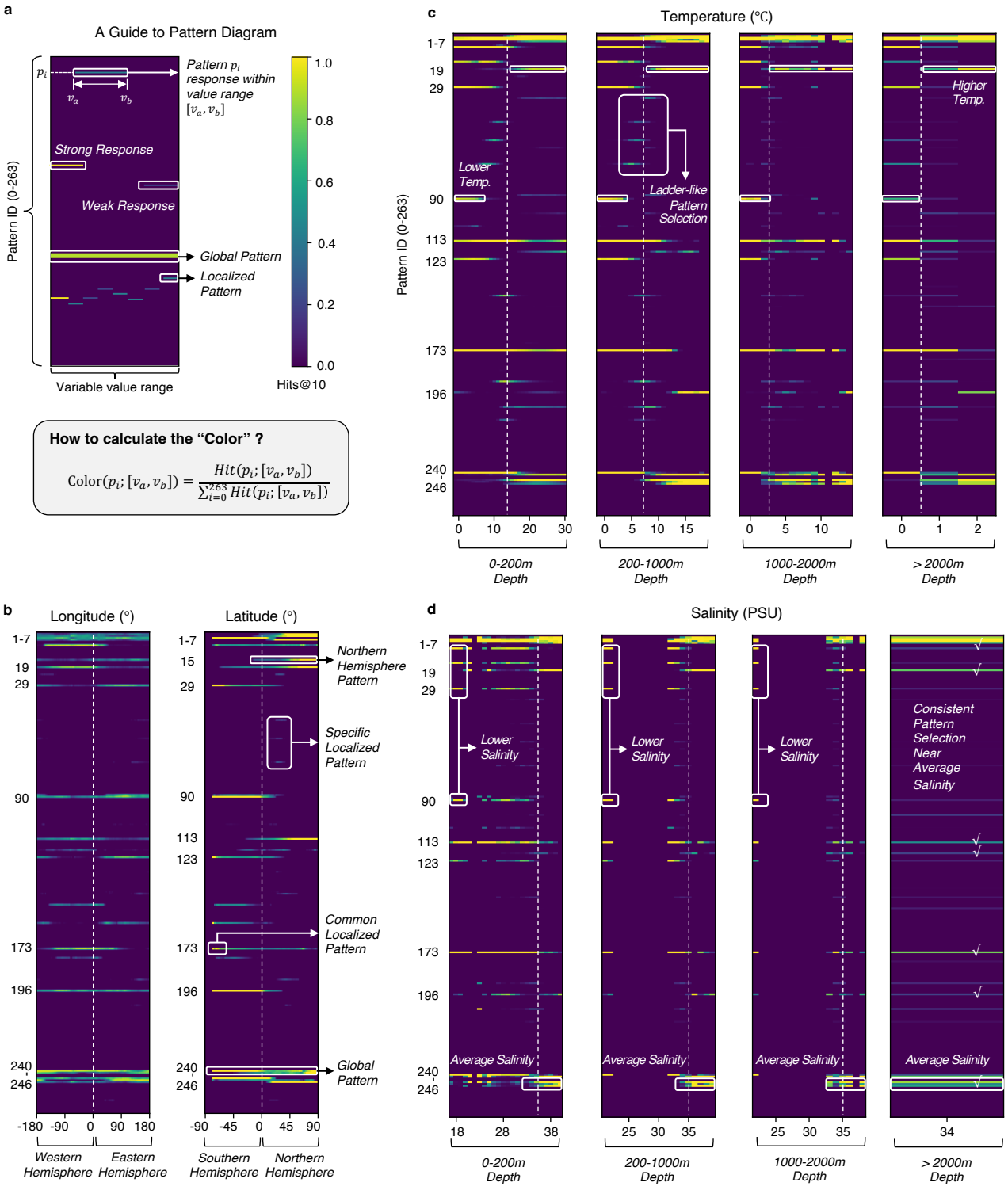

Figure 4. Correlation analysis of different high-dimensional patterns in pattern bank in response to various environmental variables. a, a guide to pattern diagram. The shared color bar on the right represents the normalized probability of belonging to the Top 10 hit patterns for different variable values. Colors close to yellow indicate a higher probability of the variable value hitting the pattern, indicating a common pattern selection for this value. Darker colors indicate specific pattern selections. If a pattern ID covers most of the numerical range, it is a global pattern; otherwise, it is a localized pattern. **b**, The relationship between different patterns with respect to longitude and latitude. **c, d**, The relationship between different patterns with respect to temperature and salinity at four depth levels.

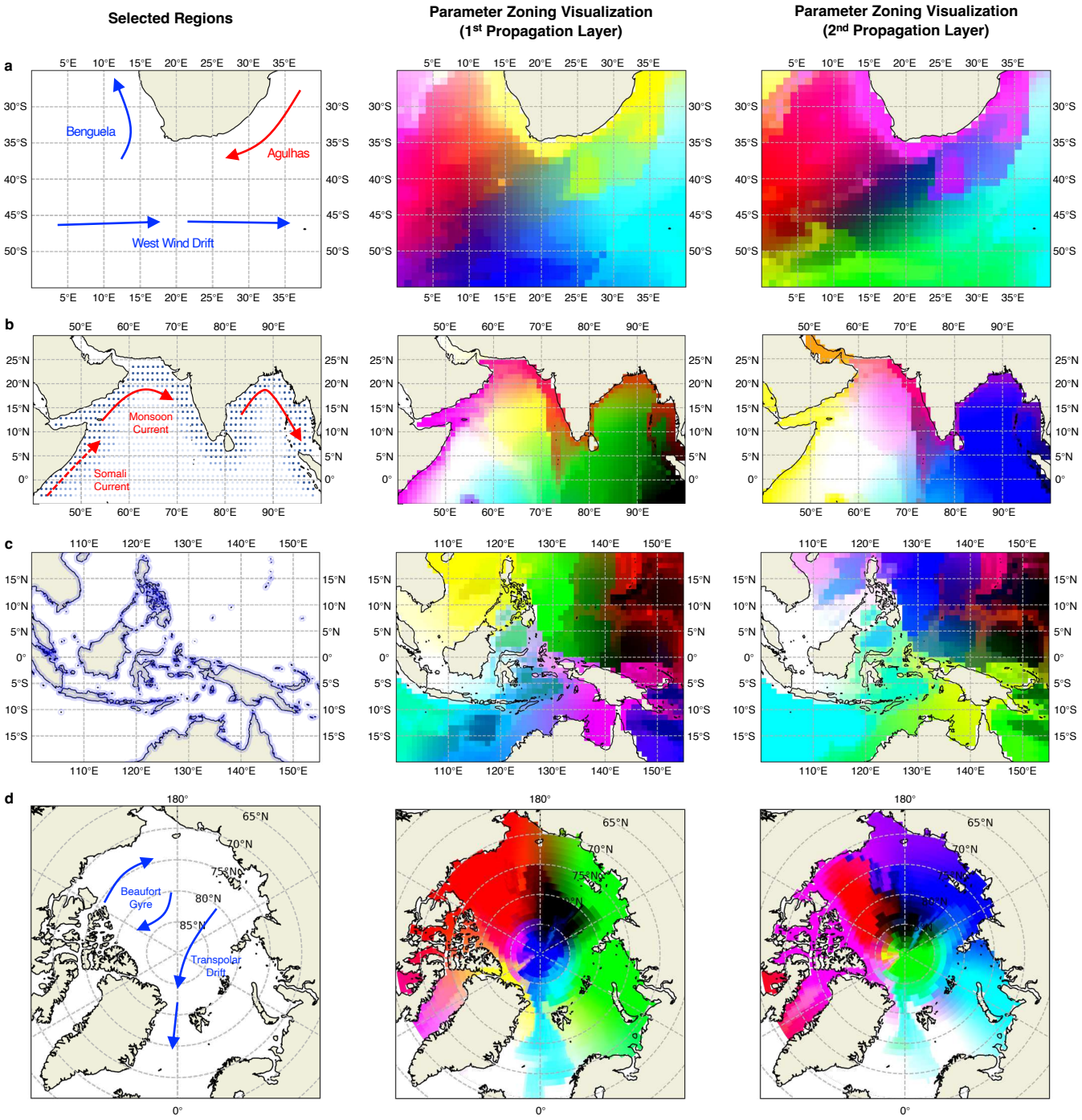

Figure 5. Visual analysis of parameter zoning in zoning-varying spatiotemporal message passing. In the three columns of images, the first column displays the location of the selected region, while the second and third columns visualize the two layer propagation parameters reduced to three-dimensional RGB parameters by TSNE. Different colors represent different zones. Jingwei automatically divides the global oceans into distinct zoning. Intriguingly, the zoning results reflect oceanic spatial variations such as currents, gyres, and distinguish the characteristics of land, coastal areas and the open sea. **a**, South Indian Ocean and South Atlantic Ocean Junction, south of Africa. **b**, North Indian Ocean region. **c**, Western Equatorial Pacific region. **d**, Arctic region. More cases are available in Figure S7.

previously reported[1]. Overall, Jingwei confirms consistent understanding of global mean and temporal trend of total ocean oxygen content, and further extends these insights to a broader spatiotemporal scale.

Oxygen minimum zones (OMZs) are regions in the ocean characterized by significantly low dissolved oxygen concentrations (Figure S8), which are ideal choices for concretely describing deoxygenation spatial variation patterns. As global oxygen levels decline, OMZs are expanding continuously[16,26], posing serious threats to marine habitats[51,52]. Existing numerical simulations show limitations in accurately modelling dissolved oxygen at intermediate depths, leading to biased quantitative analysis of OMZs. To better understand the horizontal and vertical expansion of OMZs, we firstly categorize their evolution into several types. These categories include three types of vertical expansion: `Bidirectional Expansion`, `Upward Expansion`, and `Downward Expansion`; along with one horizontal expansion referred to as `Emerging OMZ`; relatively stable regions known as `Stable OMZ`; and contracting regions termed `Shrinking OMZ`. In this paper, we specifically focus on the less understood vertical expansion patterns of OMZs. In particular, the bidirectional vertical expansion of OMZs is influenced by different mechanisms. Climate change-induced intensified stratification and thermocline ascension affect the exchange between surface oxygen-rich waters and subsurface waters, leading to the upward expansion of OMZ boundaries. Conversely, the downward changes are primarily linked to oxygen consumption during organic matter decomposition and deep ocean circulation processes.

Our analysis concentrates on two types of key areas: OMZ_{20} ($O_2 \leq 20\,\mu mol/kg$) representing the core of OMZs, and OMZ_{60} ($O_2 \leq 60\,\mu mol/kg$) representing hypoxic waters. As illustrated in Figure 6a, OMZ_{20} displays a north-south asymmetric distribution, notably concentrated in the North Indian Ocean, North Pacific, and Equatorial Pacific. Since 1960, the core OMZ area has expanded by 12.77 million km², roughly 1.3 times the size of the United States, with significant horizontal expansion both toward the coast and into the open ocean, particularly in the North Pacific. Among the five types of vertical evolution, bidirectional expansion is the most prevalent, accounting for 59.58%, followed by upward expansion at 24.98%, downward expansion at 7.78%, stable regions at 6.78%, and shrinking regions at 0.87%. As for OMZ_{60} areas, the hypoxia water in Indian Ocean is expanding southward, while expanding westward in Pacific Ocean and Atlantic. Bidirectional expansion (51.63%) and upward extension (21.17%) are dominant for OMZ_{60}, which has led to increased mortality and migration of marine life.

Figure 6b presents a OMZ_{20} scatter plot depicting the expansion rates of both the upper and lower boundaries across different regions (detailed values are shown in Table S7). The second quadrant (upper left area) illustrates the bidirectional expansion of OMZs. The dashed line indicates equal velocity expansion of both boundaries. The highest rates of bidirectional expansion are found in the North Indian Ocean and its coastal waters. In this region, the upper boundary rises by an average of 9.31 meters per decade, while the lower boundary descends by 16.58 meters per decade, resulting in an average OMZ core thickness of 680.93 meters. In the Coastal North Indian Ocean, the variations closely mirror those of offshore areas, making the expansion rate particularly significant due to the limited depth of coastal waters. In the South Pacific, including coastal regions, OMZ core expansion primarily depends on the descent of the lower boundary. Notably, the OMZ core in the Black Sea remains stable, while the Equatorial Pacific and Coastal North Pacific show similar expansion rates for both boundaries. Additionally, the scatter plot of OMZ_{60} is shown in Figure S10. We also select several representative regions to examine specific temporal changes in depth, as depicted in Figure 6c and 6d. In the Northern Indian Ocean and Equatorial Pacific, the OMZ core area shows a bimodal increase over time, approaching both upper and lower boundaries. In the Equatorial Pacific, OMZ_{20} area variation peaks occur at depths of 200 meters and 700 meters, whereas in the Coastal North Indian Ocean, they are found at 150 meters and 1100 meters. Additionally, in the South Pacific and Coastal Equatorial Pacific, the OMZ area decreases near the upper boundary, while expansion is noted in deeper regions.

Overall, various patterns of OMZ expansion exist across different areas, characterized by changes in the vertical depths of upper and lower boundaries as well as the horizontal area at various depths. Even within the same pattern of expansion, significant differences can arise in the specific depth ranges of rapid changes. Therefore, comprehensive and accurate data on dissolved oxygen are essential for identifying endangered species and regions experiencing severe deoxygenation, thereby informing targeted scientific monitoring and policy efforts for protection.

Discussion

Jingwei presents a pioneering effort to propose hybrid graph learning for reconstructing global ocean oxygen trends over the past 60 years by integrating observational data with numerical simulations. This approach not only enhances DO reconstruction accuracy for oceanography but also provides a valuable direction for deep learning study. Given our distinct methodology, Jingwei's result can serve as an independent and complementary resource to model-based simulations. The findings from Jingwei enable marine scientists to gain a clearer understanding of the spatiotemporal changes in dissolved oxygen levels and to investigate deoxygenation mechanisms, such as the vertical expansion of oxygen minimum zones (OMZs) in various regions. Furthermore, Jingwei illustrates the versatility of graph learning for large-scale spatiotemporal analyses, suggesting its applicability to other areas of ocean system modelling, including ocean biogeochemical cycles modelling and full-depth heat content reconstruction. As a result, we have also developed an open-source online platform (https://jingwei.acemap.info/) for

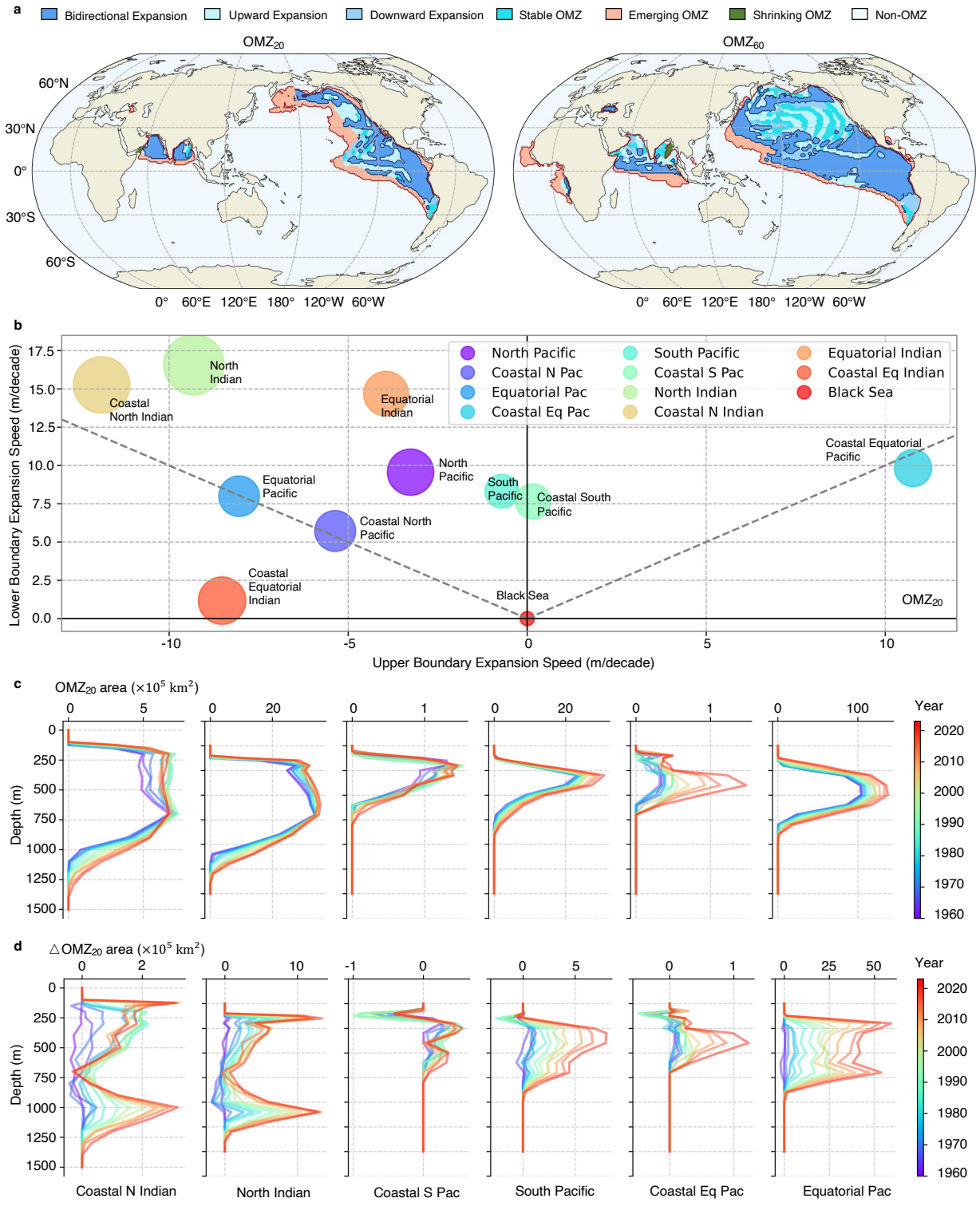

Figure 6. Analysis of global ocean oxygen spatiotemporal changes. a, Cartography showing the evolution of ocean oxygen minimum zones (OMZs) categorizes regions into six distinct types, with OMZ_{20} on the left and OMZ_{60} on the right. **b**, Vertical expansion of the OMZ_{20} core area is illustrated, with two axes indicating expansion rates of the upper and lower boundaries, respectively. The size of each point reflects the average thickness of the OMZ; a negative expansion rate indicates upward movement, while a positive rate indicates downward movement. **c**, The vertical distribution of the average OMZ_{20} area every 5 years in selected regions is shown, with colors ranging from cool to warm tones to represent time progression—from blue (1960) to red (2023). **d**, Average changes in the vertical distribution of the OMZ_{20} area for every 5 years in selected regions are compared to the area in 1960. The boundaries of each divisions are shown in Figure S9.

data visualization, resource sharing, and continuous updates.

Despite these advancements, there are several limitations to address. Jingwei currently relies on limited oceanic observational variables—specifically dissolved oxygen, temperature, and salinity. Other important variables, such as nitrates, phosphates, and chlorophyll, are not included due to their data scarcity. Future research could enhance Jingwei's insights by integrating these observational data or other simulation outputs, such as ocean current. Additionally, the adaptive zone partitioning in the knowledge propagation module currently operates across full depths without accounting for variations among different layers. Enhancing this feature to recognize spatial heterogeneity at varying depths could provide deeper insights into the dynamics of surface, thermocline, and deep ocean layers. Finally, while this study explores the applications of Jingwei in analyzing OMZ evolution, there are numerous avenues for further investigation by ocean scientists.

Looking ahead, hybrid models that combine observational data with numerical simulations may shape the next generation of Earth system models. Current advanced data assimilation practices mainly rely on numerical simulations, supplemented by observations as proxy data for model calibration. In contrast, Jingwei emphasizes observational data, utilizing simulations for knowledge injection. Achieving an effective balance between these two data types poses significant challenges and merits further exploration. We believe that such integration will not only enhance reconstruction performance but also facilitate new discoveries in oceanography, meteorology, and related Earth sciences.

Methods

Data Sources and Processing

Observation Data. In our work, we collect comprehensive dissolved oxygen observation data from five publicly available databases (Table S1). Additionally, we include two fundamental variables, temperature and salinity, in our dataset, with all observational data containing specific information on longitude, latitude, depth, and time. All the data are quality controlled. To ensure the validity of the data, we establish a unified quality control mapping rule. We map the quality FLAG settings of different databases to a common standard and retain only the "good quality" data for subsequent training and testing (Table S2). After quality control, there are a total of 566,375,179 records of dissolved oxygen (Figure S1), 3,200,815,824 records of temperature, and 2,263,189,180 records of salinity. We assume no systematic error in the data over the last six decades, according to previous reference[1]. We follow the previous method[1,15] to grid the observational data, where the temporal resolution is annual, the spatial resolution is $1^\circ \times 1^\circ$ (latitude $\times$ longitude). The depth level is divided into 33 layers ranging from 0 to 5500 meters (Table S5). To avoid biases by clusters of data, all profiles within 0.25° in latitude and longitude and 3 months are median-binned. We also use bedrock elevation data[53] to represent the changes in seabed depth range in different areas.

Model-based Simulations. To compare the reconstruction performance of Jingwei, we select three numerical simulations which involve historical simulations of ocean dissolved oxygen distribution including CESM2-omip1[12], CESM2-omip2[12] and GFDL-ESM4[13] in Coupled Model Inter-comparison Project Phase 6 (CMIP6), accessed from https://esgf-node.llnl.gov/search/cmip6/ (Table S3). Particularly, these three simulation results are characterized by a spatial resolution of $1^\circ \times 1^\circ$, representing annual averages. And the depth level bounds are consistent with Jingwei's result, facilitating fair and unified comparisons.

- **CESM2 omip1**: The Community Earth System Model (CESM) is a climate/Earth system coupled model used to simulate past, present, and future climates. The ocean component of CESM2 undergoes various physical model and numerical computation improvements, while utilizing the Marine Biogeochemistry Library (MARBL) to represent ocean biogeochemistry. The experiment, omip1, is driven by the CORE-II (Coordinated Ocean - ice Reference Experiments) atmospheric data, and initialized with physical and biogeochemical ocean observations to conduct ocean dissolved oxygen simulations.

- **CESM2 omip2**: Experiment omip2 shares the same model configuration as omip1 but is forced by the JRA-55 atmospheric data with higher spatial resolution than CORE-II.

- **GFDL-ESM4**: The Earth System Model4 (ESM4) is the fourth-generation chemistry-carbon-climate coupled climate model developed by the Geophysical Fluid Dynamics Laboratory (GFDL). Compared to previous versions, this model more comprehensively represents chemical cycling and ecosystems. Particularly, the model incorporates interactions between ocean ecology and biogeochemistry. The historical experiment utilizes rich climate observational data from 1850 to the present, imposing environmental change conditions consistent with observations to obtain historical simulations of ocean dissolved oxygen.

Jingwei Architecture

In this section, we introduce the detailed architecture design of our proposed Jingwei in Figure 1.

The multi-source gridded DO observation can be represented using four-dimensional coordinates, namely latitude, longitude, depth, and time. Let $\Omega = (\omega_{i,j,d,t})_{i,j,d,t} \in \{0,1\}^{L\times G\times D\times T}$ 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 observation matrix $\mathbf{X}$ as follows:

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

where $\mathbf{X}^{(obs)} \in \mathbb{R}^{L\times G\times D\times T}$ contains all observed entries, NA denotes the indicator of missing data observations, $\odot$ represents the element-wise product operation and $\mathbf{1}_{L\times G\times D\times T}$ is an $L\times G\times D\times T$ matrix filled with ones.

Constrained by the land plate and seabed topography, the dissolved oxygen grid data does not form a regular cube. At the same time, due to the high sparsity of the observation, we replace the traditional CNN-based grid modelling approach with a graph-based modelling to reconstruct DO changes, expanding the receptive field of the AI model.

Graph Modeling. We represent dissolved oxygen profiles as nodes $\mathbf{V}$ in the graph $\mathbf{G}$, and the relationships between the dissolved oxygen profiles as edges $\mathbf{E}$. We further consider two types of edge relationships, defined as proximity neighbors and information hubs:

- **Proximity Neighbors.** Proximity neighbors are based on the *First Law of Geography*, which states that "everything is related to everything else, but near things are more related than distant things." This law suggests that nearby locations exhibit similar characteristics. Thus, we consider the proximity neighbors as the data grids within a small range, e.g., 1° in both latitude and longitude. Regarding the irregular boundaries, we adopt the bedrock elevation data to ensure the rationality.

- **Information Hubs.** Information hubs refer to nodes that, although geographically distant, possess rich observational information. We define the observation completeness C_{obs} over a time span of T timesteps before and after the current moment, i.e., $C_{\text{obs}} = \frac{\|\omega_{i,j,d,t-T:t+T}\|_1}{2T} = \frac{\Sigma_\tau \|\omega_{i,j,d,\tau}\|}{2T}$. When the observation completeness C_{obs} exceeds the threshold c_0, nodes within a distance of up to 5° are regarded as information hubs.

Once the graph modeling is complete, we formulate the DO reconstruction as a node regression problem. Jingwei, as a hybrid learning model, simultaneously utilizes both observation data $\mathbf{X}$ and numerical simulation data $\mathbf{D}_{\text{sim}}$ for model input. Specifically, at time t, we use the observed data $X_{t-T:t-1}$ from the previous T steps and the subsequent T steps $X_{t+1:t+T}$ to expand the observation window. It is important to emphasize that, even if there exists observation for one node at time t, we do not explicitly feed it as input into the Jingwei model. Instead, it is used as a supervisory signal to train the model. To this end, we aim to optimize a model M with parameter θ, i.e., Jingwei, by training on sparse observations to minimize the following objective:

$$\min_\theta \sum_t \mathscr{L}_{\text{ERROR}}(M(X_{t-T:t-1}, X_{t+1:t+T}, \mathbf{G}, \mathbf{D}; \theta) \odot \Omega_t, X_t^{(obs)} \odot \Omega_t), \tag{2}$$

where $\mathscr{L}_{\text{error}}$ is the loss function used to evaluate the deviation between the model's reconstruction results and the true observations. In this study, we use the mean squared error (MSE) loss.

Knowledge Retrieval Module. Ocean climate models encompass simulations of complicated physical, chemical, and biological processes that collectively reflect the ocean deoxygenation mechanisms. Jingwei learns and encodes the governing patterns via data-driven hidden law discovery, hereby predicting unobserved profiles based on essential environmental variables. Knowledge retrieval module aims to transfer the vertical profile patterns from the numerical simulations $\mathbf{D}_{\text{sim}}$ to obtain differentiated and effective node representations without observational dissolved oxygen data. The numerical simulations $\mathbf{D}_{\text{sim}} = \{l_1, l_2, \cdots, l_{|D_{\text{sim}}|}\}$, where $l_i \in R^{d_l}$ is the i-th DO profile, d is the depth range of profile.

First, for each profile l_i we use the Transformer model $\mathscr{F}_{\text{Transformer}}$ as a feature extractor, along with a multi-layer perceptron (MLP) for feature mapping, to perform self-supervised learning on the numerical simulation data. We randomly mask a portion of the vertical profile data (we set the mask ratio to 75% according to the real-world observation missing ratio) and reconstruct it using the vertical sequence variation patterns:

$$h_i = \mathscr{F}_{\text{TRANSFORMER}}(l_i \odot m_i + \text{PE}_i), \quad \hat{l}_i = \text{MLP}(h_i), \quad \mathscr{L}_{\text{self}} = \|l_i - \hat{l}_i\|^2, \tag{3}$$

where $h_i \in R^{d_h}$ is the transformer embedding of DO profile, PE is the sinusoidal positional encoding, m_i is the masking vector, $\mathscr{L}_{\text{self}}$ is the self-supervised loss function. After completing the self-supervised training, we fix the parameters of

the Transformer model and obtain the representations $\mathbf{H}_{\text{sim}} \in R^{|D_{\text{sim}}| \times d_h}$ of the numerical simulation profiles. These high-dimensional representations encompass the diversity of vertical distribution patterns and filter out certain noise fluctuations. Additionally, due to the similarity and redundancy in the profiles, we further merge the profile representations into multiple mutually independent patterns through principal component analysis (PCA) reduction, thereby forming a pattern bank $\mathbf{B} = \text{PCA}(\mathbf{H}_{\text{sim}}) = [b_1, b_2, \cdots, b_{n_B}]$, where $b_m \in R^{d_h}$ is the m-th pattern. Finally, for any given node v_i, we query the pattern bank using its geo-information x_i^{geo} and physicochemical factors x_i^{factor} as a query, selecting the top K most relevant patterns to obtain the node's representation through linear combination.

$$\alpha_{i,m} = [\mathbf{W}_Q \cdot \textsc{Concat}(x_i^{\text{geo}}, x_i^{\text{factor}}) + \mathbf{b}_Q]^T \cdot (\mathbf{W}_K \cdot b_m + \mathbf{b}_K), \tag{4}$$

$$\mathbf{I}_K = \textsc{TopK}(\alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,n_B}; K), \quad z_i = \sum_{k \in \mathbf{I}_K} \frac{\alpha_{i,k}}{\sum_{j \in \mathbf{I}_K} \alpha_{i,j}} b_k, \tag{5}$$

where $\alpha_{i,m}$ is the correlation between node environmental factors and pattern, $\mathbf{W}_Q$, $\mathbf{b}_Q$, $\mathbf{W}_K$, $\mathbf{b}_K$ are model parameters, z_i is the node representation of node v_i.

Knowledge Propagation Module. Discrete in-situ observations reveal the actual spatiotemporal variations of dissolved oxygen. Jingwei establishes abundant connections to perceive the dynamics and spatially heterogeneous oxygen changes on graph topology, further improving reconstruction results under hidden law constraints. Knowledge propagation module aims to perceive the heterogeneous 4D spatiotemporal correlations through adaptive graph message-passing, hereby achieving global and long-term DO reconstruction. Firstly, in order to capture the temporal variation of dissolved oxygen, we adopt a bidirectional LSTM[54] model $f_{\text{Bi-LSTM}}$ to encode both historical and future T timesteps DO observation at time τ as follows:

$$\overrightarrow{Z^{\text{DO}}}(\tau) = \textsc{Bi-Lstm}(X_\tau, \overrightarrow{Z^{\text{DO}}}(\tau - 1); \overrightarrow{\theta}_{\text{Bi-LSTM}}), \quad \overleftarrow{Z^{\text{DO}}}(\tau) = \textsc{Bi-Lstm}(X_\tau, \overleftarrow{Z^{\text{DO}}}(\tau + 1); \overleftarrow{\theta}_{\text{Bi-LSTM}}), \tag{6}$$

where $\overrightarrow{Z^{\text{DO}}}(\tau)$ and $\overleftarrow{Z^{\text{DO}}}(\tau)$ denote the hidden states from two directions, $\overrightarrow{\theta}_{\text{Bi-LSTM}}$ and $\overleftarrow{\theta}_{\text{Bi-LSTM}}$ denote the Bi-LSTM parameters. $Z(\tau) = \overrightarrow{Z^{\text{DO}}}(\tau) \| \overleftarrow{Z^{\text{DO}}}(\tau)$ is the temporal representation at time τ, and $\|$ denotes feature concatenation. The perception of bidirectional time series can simultaneously capture changes in the temporal dimension from both historical patterns and temporal evolution.

Secondly, we propose an adaptive zoning-varying graph message passing to perceive spatial heterogeneity. Different regions of the global ocean exhibit diverse dissolved oxygen distributions influenced by physical processes, biogeochemical processes, and human activities, making it challenging for a unified model to capture spatial changes across various areas. Inspired by the zone partitioning in oceanography[35–37], we find that neural networks can effectively learn different zonings from data. Furthermore, by adjusting the parameters of the neural network based on these different zonings, we can achieve adaptive spatial feature extraction. Specifically, the core of graph neural networks (GNNs) is the message passing mechanism, which involves the transform and aggregation of messages based on the attributes of different nodes. However, vanilla GNNs share parameters for different message transforms, limiting their ability to perceive only similar relationships and preventing them from capturing geographical spatial information. Therefore, our approach is to introduce geographically heterogeneous transforms to the messages. We use geographical coordinates $\mathbf{g} \in R^{N \times d_{geo}}$ as input to generate different message variation matrices $\mathbf{W}^g$ and $\mathbf{b}^g$. For a node v_i, its message transform matrix $\mathbf{W}_i^g = \text{MLP}_W(g_i; \theta_W^g)$ and $\mathbf{b}_i^g = \text{MLP}_b(g_i; \theta_b^g)$. $\mathbf{W}_i^g \in R^{d_{in} \times d_{out}}$ denotes message rotation matrix, and $\mathbf{b}_i^g \in R^{d_{out}}$ denotes message bias vector. Subsequently, compared to vanilla GNNs, we apply different affine transformations to the messages, where the distinct affine parameters encapsulate the neural network's understanding of spatial relationships:

$$\tilde{z}_i^{(l)}(\tau) = \mathbf{W}_i^g \cdot z_i^{(l)}(\tau) + \mathbf{b}^g, \quad z_i^{(l+1)}(\tau) = \frac{1}{|N_i|} \sum_{m \in N_i} \gamma_{i,m} \mathbf{W} \tilde{z}_i^{(l)}(\tau) \tag{7}$$

where $z_i^{(l)}(\tau)$ and $\tilde{z}_i^{(l)}(\tau)$ is the l-th layer GNN embedding and the embedding after affine transformation respectively, N_i is the neighbor set of node v_i. The edge weight $\gamma_{i,m}$ is derived by nonlinear projection of edge feature by a multi-layer perceptron: $\gamma_{i,m} = \textsc{Mlp}(e_{i,m})$, where the edge feature $e_{i,m}$ includes differences in distance, density, longitude, and latitude.

When we concatenate the $\mathbf{W}^g$ and $\mathbf{b}^g$ parameters of different nodes, reduce their dimensions, and normalize them into RGB proportional coefficients, we can visualize the spatial perception capabilities of the neural network. Figures 5 and Figure S7 present the results of spatial perception across multiple global regions, where different colors represent the computational results of the neural network's various zonings. We observe that the neural network's understanding of spatial relationships is multifaceted, encompassing aspects such as ocean currents, gyres, offshore distances, etc.

Multi-task Learning. We employ two objectives for training. The first objective is to directly optimize Jingwei's reconstruction capability by training on observational samples, using root mean square error (RMSE) loss $\mathscr{L}_{reconstruction}$. The second objective involves designing a self-supervised coordinate regression auxiliary task to enhance the supervision signal. Given the very low proportion of training observational data, the overall training of the model lacks effective supervision. However, coordinate position information is readily available, and if the neural network can distinguish between different spatial locations, its understanding of spatial heterogeneity and the characteristics of dissolved oxygen can be enhanced and validated. Therefore, we add an additional output head to the representation layer of Jingwei to regress the node's coordinate values and employ another RMSE loss $\mathscr{L}_{self-supervised}$ as a constraint. The overall loss function of Jingwei is denoted as $\mathscr{L} = \mathscr{L}_{reconstruction} + \lambda \mathscr{L}_{self-supervised}$, where λ is the ratio coefficient for the two loss functions.

Jingwei Optimization

Implementation Details. To train and validate the performance of Jingwei, we divide the multi-source gridded observational data into training, validation, and testing sets in a ratio of 6:1:3. We use the Adam optimizer for training, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay set to 1e-5 for regularization. We adopt a dynamic learning rate scheduling strategy to accelerate model convergence and parameter searching. The initial learning rate was set to 0.1 and decayed to 1e-4 at a rate of 0.95 as the number of training epochs increased. During training, we randomly shuffle the years and conduct grouped training. Due to the large scale of the graph nodes, we sample the graph at each training step, selecting 30 nodes and their corresponding second-order subgraphs. Every 30 years is treated as one validation round for performance assessment. We employ an early stopping strategy to select the optimal model parameters, with patience set to 10. We implement Jingwei model based on Pytorch 2.1 framework and PyG (PyTorch Geometric 2.4) library. All the experiments are implemented on a server with NVIDIA GeForce RTX 4090 GPU (24 GB memory).

Hyperparamter Setting. The number of layers for zoning-varying graph message passing is set to 2, with the hidden layer dimension set to 64. Both historical and future time steps T are set to 5, and the ratio coefficient λ for the two loss functions is set to 0.1. All geographical location coordinates are converted to spherical coordinates. In the knowledge retrieval transfer, the Transformer encoder is set to 2 layers due to the fact that the number of features in the sequence is only 1; deeper layers may lead to overfitting. The combination of the top K vertical profile patterns is set with $K = 10$.

The name of 'Jingwei'

Jingwei is a classic figure in Chinese mythology, featured in the "Shan Hai Jing" (see https://en.wikipedia.org/wiki/Jingwei). The story tells of Jingwei, the daughter of Emperor Yan, who drowned in the East Sea. She was reborn as a bird and decided to fill the sea with pebbles and twigs, endeavoring to prevent similar tragedies. Today, Jingwei symbolizes perseverance and determination, embodying the spirit of never giving up despite difficult challenges.

In our work, we thank all marine scientists, researchers, and technicians who tirelessly venture into the field to measure oceanic dissolved oxygen data. Although their collected data might seem as extremely sparse as the pebbles and twigs Jingwei used, their persistent efforts aim to reveal the patterns of global ocean deoxygenation. To honor the spirit of these ocean explorers, we have named the AI-driven algorithm for reconstructing ocean deoxygenation proposed in this paper "Jingwei" and designed its logo as shown in Figure S11.

Data availability

In our work, we aggregate comprehensive observation datasets of global ocean from 5 publicly available database as follow: World Ocean Database (WOD 2018, https://www.ncei.noaa.gov/products/world-ocean-database), CLIVAR and Carbon Hydrographic Database (CCHDO, https://cchdo.ucsd.edu/), Argo (https://argo.ucsd.edu/data/), Global Ocean Data Analysis Project version2.2022 (GLODAPV2 2022, https://glodap.info/), Geotraces IDP (https://www.geotraces.org/). To ensure the availability and accuracy of the analysis data from various sources, we have set unified quality control standards. In addition, we access the numerical simulation data from the Coupled Model Inter-comparison Project Phase 6 (CMIP6), which are available at https://esgf-node.llnl.gov/search/cmip6/.

Code availability

The source code for the Jingwei model is available on GitHub at the following link: https://github.com/RobinLu1209/Jingwei. The repository includes methods for code training and testing, as well as the versions of the libraries on which the code depends. Additionally, we provide a visualization website where users can browse the reconstructed results of dissolved oxygen.

References

1. Schmidtko, S., Stramma, L. & Visbeck, M. Decline in global oceanic oxygen content during the past five decades. *Nature* **542**, 335–339 (2017).
2. Breitburg, D. *et al.* Declining oxygen in the global ocean and coastal waters. *Science* **359**, eaam7240 (2018).
3. Helm, K. P., Bindoff, N. L. & Church, J. A. Observed decreases in oxygen content of the global ocean. *Geophys. Res. Lett.* **38** (2011).
4. Grégoire, M. *et al.* A global ocean oxygen database and atlas for assessing and predicting deoxygenation and ocean health in the open and coastal ocean. *Front. Mar. Sci.* **8**, 724913 (2021).
5. Keeling, R. F., Körtzinger, A. & Gruber, N. Ocean deoxygenation in a warming world. *Annu. review marine science* **2**, 199–229 (2010).
6. Morée, A. L., Clarke, T. M., Cheung, W. W. & Frölicher, T. L. Impact of deoxygenation and warming on global marine species in the 21st century. *Biogeosciences* **20**, 2425–2454 (2023).
7. Ricour, F., Guidi, L., Gehlen, M., DeVries, T. & Legendre, L. Century-scale carbon sequestration flux throughout the ocean by the biological pump. *Nat. Geosci.* **16**, 1105–1113 (2023).
8. Li, C. *et al.* The ocean losing its breath under the heatwaves. *Nat. Commun.* **15**, 6840 (2024).
9. Gong, H., Li, C. & Zhou, Y. Emerging global ocean deoxygenation across the 21st century. *Geophys. Res. Lett.* **48**, e2021GL095370 (2021).
10. Oschlies, A., Brandt, P., Stramma, L. & Schmidtko, S. Drivers and mechanisms of ocean deoxygenation. *Nat. Geosci.* **11**, 467–473 (2018).
11. Boyer, T. *et al.* World ocean database 2018 (2018).
12. Danabasoglu, G. *et al.* The community earth system model version 2 (cesm2). *J. Adv. Model. Earth Syst.* **12**, 10.1029/2019ms001916 (2020).
13. Dunne, J. P. *et al.* The gfdl earth system model version 4.1 (gfdl-esm 4.1): Overall coupled model description and simulation characteristics. *J. Adv. Model. Earth Syst.* **12**, 10.1029/2019ms002015 (2020).
14. Reagan, J. R. *et al.* World ocean atlas 2023 (2024). Dataset: NCEI Accession 0270533.
15. He, G. *et al.* World ocean atlas 2018, volume 3: Dissolved oxygen, apparent oxygen utilization, and dissolved oxygen saturation. Report, National Centers for Environmental Information (NCEI) (2019).
16. Zhou, Y., Gong, H. & Zhou, F. Responses of horizontally expanding oceanic oxygen minimum zones to climate change based on observations. *Geophys. Res. Lett.* **49**, e2022GL097724 (2022).
17. Acharya, S. S. & Panigrahi, M. K. Eastward shift and maintenance of arabian sea oxygen minimum zone: Understanding the paradox. *Deep. Sea Res. Part I: Oceanogr. Res. Pap.* **115**, 240–252 (2016).
18. Stramma, L., Johnson, G. C., Sprintall, J. & Mohrholz, V. Expanding oxygen-minimum zones in the tropical oceans. *science* **320**, 655–658 (2008).
19. Giglio, D., Lyubchich, V. & Mazloff, M. R. Estimating oxygen in the southern ocean using argo temperature and salinity. *J. Geophys. Res. Ocean.* **123**, 4280–4297 (2018).
20. Stramma, L., Schmidtko, S., Levin, L. A. & Johnson, G. C. Ocean oxygen minima expansions and their biological impacts. *Deep. Sea Res. Part I: Oceanogr. Res. Pap.* **57**, 587–595 (2010).
21. Eyring, V. *et al.* Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geosci. Model. Dev.* **9**, 1937–1958 (2016).
22. Long, M. C. *et al.* Simulations with the marine biogeochemistry library (marbl). *J. Adv. Model. Earth Syst.* **13** (2021).
23. Frölicher, T. L. *et al.* Contrasting upper and deep ocean oxygen response to protracted global warming. *Glob. biogeochemical cycles* **34**, e2020GB006601 (2020).
24. Koelling, J., Atamanchuk, D., Wallace, D. W. R. & Karstensen, J. Decadal variability of oxygen uptake, export, and storage in the labrador sea from observations and cmip6 models. *Front. Mar. Sci.* (2023).
25. Huang, S. *et al.* Reconstruction of dissolved oxygen in the indian ocean from 1980 to 2019 based on machine learning techniques. *Front. Mar. Sci.* (2023).

26. Busecke, J. J., Resplandy, L., Ditkovsky, S. J. & John, J. G. Diverging fates of the pacific ocean oxygen minimum zone and its core in a warming world. *AGU Adv.* **3**, e2021AV000470 (2022).

27. Kwiatkowski, L. *et al.* Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from cmip6 model projections. *Biogeosciences* **17**, 3439–3470 (2020).

28. Bopp, L., Resplandy, L., Untersee, A., Le Mezo, P. & Kageyama, M. Ocean (de) oxygenation from the last glacial maximum to the twenty-first century: insights from earth system models. *Philos. Transactions Royal Soc. A: Math. Phys. Eng. Sci.* **375**, 20160323 (2017).

29. Chen, M. *et al.* Iterative integration of deep learning in hybrid earth surface system modelling. *Nat. Rev. Earth & Environ.* **4**, 568–581 (2023).

30. Eyring, V. *et al.* Pushing the frontiers in climate modelling and analysis with machine learning. *Nat. Clim. Chang.* 1–13 (2024).

31. Wu, Z. *et al.* A comprehensive survey on graph neural networks. *IEEE transactions on neural networks learning systems* **32**, 4–24 (2020).

32. Wang, S., Cao, J. & Philip, S. Y. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge data engineering* **34**, 3681–3700 (2020).

33. Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M. & Zitnik, M. Multimodal learning with graphs. *Nat. Mach. Intell.* **5**, 340–350 (2023).

34. Tobler, W. R. A computer movie simulating urban growth in the detroit region. *Econ. geography* **46**, 234–240 (1970).

35. Fay, A. & McKinley, G. Global open-ocean biomes: mean and temporal variability. *Earth Syst. Sci. Data* **6**, 273–284 (2014).

36. Reygondeau, G. *et al.* Climate change-induced emergence of novel biogeochemical provinces. In *Frontiers in Marine Science* (2020).

37. Sonnewald, M., Dutkiewicz, S., Hill, C. N. & Forget, G. Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Sci. Adv.* **6** (2020).

38. Ha, D., Dai, A. M. & Le, Q. V. Hypernetworks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (OpenReview.net, 2017).

39. Lu, B. *et al.* Spatio-temporal graph few-shot learning with cross-city knowledge transfer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1162–1172 (2022).

40. Wang, J. *et al.* Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge data engineering* **35**, 8052–8072 (2022).

41. Odena, A., Dumoulin, V. & Olah, C. Deconvolution and checkerboard artifacts. *Distill* **1**, e3 (2016).

42. Liu, Y., Ganguly, A. R. & Dy, J. Climate downscaling using ynet: A deep convolutional network with skip connections and fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3145–3153 (2020).

43. Kanopoulos, N., Vasanthavada, N. & Baker, R. L. Design of an image edge detection filter using the sobel operator. *IEEE J. solid-state circuits* **23**, 358–367 (1988).

44. Ke, J., Wang, Q., Wang, Y., Milanfar, P. & Yang, F. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5148–5157 (2021).

45. Su, S. *et al.* Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3667–3676 (2020).

46. Schlitzer, R. Interactive analysis and visualization of geoscience data with ocean data view. *Comput. & geosciences* **28**, 1211–1218 (2002).

47. Bao, Y. & Li, Y. Simulations of dissolved oxygen concentration in cmip5 earth system models. *Acta Oceanol. Sinica* **35**, 28–37 (2016).

48. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. machine learning research* **9** (2008).

49. Garcia, H. E., Boyer, T. P., Levitus, S., Locarnini, R. A. & Antonov, J. On the variability of dissolved oxygen and apparent oxygen utilization content for the upper world ocean: 1955 to 1998. *Geophys. Res. Lett.* **32** (2005).

50. Ito, T., Minobe, S., Long, M. C. & Deutsch, C. Upper ocean o2 trends: 1958–2015. *Geophys. Res. Lett.* **44**, 4214–4223 (2017).

51. Sperling, E. A. *et al.* Oxygen, ecology, and the cambrian radiation of animals. *Proc. Natl. Acad. Sci.* **110**, 13446–13451 (2013).

52. do Rosário Gomes, H. *et al.* Massive outbreaks of noctiluca scintillans blooms in the arabian sea due to spread of hypoxia. *Nat. communications* **5**, 4862 (2014).

53. NOAA National Centers for Environmental Information. Etopo 2022 60 arc-second global relief model (2022).

54. Graves, A., Fernández, S. & Schmidhuber, J. Bidirectional lstm networks for improved phoneme classification and recognition. In *International conference on artificial neural networks*, 799–804 (Springer, 2005).

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (No. T2421002, 62020106005, 62272301, 623B2071), National Key R&D Program of China (No.2022YFB3904204), National Natural Science Foundation of China (No. 42125601, 42276201, 42050105, 62061146002), and Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University. We thank Hongjing Gong, Julius Busecke, Huanting Hu, Zhuoyi Zhu for their in-depth discussions and suggestions for the paper. We thank Fangchao Zhou for designing the logo of Jingwei model, shown in Figure S11.

Author contributions statement

J.Z. and X.W. conceived and designed this project. B.L. developed Jingwei model, designed experiments, and conducted performance evaluation. L.H., Z.Z., H.Q. contributed to the data processing and benchmark comparisons. L.H., L.Q., Y.Z. and L.Z. performed the oceanographic data analysis and interpretation of the results. X.G. and F.L. supervised the project and investigated the methodology. B.L. took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryvideo.mp4
- NatureMachineIntelligenceSI.pdf