https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Open Access Earth System Science Data Discussions

A Novel Global Gridded Ocean Oxygen Product Derived from a Neural Network Emulator and in-situ observations

Said Ouala[1], Oussama Hidaoui[2], and Zouhair Lachkar[3]

[1]IMT Atlantique, CNRS UMR Lab-STICC, INRIA Team Odyssey, Brest, France
[2]African Institute for Mathematical Sciences, South Africa
[3]Mubadala Arabian Center for Climate and Environmental Sciences, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

Correspondence: Said Ouala (said.ouala@imt-atlantique.fr)

Abstract.

Ocean deoxygenation, driven by climate change, poses significant challenges to marine ecosystems and can profoundly alter nutrient and carbon cycling. Quantifying the rate and regional patterns of deoxygenation relies on spatio-temporal interpolation tools to fill gaps in observational coverage of dissolved oxygen. However, this task is challenging due to the sparsity of observations, and classical interpolation methods often lead to high uncertainty and biases, typically underestimating long-term deoxygenation trends. In this work, we develop a novel gridded dissolved oxygen product by integrating direct oxygen observations with machine-learning-based emulated oxygen estimates derived from temperature and salinity profiles. The gridded product is then generated through optimal interpolation of both the observed and emulated data. The resulting product shows strong agreement with baseline climatology and captures well-known patterns of seasonal variability and long-term deoxygenation trends. It also outperforms current state-of-the-art products by more accurately capturing dissolved oxygen variability at synoptic and decadal scales, and by reducing uncertainty around long-term changes. This study highlights the potential of combining machine learning with classical interpolation methods to generate improved gridded biogeochemical products, enhancing our ability to study and understand ocean biogeochemical processes and their variability under a changing climate.

1 Introduction

The global oxygen content of the ocean has been declining over recent decades Ito et al. (2017); Schmidtko et al. (2017) and is projected to continue decreasing throughout the current century Bopp et al. (2013); Kwiatkowski et al. (2020), leading to detrimental consequences for marine organisms Rabalais et al. (2002); Vaquer-Sunyer and Duarte (2008); Laffoley and Baxter (2019) and profound changes in biogeochemical cycles Codispoti et al. (2001). These changes can affect the ocean's emissions and uptake of greenhouse gases, thereby influencing Earth's climate Gruber (2008); Keeling et al. (2010); Lachkar et al. (2024). Recent estimates of the global decline in oceanic oxygen range from 0.5% to 3.3 % relative to the climatology for the period 1970–2010 Schmidtko et al. (2017). However, these estimates are subject to significant uncertainty, particularly in data-sparse

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

regions. A major source of uncertainty in assessing and understanding ocean deoxygenation is the limited spatial and temporal coverage of dissolved oxygen observations.

Despite advances in autonomous profiling floats, underwater vehicles, and large-scale ocean sensing programs such as ARGO, dissolved oxygen observations remain insufficient to accurately estimate deoxygenation rates at both global and regional scales Gruber et al. (2010); Claustre et al. (2020). Large regions, particularly in the South Pacific and Indian Ocean, remain undersampled Hermes et al. (2019); Grégoire et al. (2021). In this context, assessing global and regional trends in ocean oxygen content requires developing interpolation methods that map available data onto a regular space-time grid. Gridded oxygen products also play a crucial role in validating ocean models, including both global models used in Earth System Models (ESMs) and regional models necessary for projecting the impact of climate change on oxygen at regional scales. However, as highlighted in Ito et al. (2024b), standard interpolation techniques commonly used to generate these products tend to underestimate oxygen trends in data-sparse regions, leading to a potential underestimation of global ocean deoxygenation.

Recently, various studies have demonstrated that machine learning (ML) techniques can outperform classical state-of-the-art methods in geosciences for applications such as forecasting Lam et al. (2023); Bi et al. (2023), simulation Nguyen et al. (2023); Dheeshjith et al. (2024); Ouala et al. (2023), and data assimilation Cheng et al. (2023). From a modeling perspective, ML techniques have been successfully used to emulate ocean models at both short Wang et al. (2024); Aouni et al. (2024) and long timescales Dheeshjith et al. (2024). Additionally, ML models show promise in developing data-driven, automated tuning methods for ESMs Ouala et al. (2024); Kochkov et al. (2024), where gridded oxygen products serve as valuable references for model calibration Sharp et al. (2022); Ito et al. (2024a). ML has also been applied to produce gridded products of ocean variables from partial and noisy ocean observations Martin et al. (2023, 2024). In this context, recent studies Sharp et al. (2022); Ito et al. (2024a) have begun exploring ML-based approaches for generating gridded ocean oxygen products. While these studies highlight the potential of ML-based emulators in generating gridded oxygen products, they typically rely on existing gridded temperature and salinity datasets. For instance, Sharp et al. (2022) derived oxygen fields from the (Roemmich and Gilson, 2009) (RG09) Argo Climatology, which inherently restricts its applicability to the ARGO era, limiting its use for long-term climate trend analysis. Similarly, Ito et al. (2024a) emulated oxygen using monthly gridded temperature and salinity datasets, obtained from the Hadley Center EN version 4 Good et al. (2013), extending the temporal coverage at the cost of excluding marginal seas. Moreover, interpolation errors inherent to these products can bias ML-based emulators and degrade the quality of the data they are trained on, particularly in data-sparse regions.

Here, we propose a novel method to generate a gridded oxygen concentration product covering the period from 1965 to 2022, using observed ocean oxygen data and emulated profiles derived from temperature and salinity measurements. Specifically, we expand the available oxygen observations by training a neural network emulator to predict oxygen concentrations from temperature and salinity profiles. After a quality control process for the emulated profiles, we apply optimal interpolation (OI) to combine the real and emulated data into a unified gridded product. This approach offers several advantages over traditional interpolation methods and recent ML-based techniques, including: (i) an increased density of observations, (ii) a stand-alone data product that is independent of existing interpolated products, and (iii) a flexible temporal and spatial resolution that can be adjusted based on the density of available observations. The resulting product Ouala et al. (2025) demonstrates good agreement

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

with expected spatial and temporal variability of dissolved oxygen, particulalry regarding global deoxygenation rates. It also exhibits improved performance over current state-of-the-art products in better resolving dissolved oxygen variability at synoptic scales of the order of $O(10^3)$km and in reproducing the climatological seasonal cycle near the ocean surface. Interestingly, when compared to classical optimal interpolation of direct dissolved oxygen measurements, our product better resolves decadal variability and significantly reduces the uncertainty of the reconstructed field, particularly in data-sparse regions, where direct measurements of dissolved oxygen are lacking.

2 Materials and Methods

We construct a gridded dissolved oxygen concentration product by combining dissolved oxygen observations with emulated estimates derived from temperature and salinity profiles. The process involves:

- **Building quality-controlled datasets:** The quality control (QC) is based on both World Ocean Database (WOD) flags and additional relevant QC criteria inspired by the work of Schmidtko et al. (2017).
- **Neural Network Emulation**: A neural network is trained to emulate oxygen profiles from temperature and salinity measurements.
- **Optimal Interpolation:** Observed and emulated data are combined using OI to produce the final gridded product.

Each of these steps is detailed in the following sections.

2.1 Data

The data used in this study was obtained from WOD, the largest publicly available collection of uniformly formatted, quality-controlled ocean profile data. We use dissolved oxygen (DO), temperature (T), and salinity (S) data sourced from various platforms and institutions (refer to Tables B1 and B2 in the SI Appendix B for a detailed description of data sources).

Standard QC from WOD is applied to the T, S, and DO profiles, retaining only observations flagged as accepted values with flag 0. Furthermore, additional QC is implemented for the oxygen data following the methodology outlined in Schmidtko et al. (2017). Specifically, oxygen profiles with a maximum-minimum difference of less than 5 $\mu mol/kg$ and those with differences of less than 0.5 $\mu mol/kg$ across 18 depth levels were excluded. Furthermore, profiles with surface oxygen concentrations below 100 $\mu mol/kg$ were removed. Additional quality control steps targeted supersaturation anomalies. Specifically, profiles exhibiting supersaturation at depths exceeding 200 m with supersaturation levels above 115% were discarded. Additionally, profiles with surface oxygen concentrations below 90% of the expected saturation level were excluded.

We constructed two distinct datasets: one for training the ML model and another for generating the gridded DO product. The dataset used to train the ML model consists of collocated pairs of T, S and (DO) data from 1965 to 2022. The dataset was split into training and validation datasets. The training dataset consists of 80 % of the total data, while the validation dataset accounts for the remaining 20%. The validation dataset consists of 23 independent $1^\circ \times 1^\circ$ test regions distributed across all

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

major ocean basins. The location of these test regions, along with the evaluation of the ML-based emulator, are provided in Figure 2. The dataset used to generate the gridded DO product consists of DO data from the preprocessed WOD profiles and emulated DO samples generated by the neural network. To ensure that the emulated DO samples used in the interpolation of the final gridded product fall within the observed range of ocean oxygen variability, the emulated DO data undergo a final QC check, as described in Section 2.3.

2.2 Machine learning algorithm training and validation

2.2.1 Model architecture

The machine learning model used to emulate DO data is a Multilayer Perception (MLP) model. This model was used in various studies to predict ocean biogeochemical variables, including nutrient concentrations and carbonate system parameters both globally Sauzède et al. (2017) and in regional configurations Fourrier et al. (2020). It was also applied to construct gridded oxygen datasets based on interpolated temperature and salinity products Sharp et al. (2022); Ito et al. (2024a).

Following the approach of Sauzède et al. (2017), the input of the neural network consists of water temperature (T), salinity (S), Dissolved oxygen saturation (DOS), latitude (LAT), longitude (LON) and month of the year (MOY). The periodicity of LON and MOY were accounted using a sinusoidal encoding of these variables. For additional details on the hyperparameters of the model, the training algorithm and the evaluation metrics, please refer to the SI appendix A.

2.2.2 Performance of the ML model

We use independent test regions to evaluate the ability of the trained neural network model to emulate dissolved oxygen concentrations. The scatter plot of the predicted vs true oxygen values highlighted in Fig. 1 shows that the trained model is able to accurately predict dissolved oxygen concentration with a root mean squared error $RMSE = 12.13\mu mol/kg$ and a correlation coefficient $R^2 = 0.98$.

An analysis of the average vertical profiles based on observations and as emulated by the ML model indicates that the model aligns well with the observed profile distribution and successfully captures dissolved oxygen variability in all independent test regions (Figure 2). In particular, the model accurately predicts low oxygen concentrations associated with Oxygen Minimum Zones (OMZs), as well as the vertical and spatial variability of oxygen in test regions located far from OMZs.

To identify potential trends and biases in the neural network model, we further evaluate the distribution of the difference between the emulated oxygen values and the observations in the independent test regions, as shown in Figure 3. Overall, the boxplots indicate that the model accurately captures the variability of ocean oxygen across both space and time. Notably, larger errors are observed in the upper ocean layers, which correspond to higher oxygen concentrations. Additionally, higher errors are seen in the earlier years, likely due to the limited number of observations during this period.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

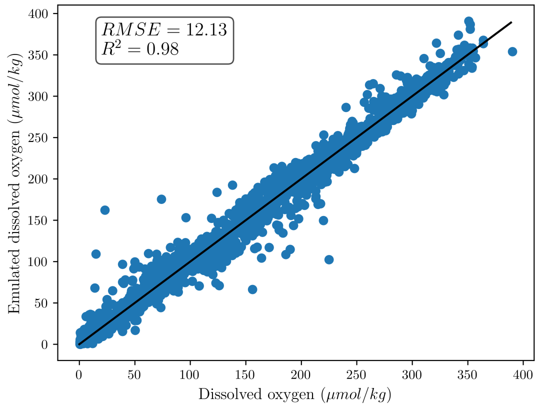

Figure 1. ***Scatter plot of the emulated oxygen data with respect to the ground truth measurements.*** The plot includes emulated oxygen values across all depths and within all independent test regions.

2.3 Validation of oxygen data emulated by machine learning

The trained ML model can generate outliers, particularly when emulating oxygen data based on temperature and salinity measurements at locations that were not represented in the training data. This phenomenon, known as out-of-distribution sampling in the ML community, can result in biases in the final interpolated product, especially in regions with a high density of emulated data but few or no actual ocean oxygen observations.

To address this issue, we design a QC framework to validate the emulated oxygen data before incorporation into the global interpolation. The primary goal of this framework is to exclude anomalous emulated oxygen data while minimizing the rejection of valid data. The QC framework begins by computing validity thresholds, derived from historical dissolved oxygen measurements. These thresholds represent the minimum (min) and maximum (max) observed values, calculated for each location and depth of the interpolation grid. They reflect the natural variability of oxygen in the ocean and serve as a baseline for identifying potential outliers (please refer to Figure C1 in the SI Appendix C for an illustration of these min/max filters.)

The computation of these min-max fields is based on a dynamic binning of the observations in the interpolation grid, where the measurements used to compute the min-max statistics of each grid cell can include neighboring grid cells until a sufficient number of 15 observations is collected. The observations are also aggregated within a $\pm$5-year window around the target year. This temporal aggregation helps increasing the number of observation points in data-poor regions and reduces the impact of short-term anomalies while capturing broader trends in oxygen variability.

Emulated oxygen values are validated by comparing them with the established thresholds for their specific location, depth, and time. Values falling outside the range defined by the min and max thresholds are flagged as outliers and excluded from the interpolation.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Open Access Earth System Science Data Discussions

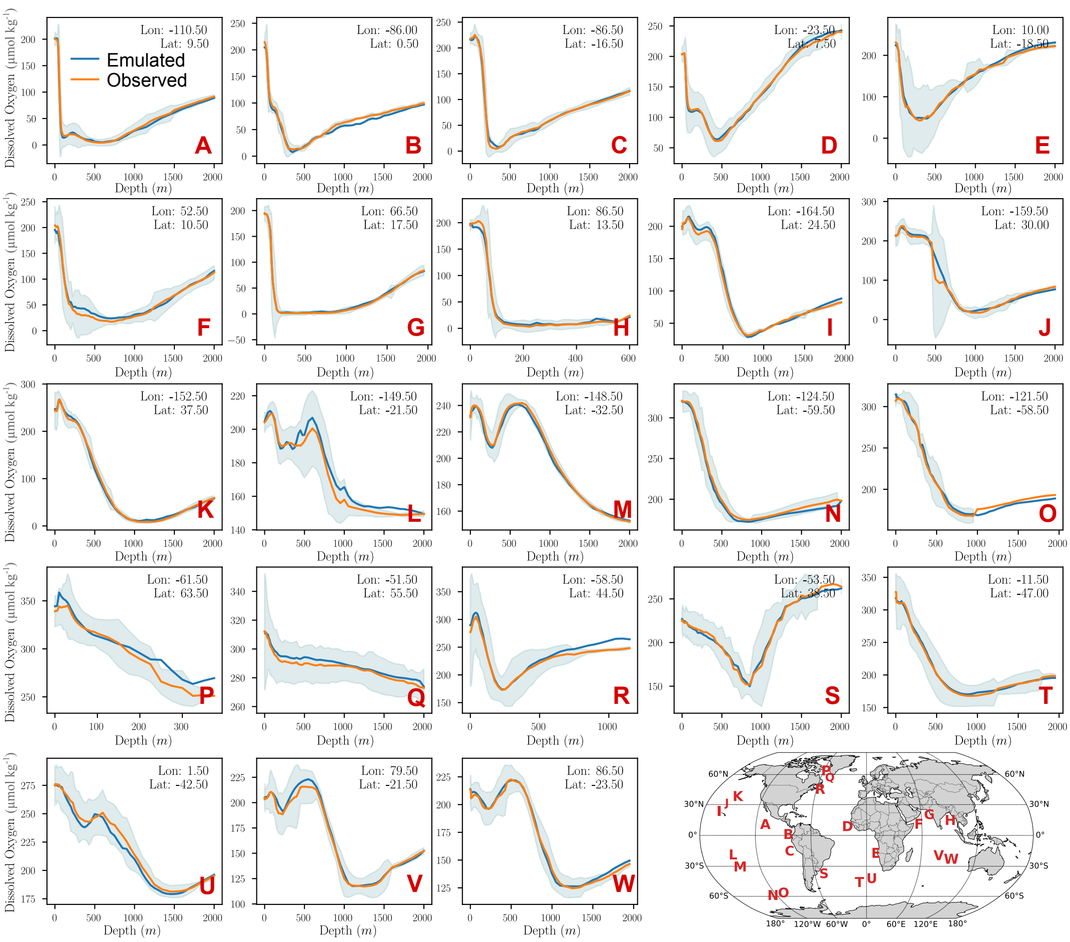

Figure 2. ***Mean emulated oxygen profiles compared to the empirical distribution of observed oxygen profiles in the independent test regions.*** The mean emulated profile (orange) against the empirical distribution of observed dissolved oxygen profiles (blue line and shading) in the independent $1^\circ \times 1^\circ$ test regions. The longitude and latitude of each of the 23 test regions are indicated in the corresponding panels. The locations of the test regions are also shown in the bottom-right corner of the figure.

2.4 Optimal Interpolation

2.4.1 General description

Quality-controlled dissolved oxygen data and emulated data are binned into a global $1^\circ \times 1^\circ$ horizontal grid with 65 standard WOD depth levels ranging from 0 m to 2000 m. These binned data are interpolated using standard Optimal Interpolation (OI) to produce an analysis field of dissolved ocean oxygen and its corresponding uncertainty.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Open Access Earth System Science Data Discussions

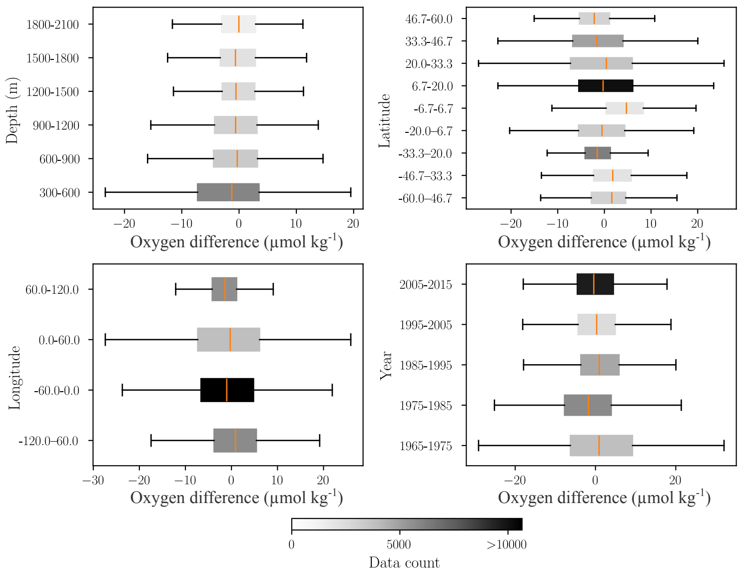

Figure 3. ***Box plots of the differences between emulated and observed oxygen concentrations in the independent test regions.*** The box plots are binned based on depth, year, longitude and latitude. The intensity of the shading in each box refers to the number of data points. For each box, the negative and positive whiskers represent the Q1–1.5*IQR and Q3 + 1.5*IQR, respectively, where Q1 is the 0.25 quantile, Q3 the 0.75 quantile, and IQR the inter-quantile range. The width of each box represents the IQR and the middle line the median of the values.

The OI method combines a background information with observations to derive the analysis state as follows:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}_b) \tag{1a}$$

$$\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}_b - \mathbf{K}\mathbf{H}\boldsymbol{\Sigma}_b \tag{1b}$$

where $\mathbf{x}_a$ is the analyzed field, $\boldsymbol{\Sigma}_a$ is the covariance of the analysis, $\mathbf{x}_b$ is the background field, $\boldsymbol{\Sigma}_b$ is the covariance of the background field, $\mathbf{y}$ are the binned observations, and $\mathbf{H}$ is the observation operator. The Kalman gain $\mathbf{K}$ is computed as:

$$\mathbf{K} = \boldsymbol{\Sigma}_b\mathbf{H}^\top \left(\mathbf{H}\boldsymbol{\Sigma}_b\mathbf{H}^\top + \boldsymbol{\Sigma}_o\right)^{-1}, \tag{2}$$

where $\boldsymbol{\Sigma}_\mathbf{o}$ are the background and observation error covariance matrices, respectively.

2.4.2 Background and observation error covariances

For convenience, we drop the dependence of the Kalman Gain matrix on the observation model by introducing the data-grid and the data-data covariance matrices $\boldsymbol{\Sigma}_\mathbf{xy}$ and $\boldsymbol{\Sigma}_\mathbf{yy}$ respectively which are defined as follows:

$$\boldsymbol{\Sigma}_\mathbf{xy} = \boldsymbol{\Sigma}_b\mathbf{H}^\top \tag{3a}$$

$$\boldsymbol{\Sigma}_\mathbf{yy} = \mathbf{H}\boldsymbol{\Sigma}_b\mathbf{H}^\top \tag{3b}$$

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

In this work, the covariance matrices $\mathbf{\Sigma_{xy}}$ and $\mathbf{\Sigma_{yy}}$ are built using the same isotropic Gaussian prior as follows:

$$\mathbf{\Sigma_{x\cdot}} = \sigma_b^2 exp(-\frac{L_{m,n}}{2L_h}) \quad (4)$$

where σ_b^2 is the total variance of the background field and $L_{m,n}$ is the distance between two grid points m and n. L_h is the e-folding horizontal length scale. In this work, we set L_h to 1000 km for the interpolation of the yearly product between 1965 and 2022, and we set L_h to 300 km for the interpolation of the monthly product between 2005 and 2022.

The background covariance matrix is assumed to be diagonal and is computed from the variance of the binned observations.

2.4.3 Observation Error Covariance Matrix

The observation error covariance matrix $\mathbf{\Sigma}_o$ is assumed to be diagonal. Each diagonal element $(\mathbf{\Sigma}_o)_{n,n}$ represents the variance of the binned observations at grid point n. This variance is derived from the following sources of uncertainty:

Gridding Uncertainty (σ_g^2): This component arises from approximating the oxygen distribution at each grid cell using the empirical distribution based on the observed and emulated samples.

Measurement Uncertainty (σ_m^2): This source of uncertainty is attributed to each dissolved oxygen measurement. It is estimated, following the methodology described in Sharp et al. (2022), as 3% of the gridded oxygen measurements.

Emulation Uncertainty (σ_e^2): This uncertainty arises from the emulated oxygen data. As discussed in section 2.2.2, the gridded emulated oxygen measurements align with gridded oxygen data. In this context, the uncertainty estimate of the emulated oxygen data is computed similarly to the dissolved oxygen measurements, with a slight increase to 4% of the gridded emulated oxygen measurements.

The total variance at grid point n is computed, assuming independence, as:

$$(\mathbf{\Sigma}_o)_{n,n} = \sigma_{g,n}^2 + \sigma_{m,n}^2 + \sigma_{e,n}^2, \quad (5)$$

2.4.4 Data aggregation

The observed and emulated oxygen data are aggregated over a time window based on the product's resolution. For the yearly (respectively, monthly) product, each grid cell at year (respectively, month) t includes data from $t \pm 2$ years (respectively, $t \pm 2$ months).

3 Results

We analyze the variability of dissolved oxygen in our gridded oxygen product and compare it with that based on the World Ocean Atlas 23 (WOA 23) and other existing ML-based products Sharp et al. (2022); Ito et al. (2024a). Specifically, we examine the annual-mean climatology, as well as the seasonal and long-term variability, and contrast our findings with those from these reference datasets.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Open Access Earth System Science Data Discussions

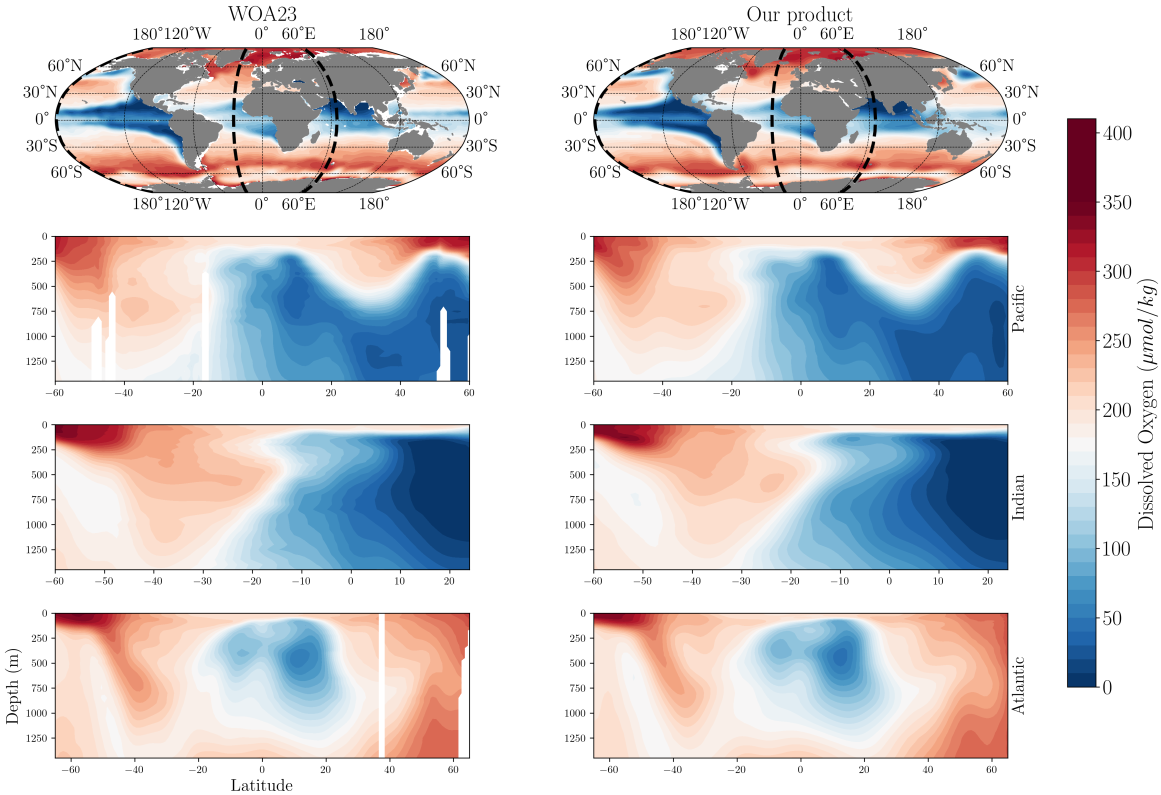

Figure 4. ***Spatial distribution of dissolved oxygen concentration in our product compared to WOA23.*** The first row shows the climatological annual-mean dissolved oxygen concentration at 200 m depth. Dashed black lines represent the locations of the meridional cross sections from the surface to 1500 m. The remaining three rows display oxygen along meridional cross-sections in the Pacific (-179 °E), Indian (65 °E), and Atlantic (-25 °E) basins in WOA23 (left) and our product (right).

3.1 Annual mean climatology

We first compare the spatial distribution of dissolved oxygen in our product to the WOA23 baseline (Figure 4). The horizontal variability of oxygen at a depth of 200 m closely matches the spatial distribution of oxygen concentrations in the WOA23 product, showing strong agreement in the location and intensity of oxygen minimum zones, as well as the high oxygen concentrations at higher latitudes. Similarly, the vertical distribution in the Atlantic, Indian, and Pacific basins aligns well with WOA23, accurately capturing vertical oxygen gradients and the depth and intensity of oxygen minimum zones.

Beyond the climatological spatial distribution, we assess the spatial resolution of the proposed DO product relative to state-of-the-art machine learning-based gridded datasets. Figure 5 presents an example of the DO anomaly field in the equatorial Pacific (-179°E to -100°E, 30°S to 30°N) alongside the vertically and monthly averaged Radially Averaged Power Spectral Density (RAPSD) of our product, compared with GOBAI-O2 and Ito et al. (2024). This region is characterized by energetic synoptic-scale variability Chelton et al. (2007), and we evaluate whether our product better captures these processes. The

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

RAPSD indicates 100-200% higher energy levels at wavelengths around $O(10^3)$ km, compared to the ML baselines, suggesting an improved representation of small-scale variability. This is further illustrated through the visual analysis of the anomaly field in Figure 5, where our product better resolves finer synoptic-scale structures on the order of $O(10^3)$ km, revealing more energetic small-scale eddies than the other ML-based products.

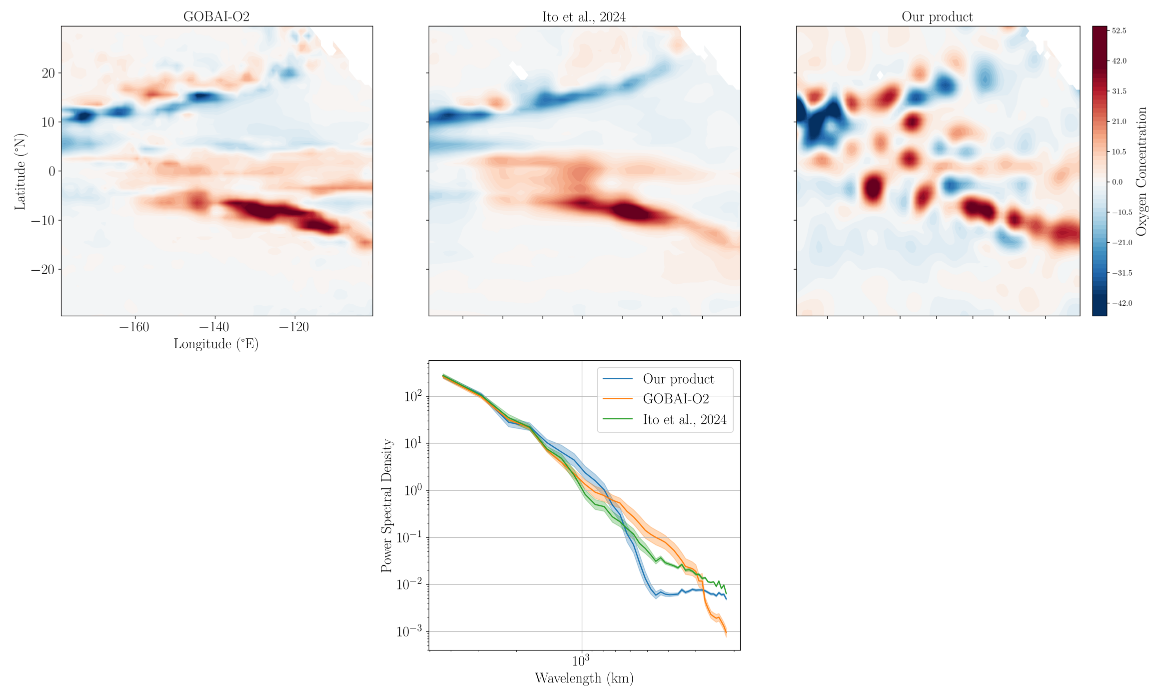

Figure 5. ***Snapshot of oxygen monthly anomalies in the equatorial Pacific.*** Snapshot of the anomaly of the gridded product in the equatorial Pacific region (-179°E to -100°E, 30°S to 30°N) in July 2015 compared to the GOBAI-O2 and Ito et al. (2024a)'s ML baselines. The upper row displays a horizontal section at 200 m depth. The bottom panel shows the distribution of the vertically and monthly averaged RAPSD computed in the equatorial Pacific region.

3.2 Seasonal variability

We also compare the climatological seasonal cycle of oxygen in our product across both hemispheres with that from WOA23 and previous ML-based products, including the GOBAI-O2 product Sharp et al. (2022) and the Ito et al. (2024) product Ito et al. (2024a) (Figure 6). Overall, the seasonal oxygen cycle in our product closely aligns with WOA23 and previous ML-based estimates. As expected, oxygen seasonality is more pronounced in the upper ocean, reflecting the strong seasonal variability in temperature (Figure 6 and Table 1). Indeed, when compared to other state-of-the-art ML-based models, namely GOBAI-O2 and Ito et al. (2024), our product shows a stronger correlation between upper-ocean oxygen seasonality and temperature variability, particularly in the Southern Hemisphere. This relationship is more pronounced relative to previous ML-based products (Table 1).

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Region	Dataset	Correlation Coefficient
Northern hemisphere	Our Product	-0.842
	WOA23	-0.778
	Ito et al., 2024	-0.476
	GOBAI-O2	**-0.852**
Southern hemisphere	Our Product	**-0.935**
	WOA23	-0.756
	Ito et al., 2024	-0.835
	GOBAI-O2	-0.890

Table 1. ***Correlation coefficients between the seasonal anomalies of oxygen and temperature in the upper ocean.*** Correlation coefficients are computed for the upper $100m$ with respect to seasonal anomalies in temperature and oxygen in both the Northern and Southern hemispheres.

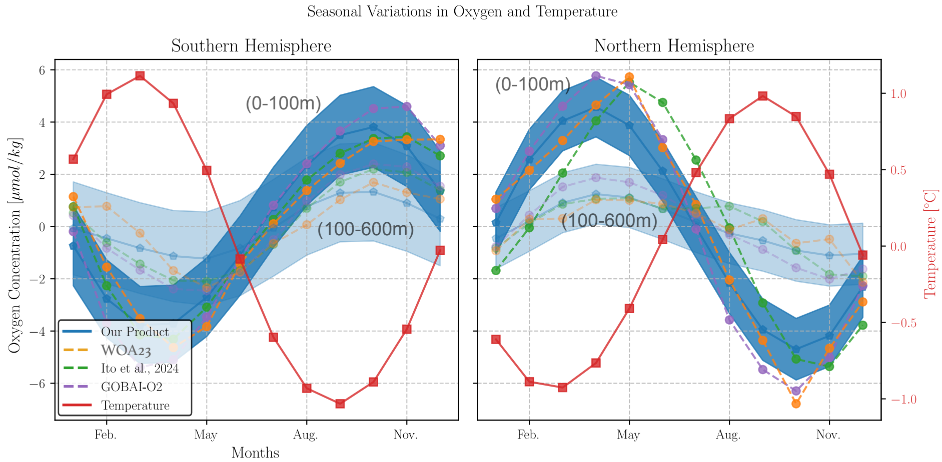

Figure 6. ***Climatological seasonal cycle in dissolved oxygen in our product compared to WOA23 and previous ML-based models.*** The seasonal cycle of dissolved oxygen in our product is shown for the upper 100 m (dark blue) and the 100–600 m subsurface layer (light blue) in the Southern (left) and Northern (right) Hemispheres. Shaded envelopes represent the uncertainty range around the seasonal means. For comparison, the seasonal oxygen cycles from WOA23 (orange) and previously published ML-based reconstructions by Ito et al. (2024) (green) and GOBAI-O2 (purple) are also displayed. Additionally, the seasonal cycle of temperature in the upper 100 m from WOA23 is shown in red.

3.3 Long-term oxygen changes and comparison with previous estimates

We analyze long-term changes in ocean oxygen levels using our product and compare them to estimates from ML-based reconstructions by GOBAI-O2 and Ito et al. (2024) (Figure 7). To ensure a consistent comparison, we use a common ocean volume across all three products, excluding regions absent in previously published ML datasets. These include marginal seas

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

(not included in Ito et al., 2024) and the Arctic (>80°N) and Antarctic (<80°S) oceans (not included in GOBAI-O2). Oxygen content computed based on the entire ocean volume of the proposed product (including marginal seas and polar regions) are depicted in Figure C5 in SI Appendix C.

Our analysis reveals a deoxygenation rate of approximately -438.37 ± 52.19 Tmol/decade in the upper 1000 m over the period 1965–2022, which closely aligns with the -452.79 Tmol/decade reported by Ito et al. (2024) for the same layer and a similar time span. When extending the analysis to the upper 2000 m, the deoxygenation rate increases to -685.53 ± 113.25 Tmol/decade.

Over the past two decades (2004–2022), our product suggests a slower deoxygenation rate, estimated at -296.54 ± 111.37 Tmol/decade in the upper 1000 m and -403.73 ± 193.85 Tmol/decade in the upper 2000 m. These values are more conservative than those from GOBAI-O2, which reports -384.48 Tmol/decade and -720.99 Tmol/decade for the same depth ranges and time period. When including marginal seas and polar regions, our estimated deoxygenation rates for the full 1965–2022 period increase to -582.40 ± 80.63 Tmol/decade in the upper 1200 m and -803.36 ± 147.20 Tmol/decade in the upper 2000 m (Figure C5 in SI Appendix C).

We also compare our deoxygenation estimates with previous results based on traditional mapping techniques. In the upper 2000 m, our estimated deoxygenation rate is comparable to the -960.4 ± 409.1 Tmol/decade reported by Schmidtko et al. (2017) Schmidtko et al. (2017) Schmidtko et al. (2017) for the period 1965–2010. However, when considering only the upper 1200 m, our estimate suggests a significantly faster deoxygenation rate than Schmidtko et al.'s estimate of -257.5 ± 185.1 Tmol/decade over the same period.

Nonetheless, our results exhibit better agreement with Schmidtko et al. (2017) when analyzed over the same time span (1965–2010), yielding a deoxygenation rate of -478.06 ± 125.51 Tmol/decade in the upper 1200 m and -644.12 ± 227.64 Tmol/decade in the upper 2000 m.

Additionally, we compare our results with estimates from Ito (2022), which were derived using optimal interpolation. Our findings indicate a substantially faster deoxygenation trend. Specifically, we estimate a rate of -340.84 ± 151.27 Tmol/decade in the upper 700 m and -763.11 ± 184.74 Tmol/decade in the upper 2000 m, whereas Ito et al. (2022) reported significantly lower rates of -100 Tmol/decade in the upper 700 m and -327 ± 45.99 Tmol/decade in the upper 2000 m. It is well established that the estimates from Ito et al. (2022) tend to be lower compared to other state-of-the-art studies Ito (2022). Recent research has shown that optimal interpolation methods can introduce biases, particularly underestimating deoxygenation in regions with sparse observational data Ito et al. (2024b). In this context, our methodology mitigates such biases by increasing the density of dissolved oxygen data through the incorporation of ML-emulated oxygen estimates.

Finally, it is worth noting that our product reveals much stronger decadal and inter-decadal variability in the rate of deoxygenation compared to previous ML-based reconstructions (Figure 7). The influence of decadal climate variability in modulating regional and global deoxygenation is well established Oschlies et al. (2018). For instance, the rate of deoxygenation in our product from 1980 to the early 2000s was substantially higher than during the 1960s and 1970s, as well as over the past two decades. These variations—largely absent in earlier ML-based reconstructions—are consistent with model-based studies suggesting that major climate variability modes, such as the Pacific Decadal Oscillation (PDO), strongly influence global oxygen content. For

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Open Access Earth System Science Data Discussions

example, Poupon et al. (2023) show that deoxygenation is favored during positive phases of the PDO (such as in the 1980s and 1990s), whereas negative PDO phases—dominant in the 1960s, 1970s, and over the past two decades—enhance oxygenation in the tropical Pacific, thereby weakening climate change–driven deoxygenation.

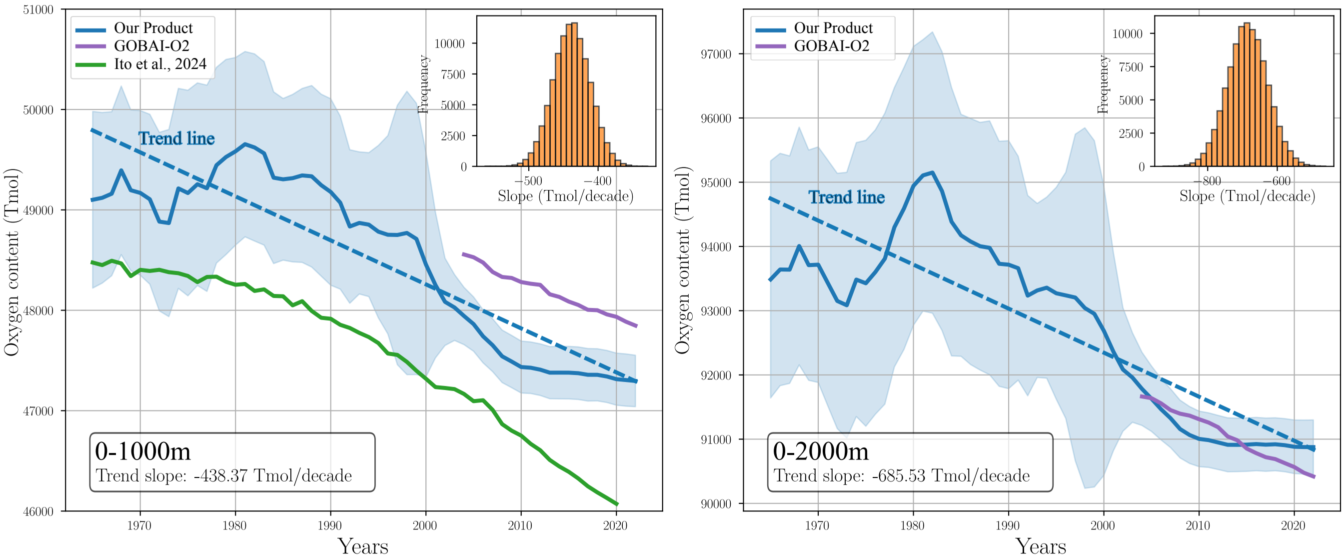

Figure 7. ***Long-term changes in globally integrated oxygen inventory in our product and other ML-based products.*** The left panel shows the inventory for the upper 1000 m, while the right panel displays the inventory for the upper 2000 m. The histograms in each panel represent the empirical distribution of the deoxygenation rate, computed from a Monte Carlo simulation with 10,000 realizations. For comparison, changes in oxygen inventories based on previously published ML-based reconstructions by Ito et al. (2024) (green) and GOBAI-O2 (purple) are also displayed.

3.4 Uncertainty estimates

We analyze the uncertainty fields associated with the proposed gridded product. The uncertainty is quantified by the covariance matrix Σ_a of the reconstructed field. As described in Section 2.4, Σ_a is a diagonal matrix, representing the variance at each grid point. This covariance is sensitive to the density of observational coverage and it decreases as data availability increases.

Our methodology substantially expands the effective observational network by emulating dissolved oxygen (DO) profiles from measured temperature and salinity data (Figure 8a). This enhanced data coverage leads to a significant reduction in the uncertainty of the reconstructed field relative to an optimal interpolation based solely on direct DO observations (Figure 8a). The improvement is particularly evident after 2000, which corresponds to the deployment of the Argo program, which significantly improved the global coverage of T/S measurements used here to emulate oxygen data.

This improvement is further evidenced by comparing the optimal interpolation standard deviation when using only the direct observations versus when including the emulated profiles as well (Figure 8b). The baseline product, based solely on

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

direct observations, exhibits high uncertainty, particularly in data-sparse regions. In contrast, the proposed product—which incorporates both direct and emulated profiles—demonstrates significantly reduced uncertainty due to the enhanced spatial coverage.

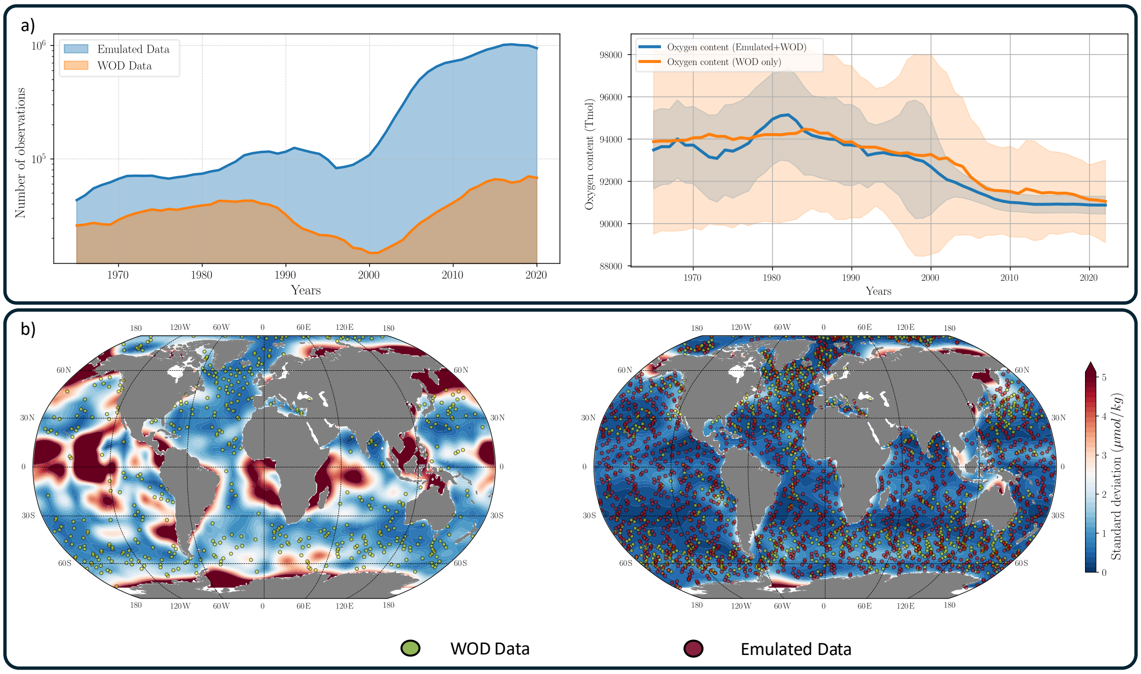

Figure 8. ***Analysis of uncertainty of the gridded oxygen product.*** Panel (a, left) shows the number of direct DO observations in the preprocessed WOD database and the number of emulated profiles at 200 m depth. Panel (a, right) compares the global DO inventory derived from two OI fields: The proposed product (in blue), which takes into account both direct and emulated data, and a baseline field (in orange) that relies exclusively on direct measurements. Panel (b) presents the standard deviation of the OI fields at 200 m depth. The left map corresponds to the OI using only direct observations, while the right includes both direct and emulated profiles. The spatial distribution of direct (green) and emulated (red) observations is displayed, with a subsampling factor of 15 applied to improve visibility.

4 Conclusion and outlook

In this study, we developed a novel gridded dissolved oxygen concentration product by integrating observational oxygen data with emulated estimates derived from temperature and salinity profiles. The proposed methodology offers several key advantages, including:

– **Increased Density of Observations:** The methodology significantly increases the density of the observational coverage, which in turn reduces the uncertainty and biases of the gridded product. The improvement is particularly strong in data-poor regions.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

- **Independent Interpolation of Oxygen Data:** Unlike recent state-of-the-art ML approaches Sharp et al. (2022); Ito et al. (2024a) that derive gridded oxygen products from interpolated temperature and salinity fields, our method directly interpolates observed and emulated oxygen data using optimal interpolation (OI). This approach ensures a more accurate representation of uncertainty in the gridded product and avoids additional sources of error that may arise from relying on pre-interpolated temperature and salinity datasets.

- **Flexible Space-Time Resolution:** The approach provides flexibility in the spatial and temporal resolution of the gridded product, which can be adjusted based on the density of available observations. Specifically, we generate two gridded products: one with a yearly resolution covering the entire study period (1965–2022) and another with a monthly resolution from 2005 to 2022, capitalizing on the denser data coverage during the Argo period. Both products are constructed with a 1°resolution. However, the effective resolution of the monthly product is enhanced by using a smaller e-folding length scale in the OI process.

The resulting product generally agrees with the reference climatology and recent ML-based products in terms of reproducing the spatial variability of dissolved oxygen, and it also aligns with previously published estimates of long-term global deoxygenation, albeit with reduced uncertainty around those estimates. Additionally, the product outperforms current state-of-the-art products by more effectively resolving dissolved oxygen variability at synoptic scales on the order of $O(10^3)$ km and by better capturing temporal variability from seasonal to interdecadal scales.

The proposed methodology can be extended to other biogeochemical variables, including, for example, nutrients, pH, phytoplankton pigments, and particle backscatter. Previous studies Sauzède et al. (2017) have shown that neural network models can predict such variables based on dissolved oxygen measurements. In this context, our study highlights the potential to extend these works by developing biogeochemical emulators based on physical measurements and emulated oxygen data. This would enable the construction of gridded products for these variables with improved spatial and temporal resolutions.

While our results confirm global deoxygenation trends that are consistent with previous studies, further research is needed to investigate regional-scale deoxygenation, particularly in areas where major OMZs are present. In this context, the increased observational density provided by our ML-emulated oxygen data can enhance our ability to monitor and document ongoing changes in major OMZs, while also reducing uncertainties in these estimates.

5 Data availability

The gridded products developed in this study are available as NetCDF files Ouala et al. (2025). The corresponding DOI (10.5281/zenodo.15478088) will be made public upon acceptance of the manuscript.

Author contributions. SO and ZL designed and structured the study. SO and OH developed the model code and generated the gridded product. SO and ZL analyzed the data and prepared the manuscript.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. SO acknowledges the support of the AI4Extremes Chair and the Horizon project AI4PEX (grant agreement 101137682).
300 ZL was supported by funding provided by Tamkeen through grant CG009 to the Mubadala ACCESS Center and funding support from Mubadala Philanthropies under XR016; their support is greatly appreciated. Data processing and machine learning model training was performed at the High Performance Computing (HPC) cluster of NYUAD, Jubail.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

References

Aouni, A. E., Gaudel, Q., Regnier, C., Van Gennip, S., Drevillon, M., Drillet, Y., and Lellouche, J.-M.: GLONET: Mercator's End-to-End Neural Forecasting System, arXiv preprint arXiv:2412.05454, 2024.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, 2023.

Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., Halloran, P., Heinze, C., Ilyina, T., Seferian, R., et al.: Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models, Biogeosciences, 10, 6225–6245, 2013.

Chelton, D. B., Schlax, M. G., Samelson, R. M., and de Szoeke, R. A.: Global observations of large oceanic eddies, Geophysical Research Letters, 34, 2007.

Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J., et al.: Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review, IEEE/CAA Journal of Automatica Sinica, 10, 1361–1387, 2023.

Claustre, H., Johnson, K. S., and Takeshita, Y.: Observing the global ocean with biogeochemical-Argo, Annual review of marine science, 12, 23–48, 2020.

Codispoti, L., Brandes, J. A., Christensen, J., Devol, A., Naqvi, S., Paerl, H. W., and Yoshinari, T.: The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene?, Scientia Marina, 65, 85–105, 2001.

Dheeshjith, S., Subel, A., Adcroft, A., Busecke, J., Fernandez-Granda, C., Gupta, S., and Zanna, L.: Samudra: An AI Global Ocean Emulator for Climate, arXiv preprint arXiv:2412.03795, 2024.

Fourrier, M., Coppola, L., Claustre, H., D'Ortenzio, F., Sauzède, R., and Gattuso, J.-P.: A regional neural network approach to estimate water-column nutrient concentrations and carbonate system variables in the Mediterranean Sea: CANYON-MED, Frontiers in Marine Science, 7, 620, 2020.

Good, S. A., Martin, M. J., and Rayner, N. A.: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, Journal of Geophysical Research: Oceans, 118, 6704–6716, 2013.

Grégoire, M., Garçon, V., Garcia, H., Breitburg, D., Isensee, K., Oschlies, A., Telszewski, M., Barth, A., Bittig, H. C., Carstensen, J., et al.: A global ocean oxygen database and atlas for assessing and predicting deoxygenation and ocean health in the open and coastal ocean, Frontiers in Marine Science, 8, 724 913, 2021.

Gruber, N.: The marine nitrogen cycle: overview and challenges, Nitrogen in the marine environment, 2, 1–50, 2008.

Gruber, N., Doney, S. C., Emerson, S. R., Gilbert, D., Kobayashi, T., Körtzinger, A., Johnson, G. C., Johnson, K. S., Riser, S. C., and Ulloa, O.: Adding Oxygen to Argo: Developing a Global In Situ Observatory for Ocean Deoxygenation and Biogeochemistry., 2010.

Hermes, J., Masumoto, Y., Beal, L., Roxy, M. K., Vialard, J., Andres, M., Annamalai, H., Behera, S., d'Adamo, N., Doi, T., et al.: A sustained ocean observing system in the Indian Ocean for climate related scientific knowledge and societal needs, Frontiers in Marine Science, 6, 355, 2019.

Ito, T.: Optimal interpolation of global dissolved oxygen: 1965–2015, 2022.

Ito, T., Minobe, S., Long, M. C., and Deutsch, C.: Upper ocean O2 trends: 1958–2015, Geophysical Research Letters, 44, 4214–4223, 2017.

Ito, T., Cervania, A., Cross, K., Ainchwar, S., and Delawalla, S.: Mapping dissolved oxygen concentrations by combining shipboard and Argo observations using machine learning algorithms, Journal of Geophysical Research: Machine Learning and Computation, 1, e2024JH000 272, 2024a.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Ito, T., Garcia, H. E., Wang, Z., Minobe, S., Long, M. C., Cebrian, J., Reagan, J., Boyer, T., Paver, C., Bouchard, C., et al.: Underestimation of multi-decadal global O 2 loss due to an optimal interpolation method, Biogeosciences, 21, 747–759, 2024b.

Keeling, R. F., Körtzinger, A., Gruber, N., et al.: Ocean deoxygenation in a warming world, Annu. Rev. Mar. Sci, 2, 199–229, 2010.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv:1412.6980 [cs], http://arxiv.org/abs/1412.6980, arXiv: 1412.6980, 2014.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., et al.: Neural general circulation models for weather and climate, Nature, pp. 1–7, 2024.

Kwiatkowski, L., Torres, O., Bopp, L., Aumont, O., Chamberlain, M., Christian, J. R., Dunne, J. P., Gehlen, M., Ilyina, T., John, J. G., 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.

Lachkar, Z., Cornejo-D'Ottone, M., Singh, A., Arístegui, J., Dewitte, B., Fawcett, S., Garçon, V., Lovecchio, E., Molina, V., and Vinayachandran, P.: Biogeochemistry of greenhouse gases in coastal upwelling systems: Processes and sensitivity to global change, Elementa: Science of the Anthropocene, 12, 2024.

Laffoley, D. D. and Baxter, J.: Ocean Deoxygenation: Everyone's Problem-Causes, Impacts, Consequences and Solutions, IUCN, 2019.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, 2023.

Martin, S. A., Manucharyan, G. E., and Klein, P.: Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies, Journal of Advances in Modeling Earth Systems, 15, e2022MS003 589, 2023.

Martin, S. A., Manucharyan, G. E., and Klein, P.: Deep learning improves global satellite observations of ocean eddy dynamics, Geophysical Research Letters, 51, e2024GL110 059, 2024.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A.: ClimaX: A foundation model for weather and climate, arXiv preprint arXiv:2301.10343, 2023.

Oschlies, A., Brandt, P., Stramma, L., and Schmidtko, S.: Drivers and mechanisms of ocean deoxygenation, Nature geoscience, 11, 467–473, 2018.

Ouala, S., Brunton, S. L., Chapron, B., Pascual, A., Collard, F., Gaultier, L., and Fablet, R.: Bounded nonlinear forecasts of partially observed geophysical systems with physics-constrained deep learning, Physica D: Nonlinear Phenomena, 446, 133 630, 2023.

Ouala, S., Chapron, B., Collard, F., Gaultier, L., and Fablet, R.: Online calibration of deep learning sub-models for hybrid numerical modeling systems, Communications Physics, 7, 402, 2024.

Ouala, S., Hidaoui, O., and Lachkar, Z.: A Novel Global Gridded Ocean Oxygen Product Derived from a Neural Network Emulator and In-situ Observations, https://doi.org/10.5281/zenodo.15478088, to be made public upon manuscript acceptance, 2025.

Poupon, M. A., Resplandy, L., Lévy, M., and Bopp, L.: Pacific decadal oscillation influences tropical oxygen minimum zone extent and obscures anthropogenic changes, Geophysical Research Letters, 50, e2022GL102 123, 2023.

Rabalais, N. N., Turner, R. E., and Wiseman Jr, W. J.: Gulf of Mexico hypoxia, aka "The dead zone", Annual Review of ecology and Systematics, 33, 235–263, 2002.

Roemmich, D. and Gilson, J.: The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program, Progress in oceanography, 82, 81–100, 2009.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Sauzède, R., Bittig, H. C., Claustre, H., Pasqueron de Fommervault, O., Gattuso, J.-P., Legendre, L., and Johnson, K. S.: Estimates of water-column nutrient concentrations and carbonate system parameters in the global ocean: A novel approach based on neural networks, Frontiers in Marine Science, 4, 128, 2017.

Schmidtko, S., Stramma, L., and Visbeck, M.: Decline in global oceanic oxygen content during the past five decades, Nature, 542, 335–339, 2017.

Sharp, J. D., Fassbender, A. J., Carter, B. R., Johnson, G. C., Schultz, C., and Dunne, J. P.: GOBAI-O 2: temporally and spatially resolved fields of ocean interior dissolved oxygen over nearly two decades, Earth System Science Data Discussions, 2022, 1–46, 2022.

Vaquer-Sunyer, R. and Duarte, C. M.: Thresholds of hypoxia for marine biodiversity, Proceedings of the National Academy of Sciences, 105, 15 452–15 457, 2008.

Wang, X., Wang, R., Hu, N., Wang, P., Huo, P., Wang, G., Wang, H., Wang, S., Zhu, J., Xu, J., et al.: Xihe: A data-driven model for global ocean eddy-resolving forecasting, arXiv preprint arXiv:2402.02995, 2024.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Appendix A: Parameterization and training of the ML model

The machine learning model used to emulate dissolved oxygen concentrations is a fully connected neural network with six layers. Its architecture includes an input layer with 10 features (detailed hereafter), four hidden layers with 256, 128, 128, 64, and 32 neurons, respectively, and an output layer with a single neuron.

The input features include water temperature (*T*), salinity (*S*), latitude (LAT), longitude (LON), month of the year (MOY), day of the month (DOY), and calendar year (Y). To capture the periodicity of LON and MOY, a sinusoidal encoding is applied:

$$\mathbf{MOY}_s = \begin{bmatrix} \cos\left(\frac{MOY \times \pi}{6}\right) \\ \sin\left(\frac{MOY \times \pi}{6}\right) \end{bmatrix} \quad ; \quad \mathbf{Lon}_s = \begin{bmatrix} \cos\left(\frac{LON \times \pi}{180}\right) \\ \sin\left(\frac{LON \times \pi}{180}\right) \end{bmatrix} \tag{A1}$$

where the subscript s refers to the sinusoidal encoding.

The hidden layers utilize the rectified linear unit activation function (*relu*). The model is trained using the mean squared error (MSE) loss function and is optimized with the Adam optimizer Kingma and Ba (2014).

Appendix B: Data sources

This section outlines the data sources used in this study. The data are organized into two distinct databases. The first database, presented in Table B1, includes data sources and instruments that provide collocated measurements of temperature, salinity, and dissolved oxygen. This dataset is utilized for training, validation, and testing of the neural network emulator. Oxygen data from this dataset are also used in the optimal interpolation.

WOD dataset	**Instrument Type**
OSD	Bottle, low-resolution CTD, low-resolution XCTD, plankton data
CTD	High-resolution CTD and high-resolution XCTD data
PFL	Profiling float data, mainly from the Argo program
DRB	Drifting buoy data from surface drifting buoys with thermistor chains and ice-tethered profilers
UOR	Undulating Oceanographic Recorder data from a CTD probe on a towed vehicle
GLD	Glider data
MRB	Moored buoy data, mainly from the Equatorial buoy arrays (TAO)

Table B1. ***Data sources used to construct the training, validation and testing dataset of the ML-based emulator.*** Data access: 6/12/2023.

The second database, presented in Table B2, consists of only temperature and salinity data. This dataset is used to generate emulated oxygen data using the trained ML emulator.

Appendix C: Additional figures

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

WOD dataset	Instrument Type
OSD	Bottle, low-resolution CTD, low-resolution XCTD, plankton data
CTD	High-resolution CTD and high-resolution XCTD data
PFL	Profiling float data, mainly from the Argo program
DRB	Drifting buoy data from surface drifting buoys with thermistor chains and ice-tethered profilers
UOR	Undulating Oceanographic Recorder data from a CTD probe on a towed vehicle
GLD	Glider data
MRB	Moored buoy data, mainly from the Equatorial buoy arrays (TAO)
SUR	Surface-only data (bucket, thermosalinograph)
APB	Autonomous Pinniped Bathythermograph - TT-D recorders and CTDs attached to elephant seals

Table B2. ***Data sources used to construct the temperature and salinity dataset used to generate the emulated oxygen data.*** Data access: 15/5/2024.

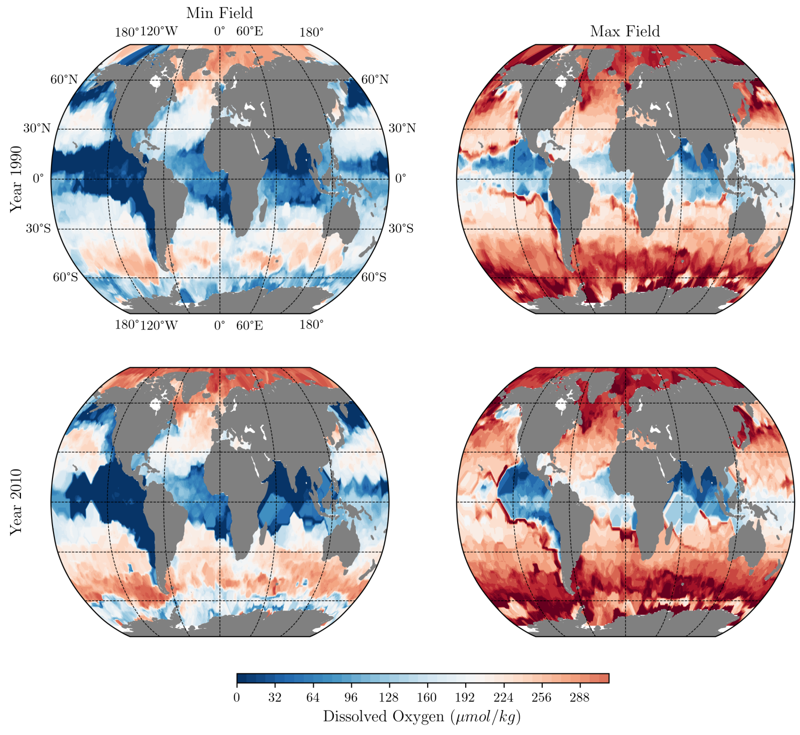

Figure C1. ***Minimum and Maximum fields used in the QC of the emulated profiles.*** First row, Minimum and Maximum fields at 200m depth, computed as described in section 2.3 for the year 1990. Second row, Same fields for the year 2010.

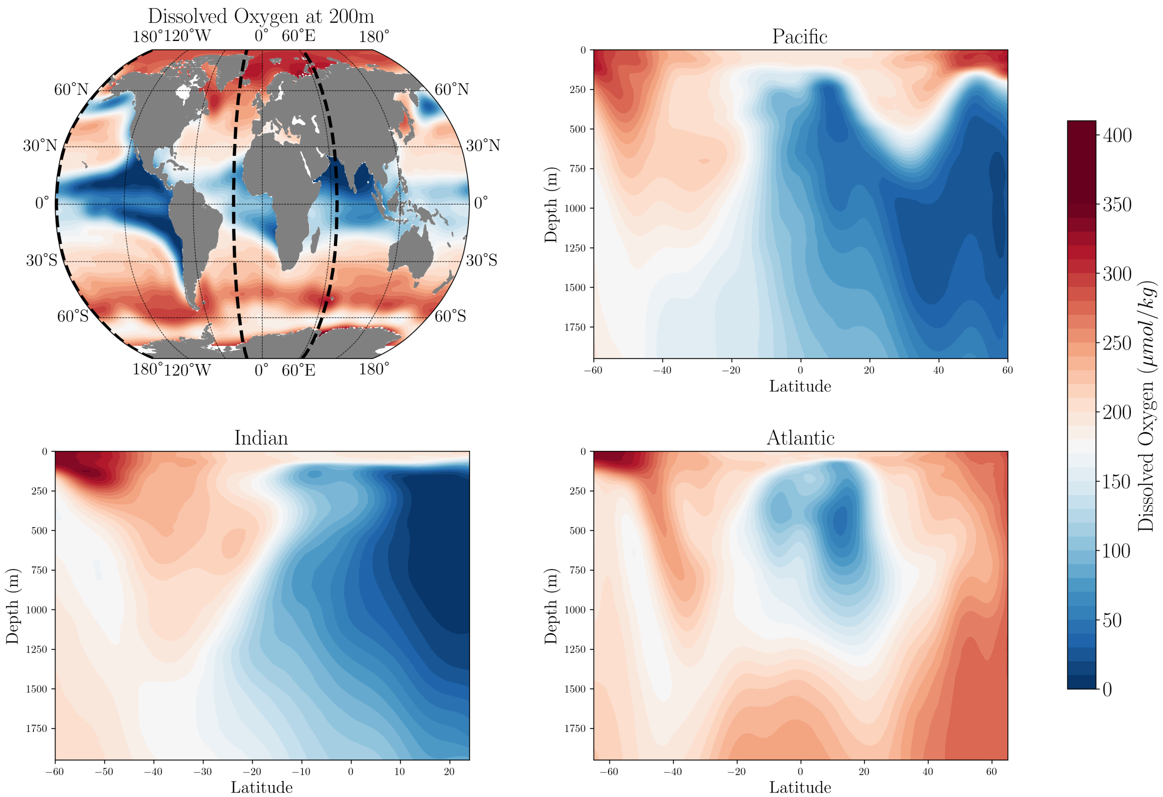

Figure C2. ***Spatial distribution of dissolved oxygen concentration of the proposed gridded product.*** The upper right figure shows the long-term mean of dissolved oxygen concentration at 200m. Dashed black lines represent the locations of the Meridional cross sections from the surface to 2000m.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

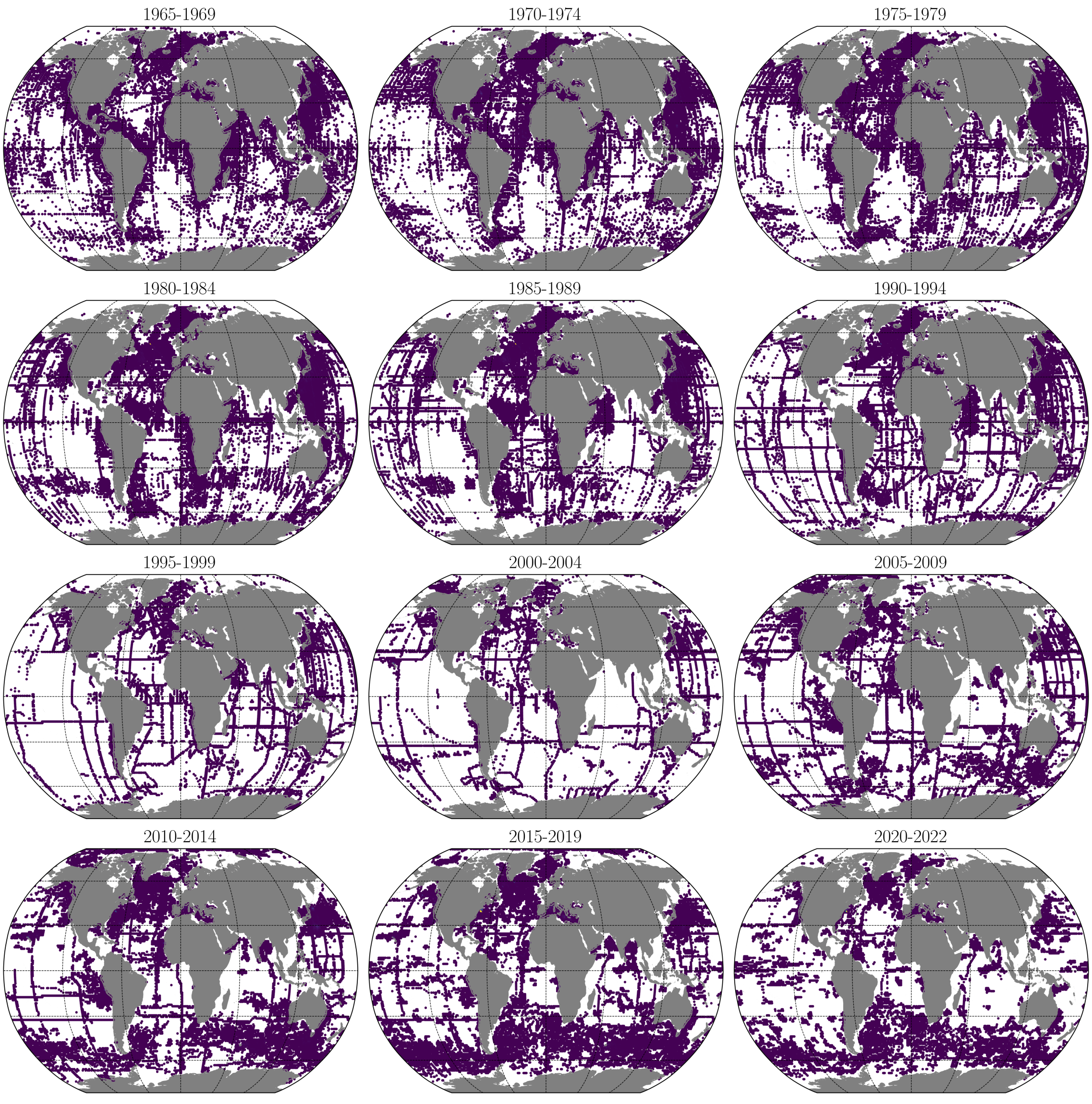

Figure C3. ***Oxygen data coverage since 1965 at 200m depth.*** Dark blue indicates $1^\circ \times 1^\circ$grid locations with oxygen measurements at 5-year data intervals.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

Open Access Earth System Science Data Discussions

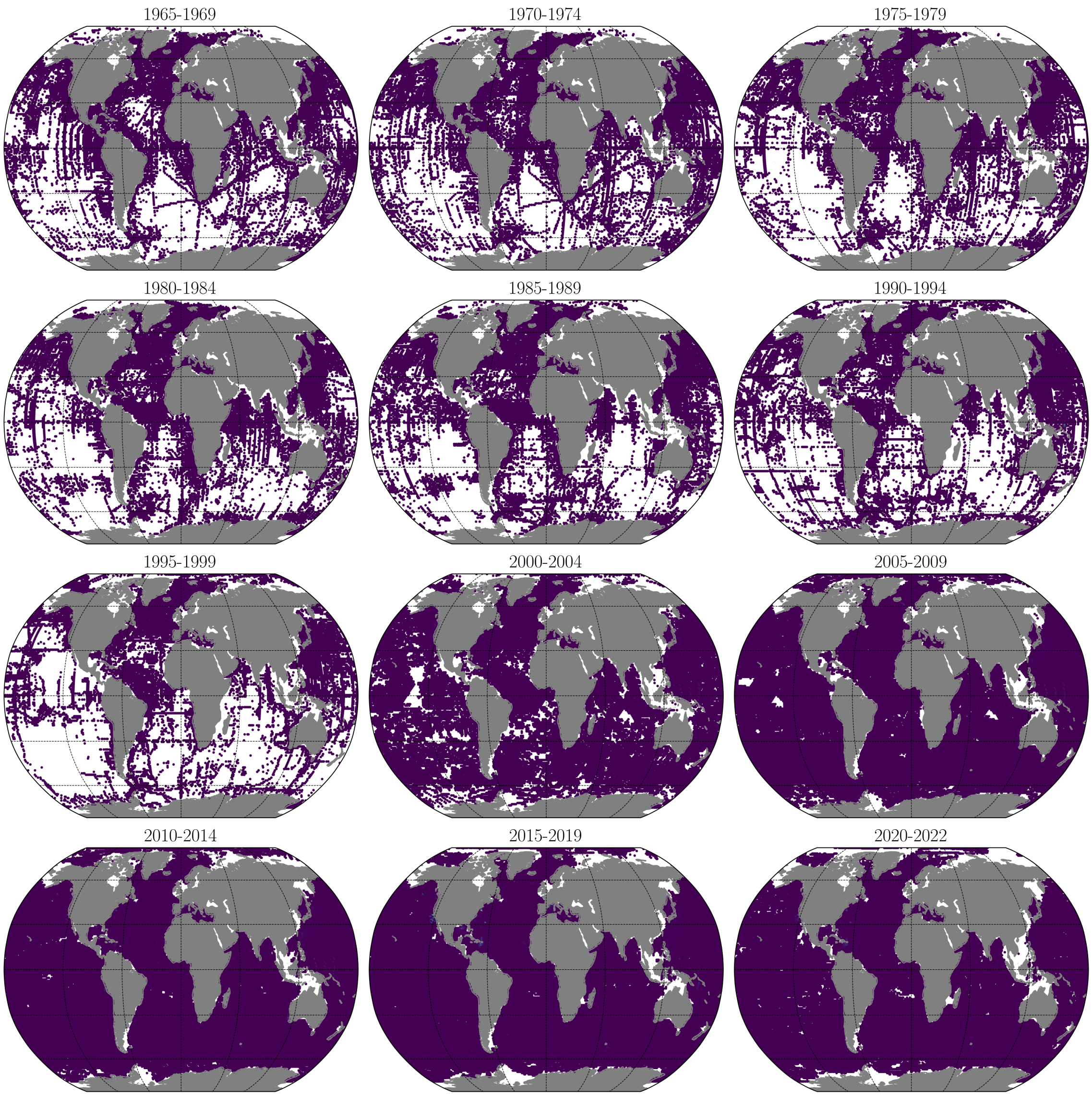

Figure C4. ***Emulated oxygen data coverage since 1965 at 200m depth.*** Dark blue indicates 1°× 1°grid locations with emulated oxygen measurements at 5-year intervals. These locations essentially correspond to the coverage of temperature and salinity measurements, excluding the emulated profiles flagged as unrealistic by the QC procedure described in Section 2.3.

https://doi.org/10.5194/essd-2025-288
Preprint. Discussion started: 16 June 2025

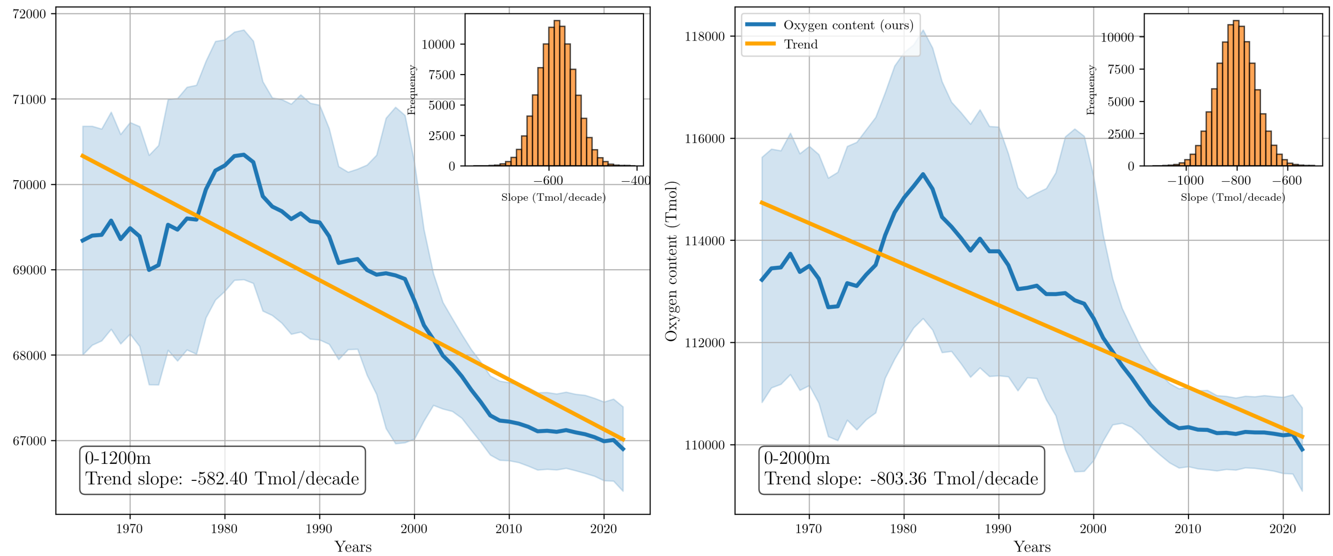

Figure C5. ***Long-term changes in globally integrated oxygen inventory.*** The left panel shows the inventory for the upper 1200 m, while the right panel displays the inventory for the upper 2000 m. The histograms in each panel represent the empirical distribution of the deoxygenation rate, computed from a Monte Carlo simulation with 10,000 realizations.