Title: Mapping dissolved oxygen concentrations by combining shipboard and Argo observations using machine learning algorithms

Authors: Takamitsu Ito(1), Ahron Cervania(1), Kaylin Cross(2), Sanika Ainchwar(3) and Sara Delawalla(1)

Affiliation: (1) School of Earth and Atmospheric Sciences, Georgia Institute of Technology
(2) School of Civil and Environmental Engineering, Georgia Institute of Technology
(3) College of Computing, Georgia Institute of Technology

Corresponding author email: taka.ito@eas.gatech.edu

Abstract: The ocean oxygen (O_2) inventory has declined in recent decades but the estimates of O_2 trend are uncertain due to its sparse and irregular sampling. A refined estimate of deoxygenation rate is developed using machine learning techniques and biogeochemical Argo array. The source data includes historical shipboard (bottle and CTD-O_2) profiles from 1965 to 2020 and biogeochemical Argo profiles after 2005. Neural network and random forest algorithms were trained using approximately 80% of this data and the remaining 20% for validation. The training data is further divided into 5-fold decadal groups to perform cross validation and hyperparameter tuning. Through different combinations of algorithm types and predictor variable sets, an ensemble of gridded monthly O_2 datasets was generated with similar skills (root-mean-square error ~ 13-18 μmol/kg and R^2 ~ 0.9). The largest errors are found in the oxycline and frontal regions with strong lateral and vertical gradients. The mapping was repeated with shipboard data only and with both shipboard and Argo data. The effect of including Argo data on the estimated global deoxygenation trends has a major impact with an 56% increase while reducing the uncertainty by 40% as measured by the ensemble spread. This study

demonstrates the importance of new biogeochemical Argo arrays in relatively data-poor regions such as the Southern Ocean.

Plain language summary

Oxygen is an essential molecule existing in the seawater. Its concentrations are declining in many parts of the oceans. The causes of the decline are not fully understood but it is thought to be linked to the recent warming of the surface ocean and its impact on the physics and chemistry of the seawater. It is difficult to accurately estimate how much oxygen has been lost from the oceans based on historical measurements because of sparse sampling density and irregular timing of measurements. This study assesses the skill of machine-learning based estimates of oxygen in the global oceans, with the specific aim of synthesizing historical ship-based measurements and new autonomous data from robotic floats. By combining these data, we were able to determine the rate of oxygen loss at finer temporal and spatial regions. Our results show that including float data substantially increases the estimate of oxygen loss while reducing its uncertainty.

Key points

- A new ensemble dataset of oxygen is developed based on observations and machine learning algorithms.
- The newly developed dataset is broadly consistent with established climatology and with deoxygenation rates from other independent studies.
- Synthesis of shipboard and Argo-oxygen data increased estimated deoxygenation rates by 56% while reducing the uncertainties by 40%.

1. Introduction

Historical observations from past decades have shown growing influences of anthropogenic perturbations on marine ecosystems and biogeochemistry (Friedland et al., 2020; Gruber et al., 2021; Pershing et al., 2015; Seidov et al., 2018). Dissolved oxygen is essential for all multicellular life and its concentration can shape the habitats of marine organisms, cycling of nutrients and trace elements, and the redox states of the seawater. There is a growing consensus in the scientific community that the global ocean O_2 inventory has declined in recent decades. Estimates of the oceanic oxygen inventory decline are in the range of 0.5-3.3% over the period of 1970-2010, equivalent of $-0.48 \pm 0.35\%$ per decade, for the upper 1,000m (Bindoff et al., 2019 and references therein). Assessing the global and regional O_2 inventories requires filling data gaps because the historical O_2 measurements are irregular in time and sparse in space. The wide range in the estimates of ocean deoxygenation can be due to the different interpolation methods, different data quality control standards, and different data sources.

There are three major groups of O_2 observations including two types of shipboard measurements and biogeochemical Argo floats. First, bottle O_2 profiles are typically measured by modified Winkler titration method with a precision of about 1 μmol/kg. Most modern oxygen chemical titration measurements are based on Carpenter's whole bottle titration method and an amperometric or photometric end-point detection with a precision of about 0.5-1 μmol/kg (Carpenter, 1965). Older bottle data prior to 1965 may have larger measurement uncertainties. Secondly, Conductivity-Temperature-Depth (CTD) instruments have been equipped with O_2 sensors since the late 1980s, and they are periodically calibrated to the bottle data (Culberson et al., 1991).

Argo is an international program that measures seawater temperature and salinity using a fleet of robotic instruments that drift with the ocean currents and periodically sample the water column by moving up to the surface, with a typical depth and cycle time of 2000m and 10 days (Roemmich et al., 2019). Biogeochemical-Argo (BGC-Argo) aims to develop the global network of biogeochemical sensors mounted on Argo floats including O_2, NO_3, pH and bio-optical properties (Bittig et al., 2019; Johnson et al., 2013; Sarmiento et al., 2023). Chemical sensors for measuring biogeochemical data require post-deployment quality control and calibration (Maurer et al., 2021). There are real-time, real-time adjusted and delayed mode data. In-situ calibration using atmospheric reanalysis/in-air measurement and empirical algorithms can bring accuracy to within 3 μmol/kg for O_2.

Calculations of basin-scale O_2 inventory requires statistical gap-filling methods to estimate O_2 for the location and time where direct measurements are not available. Such gap-filling techniques include objective analysis such as the multi-pass Barnes method (Barnes, 1964) and optimal interpolation or kriging (Wunsch, 1996). Irregular and uneven distribution of observational data are known to cause increased uncertainties and underestimation of trends in the data-poor regions (Ito et al., 2023). Recently, machine learning (ML) has become a powerful tool in climate and ocean sciences (Chen et al., 2019; Gloege et al., 2021; Reichstein et al., 2019). In marine biogeochemistry, ML has been used to generate the maps of partial pressure of carbon dioxide (Chen et al., 2019; Gloege et al., 2021; Landschützer et al., 2013; Moussa et al., 2016; Sharp et al., 2022; Zeng et al., 2015), oxygen (Sharp et al., 2023), alkalinity (Broullón et al., 2019), dissolved iron (Huang et al., 2022), phytoplankton concentrations (Chen et al., 2020) and nutrients (Sauzède et al., 2017). Typically, data gaps are filled by some form of nonlinear regression models trained by available observational data. The underlying assumption is that

there are significant, regional relationships between biogeochemical variables and other input data such as temperature, salinity, pressure and/or geographic coordinates. With a large amount of training data, ML algorithms can learn detailed relationships from existing observations. Once the algorithm is trained and validated, it can be used to reconstruct gridded biogeochemical fields. Sharp et al., (2023) recently developed gridded maps of global O_2 distribution from 2004 to 2022 using two ML approaches including two-layer Neural Network (NN) and Random Forest (RF) regression models. They found a global deoxygenation trend of -0.82 ± 0.11 % per decade from 2004 to 2022 based on the machine learning technique and Argo-O_2 and GLODAP observational datasets. This estimate is larger than that assessed by Bindoff et al. (2019) of -0.48 ± 0.35 % per decade over a different period (1970 to 2010) but these estimates overlap within the uncertainties.

Since the mid-2000s, a significant number of O_2 profiles are measured by biogeochemical Argo floats and its share is increasing. The calibration of Argo-O_2 data is still under development, especially for the response time of optode sensors in the upper ocean oxycline (Bittig & Körtzinger, 2017). Despite these potential biases and uncertainties, there can be significant advantage gained by including the quality-controlled Argo-O_2 data to better estimate the O_2 inventory by combining it with historical shipboard observations. The objective of this study is two-fold. First, we aim to develop four-dimensional (3-dimensional space and time) reconstructions of gridded O_2 datasets using multiple ML approaches. This work is different from Sharp et al. (2023) who focused on the Argo O_2 profiles after 2004. This study covers a significantly longer period from January 1965 using the combination of Argo-O_2 and historical shipboard observations. This study documents the development of the ML based O_2 mapping, leading to the formation of an ensemble of O_2 reconstructions selected from a large

number of trained algorithms with different input variable sets and ML parameters. Secondly, we aim to quantify the impact of including Argo-O_2 data. Separate sets of ML-based O_2 ensembles are formed based on the shipboard data only (without Argo) and with the shipboard and Argo-O_2 data. The comparison of deoxygenation trends and the ensemble spread quantifies the potential impacts on the estimates of deoxygenation trends.

2. Methods

This methods section first describes the data sources for dissolved oxygen and other input variables in section 2.1. We then provide the description of the machine learning approaches in section 2.2 followed by the experimental design and workflow in section 2.3.

2.1 Data Sources

Figure 1 shows the distribution of shipboard and Argo-O_2 measurements based on World Ocean Database 2018 (WOD18, Boyer et al., 2018) for the period of January 1965 to December 2020, downloaded in October 2023. The displayed profile count includes those profiles that passed the quality control step as discussed below. WOD18 is an international collaboration among national data centers, oceanographic research institutions and investigators to provide a comprehensive dataset of quality-controlled oceanographic variables. The preprocessing of the data includes data quality checks indicated by the WOD18 quality control (QC) flags of 0 through 9. This study uses the accepted values (QCflag = 0) only. Data with the QC flags of 1 through 9 are not used in this study as they are outliers or questionable data with several different criteria. The number of profiles taken each year/month fluctuates significantly. Prior to 1990, most O_2 profiles are taken by ship-based bottle measurements. After the 1990s, the number of

CTD-O_2 profiles increased and became a major O_2 data source. Since the mid-2000s, the number of Argo-O_2 profiles has steadily increased. Argo-O_2 data is obtained from the Argo Global Data Assembly Center (GDAC) including the time, location, quality control flags, and descriptions of calibration methods for each O_2 sensor. The entire archive of BGC Argo floats was downloaded in October, 2023. We specifically searched for floats containing delayed-mode O_2 profiles using two standard methods of bias correction including in-air pO_2 measurement with atmospheric reanalysis data (Bushinsky & Emerson, 2015; Johnson et al., 2015) and climatological air-sea disequilibrium of surface O_2 (Takeshita et al., 2013). There are 1,366 BGC-Argo floats that satisfy this condition globally, and from these floats, O_2 data points with acceptable QC flag indicated as QC flag of 1, 2 or 8 are selected (1="good data", 2="probably good data" and 8="estimated data"). BGC-Argo and its calibration methods are still evolving. Sensor calibration bias for Argo-O_2 observations can also include finite response time of optode sensors, which may cause systemic bias in the oxycline regions (Bittig et al., 2014; 2018).

(Figure 1 here)

Figure 1. Sampling density (A,B) Logarithm (base 10) of the cumulative profile count within each 1°x1° longitude-latitude cell for oxygen (O_2) based on the World Ocean Database 2018 (Boyer et al., 2018) downloaded in October 2023. The color saturates at 2 (more than 100 profiles) per cell since 1965. (C) The cumulative profile count for the BGC-Argo O_2 data. These profile counts only includes the profiles that passed the quality control step.

Including all three platforms, approximately one million (963,412) quality-controlled O_2 profiles are used in this study with 69 % bottle, 17 % CTD-O_2 and 14 % Argo-O_2 measurements from 1965 to 2020. **Figure 2** shows the year-by-year temporal evolution of the profile counts from the three sources globally and in five open-ocean basins including the Atlantic, Pacific, Indian, Southern and Arctic Ocean. There are profiles taken in the marginal seas and coastal waters, that are not included in this basin-scale breakdown.

(Figure 2 here)

Figure 2. Yearly evolution of the O_2 profile count. The number of quality-controlled profiles are displayed as a function of time as "stacked" bar chart where Bottle profile count (blue) is placed at the bottom, upon which CTD profile count (orange) is placed. Argo-O_2 profile count (green) is placed on the top without overlap. The vertical axis is in the units of thousands of profiles per year. The definition of ocean basin is taken from the basin mask of the World Ocean Atlas 2018 (Garcia et al., 2018).

Globally, there are 10-20k O_2 profiles per year, but their measurement platforms have evolved over time. Bottle data dominated during the earlier periods but it has been declining after 1990. The decline of bottle profile count was partially compensated by the increase of CTD-O_2 profiles after 1990s and then Argo-O_2 profiles after 2010s. The Atlantic and Pacific Oceans have the largest number of profiles (213k and 253k respectively), exhibiting similar evolution as the global data. The profile counts in the Indian (55k) and Southern Ocean (43k) significantly

increased in the last decade owing to the Argo-O_2 profiles. In contrast, Arctic profiles (46k) mainly come from before 1990s and are highly skewed towards the Atlantic sector.

As a part of pre-processing, the original WOD18 standard-depth profiles with 102 depth levels are placed into monthly bins which are 1°x1° longitude-latitude grid cells. We focus on the upper 47 levels for 0-1,000 m of the water column. Argo-O_2 data is interpolated onto the same standard depths, and placed into the 1°x1° longitude-latitude grid cells. The binning was performed separately for shipboard and Argo-O_2 data, allowing them to be to mapped them together or separately. We focus on the five major ocean basins including Atlantic, Pacific, Indian, Southern and Arctic Oceans according to the definition of ocean basins taken from the World Ocean Atlas 2018 (WOA18; Garcia et al., 2018). The basin boundaries are shown in **Figure 3**. The definition of the Indian Ocean includes the Bay of Bengal.

(Figure 3 here)

Figure 3. Basin definition. The five major basins are filled with different color. This definition is taken from the basin mask of WOA18. Each basin is assigned a number. Here, we use Atlantic (1), Pacific (2), Indian (3, including Bay of Bengal of 56), Southern (10) and Arctic (11) basins.

The target analysis period is after 1965 when the modern oxygen titration method is established by Carpenter (1965). Prior to 1987, only the bottle O_2 data is selected for the shipboard profiles due to the concern that very early CTD-O_2 data may contain larger uncertainties. After 1987, the bottle and CTD-O_2 profiles are averaged within the 1°x1° bins weighted by the profile counts.

206

207 **2.2 Machine learning algorithms**

208 This study aims to extract regional relationships that allow filling data gaps in O_2 using
209 surrogate (predictor) variables such as temperature (T), salinity (S), and pressure using machine
210 learning approaches. As a basis for the surrogate variables, optimally interpolated monthly
211 gridded T/S fields are obtained from the Hadley Centre EN version 4 dataset (hereafter, EN4,
212 Good et al., 2013). It is a global gridded dataset from 1900 to present at the horizontal resolution
213 of 1°x1° in longitude-latitude grid and with 42 vertical depth levels (20 levels within the 0-
214 1,000m). In supervised learning, a computer program is designed to learn the relationship
215 between a large number of paired input-output examples. The output (predictand) variable is the
216 O_2 concentration, and the input (predictor) variable can include physical variables and
217 coordinates. The potential predictor variables include absolute salinity, conservative temperature,
218 pressure, potential density, Brunt-Väisälä frequency, longitude, latitude, time, and month. Some
219 of these variables are coordinates and others are derived from the EN4 dataset. Using the
220 Thermodynamic Equation of State 2010 (TEOS-10), conservative temperature (Θ) and absolute
221 salinity (Sa) are calculated. Potential density is a non-linear function of Θ, Sa and pressure
222 (depth) and is calculated following TEOS-10. Tracer transport in the interior ocean is primarily
223 oriented along the potential density surfaces. While it can be computed from Θ, Sa, and pressure,
224 including potential density may improve the machine learning algorithm. Brunt-Väisälä
225 frequency measures the local stratification, determined from the vertical density gradient. Since
226 stratification can be linked to turbulent mixing, Brunt-Väisälä frequency may potentially
227 improve the algorithm. Having said this, however, it is not clear whether including all above
228 variables will improve the estimation of O_2. The performance may depend on various factors

including the choice of input variables and specific configuration of algorithms. Gregor et al. (2019) showed biases and discrepancies between different methods to gap-fill pCO_2 data in regions where training data is sparse. Applications of ML to ocean biogeochemistry often struggles in data-sparse areas, and care must be taken to choose the algorithms that are best fit to the specific problem (Brunton & Kutz, 2019).

Artificial neural networks and random forest regression are commonly used algorithms for supervised learning, but they have distinct characteristics and operate in different ways. Neural Networks (hereafter, NN) are composed of interconnected nodes (neurons) arranged in layers including input, hidden, and output layers (LeCun et al., 1998). NN is capable of representing complex, nonlinear relationships and can capture intricate patterns, but it requires a large amount of training data. In contrast, Random Forest (hereafter, RF) is an ensemble learning method that combines multiple decision trees to make predictions (Ho, 1995; Kleinberg, 2000). RF can capture complex relationships, but it may struggle with very subtle patterns. RF can handle missing data effectively by using surrogate splits, which means it may outperform NN in data-poor regions. In addition, RF can provide feature importances which can help interpret the results.

In this study, we will employ the Scikit-Learn version 1.3 (Pedregosa et al., 2011) for the Python implementation of NN and RF regression models. For each type of algorithm, there are several free parameters (hyperparameters) that cannot be learned from the data and must be selected before training. These parameters govern the learning process and influence how the model learns the relationship between the predictor and predictand variables. In practice, it's hard to know in advance which algorithm/hyperparameter set works better for a particular problem, and it requires testing multiple algorithms to make a good model choice by experimentation.

Examples of hyperparameters include the number of nodes for each hidden layer in neural networks, the regularization parameter, learning rate, or maximum features in RF. Hyperparameter tuning involves selecting the best combinations of these settings to achieve the best performance.

2.2.1 Train-test split by randomly selecting data from specific years

In oceanographic data, observations always contain some level of noise, which can come from sensor accuracy and sampling uncertainty (spurious noise) as well as unexplained natural variability. Overfitting occurs when an algorithm fits the noises in the training data rather than capturing the signal, and as a result, it negatively impacts its ability to generalize to new, unseen data. Overfitting could occur when a model is too complex relative to the size of the training data and the noise level. In this study, we employ two types of strategies that the algorithms are not overfit by evaluating their ability to generalize unseen data. First, approximately 80% of the observed O_2 profiles are selected to train the algorithms, and the remaining 20% are withheld as to measure how well the trained algorithms can reconstruct the profiles that are not used during the training. Since oceanographic data is correlated in time and space, O_2 measurements from similar region and time should not be shared between the training and test data. During the pre-processing, O_2 profiles within 1°x1° grid cell and within the same month are averaged into a single bin. This reduces the possibility of having similar set of values between the train and test data. The 80-20 split is implemented by randomly selecting 11 years out of the 55-year input data (1965-2020), such that the performance of algorithms are measured by their ability to reconstruct the 11 years of data unused in the training of the algorithm. The selections of the test

data are randomized for each combination of input data set, algorithm type and hyperparameter set.

2.2.2 Decadal group K-fold Cross Validation

The second level of protection against overfitting is the Decadal-group K-fold Cross Validation (hereafter, DKCV), which is a resampling procedure that helps estimating how well an algorithm will perform on unseen data. **Figure 4** visually illustrates this procedure as we apply DKCV for hyperparameter tuning. The training data (~80% of oxygen profiles) are split into K decadal groups (K=5 in this study) and each set of hyperparameters is trained K times using different (K-1) groups of training data, and its performance is validated by measuring how well the trained algorithm reconstructs the one 10-year group that is withheld from the training in terms of R^2 score and root mean square error. Considering the long memory of the ocean properties, the K groups are defined by the decades including 1965-1974, 1975-1984, 1985-1994, 1995-2004 and 2005-2020. The last segment is a 15-year long period for practically covering the entire Argo-O_2 data. In this procedure, a decade of data are dropped from the training set, avoiding the overlap within a 10-year period between the training and validation. This procedure is repeated for all possible combinations of the hyperparameter set in consideration, allowing to select the best configuration while minimizing the possible occurrence of overfitting.

(Figure 4 here)

Figure 4. Decadal group K-fold Cross Validation (DKCV). Training data is divided into K (=5) groups as illustrated by different horizontal bars, which is a suitable approach since autocorrelation is suspected in oceanographic measurements. This ensures that data from the same decade are not shared between training and testing of the algorithms.

2.3 Workflow and experimental design

Building on the procedures discussed in Section 2.2, a workflow is developed for a suite of ML algorithms for predicting the O_2 distribution. **Table 1** organizes different combinations of input/output variables as experiments (Exp) 1 through 6. All experiments use shipboard O_2 as the predictand variable, and Argo-O_2 is also included in Exp 4 through 6. All experiments also include conservative temperature (Θ), absolute salinity (Sa), longitude, latitude, pressure (P or depth), and time as predictor variables. Time is counted as the number of months since January 1965. We also include the sine and cosine of the month of year (mon) to capture annual cycle with 12-month periodicity as $\cos(\pi\, mon/6)$ and $\sin(\pi\, mon/6)$. Exp 2 and 5 includes potential density (σ_θ) and Exp3 and 6 additionally include the strength of stratification as the square of Brunt-Väisälä frequency (N^2) which is proportional to the vertical density gradient. There are some redundancies in the predictor variables where time can include month, and σ_θ and N^2 can be calculated as non-linear functions of T and S. However, these factors are explicitly included because of their predictive potential. The seasonal cycle can be important for O_2 especially in the

near-surface layer for biological O_2 production. Isopycnal surfaces and water column stratification can be important indicators of O_2 ventilation and transport. Comparing Exp 1-3 versus 4-6 can inform the importance of including the Argo-O_2 data.

	Θ	Sa	long	lat	time	P	mon	σ_θ	N^2	Argo
Exp 1										
Exp 2										
Exp 3										
Exp 4										
Exp 5										
Exp 6										

Table 1. Input variables. Experiments highlighted green contain only shipboard O2, while experiments highlighted red contain both shipboard and Argo-O_2. "Θ" is conservative temperature (°C). "Sa" is absolute salinity (g/kg). "long" is longitude and "lat" is latitude, both in degrees except for Southern and Arctic Ocean where the polar stereographic coordinates are used. "P" is pressure (dbar). "σ_θ" is potential density (kg/m^3), and "N^2" is the square of Brunt-Väisälä frequency (s^{-1}). "time" is measured as the number of month since January 1965. "mon" is the month of year implemented as sine and cosine functions.

Two types of algorithms, NN and RF are trained for each experiment (Exp1-6). For each algorithm, a suite of hyperparameters sets is considered (18 sets each for NN and RF), thus a total of 216 algorithms are trained for different combinations of algorithm type, hyperparameter sets, and input/output parameter choices. This was applied separately for each of the 5 basins (Atlantic, Pacific, Indian, Southern and Arctic), leading to the total of 1080 basin-scale algorithms. For NN, the number of nodes in hidden layers and the regularization parameter are systematically changed (see **Supplementary Table 1**). A wide range of hidden layers are

considered including 10-10-10-10, 20-20-20-20, 40-40-40-40, 60-60-60-60, 60-40-20-10, 20-20-20-20-20-10-5, and three different regularization parameters are considered including 0.001, 0.01 and 0.1. Increasing the number of nodes and layers allows more complexity whereas increasing the regularization parameter prevents the model from becoming too complex. Regularization parameter is a coefficient multiplying the sum of squared weights, which is added to the loss function. A larger regularization parameter tends to regulate the magnitude of weights. The combination of hyperparameters results in 18 different configurations of the NN algorithm. The NN algorithm used in this study trains using backpropagation with no activation function in the output layer. It uses the square error as the loss function, and the output is a set of continuous values. Weights are randomized at the initialization.

For RF, the reference hyperparameters are taken from Probst et al., (2019), and different configurations are explored for two variables (see **Supplementary Table 2**) including the minimum samples for split (min_samples_split) and the maximum features (max_features). The value of min_samples_split is varied over a wide range from 2 to 64. If the number of samples in a node is less than the specified "min_samples_split," the node will not be split, and it will become a leaf node, effectively halting the tree's growth in that branch. Thus, increasing the value of the min_samples_split prevents the model to become too complex and reduces the overfitting. The maximum features (also referred to as “mtry” in literature) determines the number of features randomly selected at each split when building the decision trees. The maximum features should be less than the total number of predictor variables, and is varied from 2 to 5 in this study. This choice covers the canonical value of $\sqrt{p} \sim 3$, where p is the number of predictor variables (Probst et al., 2019). Limiting the maximum features reduces overfitting by increasing the randomness and diversity among the trees.. A large number of trees avoids

overfitting and stabilizes the algorithm, and it is set to 500 in this study. The combination of these hyperparameters results in 18 different configurations of the RF algorithm.

A wide range of ML model complexity are explored through the diverse set of hyperparameter sets. As stated in section 2.1, observed O_2 profiles are averaged into 1°x1° longitude-latitude bins during pre-processing, which combines O_2 profiles from close proximity in space and time into a single profile. Furthermore, DKCV is performed to select the best possible configuration of the hyperparameters. Then we use ~20% of data unused during the algorithm training to assess the algorithm skill and uncertainty of the resulting mapping products. The schematic diagram (**Figure 5**) shows the overall workflow. The best performing algorithm is selected after training of all possible combination of hyperparameters for each combination of input/output variables and algorithm type. The performance metrics are root-mean-square error (RMSE) and R^2 values. Once the best performing hyperparameters are found, the algorithms are further evaluated with additional performance metrics including mean bias, root-mean-square-error (RMSE), and R^2 value using the 20 % of the data that are held out from the training. Using all of these factors, the ML algorithms' performances are measured, and the gridded O_2 datasets are generated by projection of predictor variables.

(Figure 5 here)

Figure 5. The workflow. This flowchart describes the preprocessing, training, tuning, and testing of the algorithm to map O_2. The shaded region is repeated for 6 experiments (with different input dataset) and 2 algorithms (NN and RF). The main outcomes are the gridded

monthly maps and its uncertainty estimates boxed with the thick line. The entire workflow is repeated for each of the 5 basins.

3. Hyperparameter tuning and performance evaluation

A total of 1080 ML algorithms is trained including 540 NN and 540 RF regression models based on different combinations of input/output variables and hyperparameter sets for the 5 ocean basins. Each of the algorithms is trained 5 times using DKCV approach, thus the total of 5,400 trainings are performed. These calculations were computationally demanding but they can be efficiently carried out in parallel computing platform using Derecho supercomputers at National Center for Atmospheric Research (CISL, 2019).

3.1 Optimization of hyperparameters

For each set of input/output variables (**Table 1**), all possible configurations of hyperparameters are explored with the DKCV approach, and the RMSE and R^2 scores are recorded. **Figure 6** and **7** shows the mean RMSE scores for the NN and RF algorithm for each basin with the hyperparameter sets listed in **Supplementary Tables 1** and **2**. The algorithms are capable of reproducing O_2 observations withheld from the training with RMSE range of 15-22 μmol/kg in all basins, which is greater than the measurement errors. It suggests that the mapping (interpolation) error is the largest source of uncertainty in this dataset, which will be discussed in detail later (**Section 5**). The R^2 scores (not shown) are approximately 0.9 and higher in all basins except for the Arctic ($R^2 \sim 0.7$). The relatively low R^2 in the Arctic basin may reflect the skewness in the sample distribution as most profiles are taken before 1990s and are primarily from the Atlantic sector, in addition to the presence of sea ice which makes the shipboard

observation difficult. The profile counts of the Southern and Indian Ocean are not significantly greater than that of the Arctic, but they have wider spatial and temporal data coverage owing to the recent deployment of BGC Argo floats.

(Figure 6 here)

Figure 6. Mean RMSE scores from the hyperparameter tuning of NN using DKCV approach. Results from Exp1 (left, S for ship-only) and Exp 4 (right, S+A for ship+Argo) are shown. Each row from top to bottom is Atlantic, Pacific, Indian, Southern and Arctic Oceans. Color bar shows the magnitude of the RMSE in the units of μmol/kg. The dots indicate the best performing hyperparameter set.

The hyperparameter sets with the lowest RMSE score are selected as the best performing algorithm, and they generally match the ones with the highest R^2 scores. The displayed cases in **Figure 6** and **7** are from Exp 1 (ship only) and 4 (ship+Argo) but the general outcomes from other experiments are similar, and are displayed as supplementary **Figure S1-4**. Comparing the results between the cases with/without Argo data, there is a noticeable difference in the overall magnitude of RMSE where the inclusion of Argo data decreased the error in NN.

In the NN algorithm, the number of hidden layers/nodes determines the complexity of the algorithm. The best performing configurations are different depending on the basin and on the inclusion of Argo data (**Figure 6**). The most complex configuration (60-60-60-60, 4 hidden layers with 60 nodes for each layer) is selected for the Atlantic basin including shipboard and Argo data where the highest regularization coefficient reduced the risk of the overfitting. The

most complex configuration (60-60-60-60 with lowest regularization) performed the poorest in general. Simpler configurations with fewer number of nodes performed better in relatively data sparse basins including Indian, Southern and Arctic basins. There are multiple configurations that exhibit similarly low value of RMSE. There are potentially multiple hyperparameter choices that perform equally well, indicative of trade-offs between regularization and model complexity.

The primary determining factor for RF is the max_features (see **Figure 7**). For this study, the canonical max_feature value is 3 (Probst et al., 2019) which was selected for the most basins. In the relatively data-rich Atlantic and Pacific Ocean, larger values of min_samples_split performed slightly better, which also avoids overfitting. Conversely, lower values of min_samples_split performed slightly better in the relatively data-sparse Indian and Southern Ocean.

(Figure 7 here)

Figure 7. Same as Figure 6 but for the RF algorithm.

3.2 Validation and quantification of uncertainties using the test data

We selected the best performing hyperparameter sets for NN and RF algorithms using DKCV, and the algorithms are re-trained using all training data. They are evaluated against the test data which consists of ~20% of all input data that are set aside and unused for training. The test data is assembled from randomly selected 11 out of 55 years, and it is randomized differently for each basin. **Figure 8** shows an example of the distribution of the test data from RF with Exp 5. These test data are used to evaluate the algorithm and to quantify the uncertainties. The

performance is evaluated using three metrics including mean bias, RMSE and R^2, and the results are listed in **Supplementary Table 3** and **4**.

The general performance of both algorithms is quite high with the overall R^2 scores of 0.9 and higher with the exception of the Arctic Ocean where it is in the range of 0.6-0.8. The mean biases are generally low for all basins, less than 2.1 μmol/kg for all algorithms. The magnitude of RMSE is in the range of 13-18 μmol/kg. RF algorithms overall performed slightly better than NN in terms of these metrics. Comparing the results from ship-only (Exp 1-3) and ship and Argo (Exp 4-6), these metrics are overall similar.

The panel C and D of **Figure 8** shows the spatial distribution of the error as calculated by the difference between the algorithm reconstructions and the test data based on the RF algorithm from the Exp 5. The specific choice of the algorithm and input data does not significantly impact on the overall structure of the error field. The major regions of disagreements are close to strong background O_2 gradients. Relatively large errors >20 μmol/kg occurs near the oxycline at the depth range where there are strong vertical gradients. Similarly large errors are found at the frontal region in the Southern Ocean and at the lateral boundaries of the tropical oxygen minimum zone.

(Figure 8 here)

Figure 8. Test data for RF algorithm from the Exp 5. (A) Spatial distribution of test profiles that are unused for training of algorithm. (B) Temporal distribution of the test profiles. (C, D) Meridional section of misfit between estimated O_2 and test data as colored dots in Atlantic and

Pacific basin taken from the boxed regions in panel A. The contours are annual mean climatological O_2 concentrations based on World Ocean Atlas 2018 (Garcia et al., 2018).

The evaluation of ML algorithms, so far, are performed based on the total O_2 concentrations including climatological means and its temporal variability and trends. The high R^2 values and very small mean bias suggest that the algorithms can reconstruct the overall spatial patterns and mean values relatively well. The detailed comparison with the test data revealed the weaknesses of the reconstructed O_2 in the oxycline regions and the lateral water mass boundaries. The amplitude of O_2 anomalies generated by displacements of water parcel scales with the magnitude of background O_2 gradient. A possible explanation for very small mean bias (<3 μmol/kg) and relatively high RMSE (13-18 μmol/kg) in the oxycline and frontal region can be that the algorithm captures the climatological O_2 distribution well but struggles to represent the spatial and temporal variability. It is difficult to assess the algorithm skill separately for background climatology and anomalies because of temporally sparse sampling. However, this is not necessarily the case at the sites of ocean time series stations. The algorithm reconstructions and test data are compared at two ocean time series stations, Station P (OSP, 145°00'W, 50°00'N) in the northeastern Pacific Ocean, and Bermuda Atlantic Time Series (BATS, 64°10'W, 31°40'N) in the subtropical North Atlantic Ocean. Larger numbers of samples taken at these stations allow focused examination of skill for the reconstruction of temporal variability at these stations.

Figure 9 shows the comparison of reconstructed and test data at the specific location of OSP and BATS. Two isopycnal layers are selected for each station. **Figure 9AB** shows sigma-theta level 26.8 and 27.0 which are in the upper/lower oxycline depths at OSP. As expected, R^2

values are in the range of 0.5 to 0.8, which is lower than the basin-scale value (~0.9). The RMSE is in the range up to 18 μmol/kg, which is comparable to the basin-scale value. **Figure 9C** shows sigma-theta level 26.3 at BATS which is within the North Atlantic Subtropical Mode Water. **Figure 9D** shows sigma-theta level 26.8 in the lower thermocline. Again, R^2 values are in the range of 0.6 to 0.8, which reflects the algorithm skill to represent variability at these specific locations, and the RMSE is in the range of 6-8 μmol/kg which is significantly lower than the basin-scale value. While we admit that the algorithms are not perfect, it is encouraging that this approach can capture a significant fraction of temporal variability at these sites.

(Figure 9 here)

Figure 9 Validation of O_2 variability at two ocean time series stations. (A,B) Ocean Station Papa and (C,D) Bermuda Atlantic Ocean Time Series. Blue dots are for the NN algorithm, and red dots are for the RF. The blue solid lines are 1:1 line.

For most input/output variable sets (**Table 1**), RF algorithms showed lower RMSE than NN but R^2 and mean biases are similar to NN, indicating slightly better skill (**Supplementary Table 3** and **4**). Comparing the algorithms trained with shipboard only (Exp 1-3) and shipboard and Argo data (Exp 4-6), there is a slight improvement in terms of RMSE or R^2 in favor of inclusion of Argo data. This is expected as Argo data contributed to the significant increase in data coverage even though it is limited to the period after 2005.

In comparison to a recently developed global dataset, GOBAI-O_2 (Sharp et al., 2023), whose global-scale RMSE is 8.8 μmol/kg, our results show a larger RMSE of 13-18 μmol/kg.

GOBAI-O_2 employs similar neural network and random forest algorithms under different configurations, and their data sources are mainly based on Argo-O_2 (with additional GLODAPv2 profiles), thus we do not expect the same uncertainties.

3.3 Evaluation of climatological O_2 distribution

Using the algorithms developed and tested in Section 3.2, we projected O_2 distributions using the gridded EN4 data from 1965 to 2020, and we further analyze the results in comparison to the well-established climatological distribution using World Ocean Atlas 2018 (WOA18). **Figure 10** shows the summary of comparison for annual mean climatology averaged over 0-1,000 m. This is not a validation in the strict sense since many of the shipboard data used to assemble World Ocean Atlas were also used in the training of the algorithms. Rather, it is reassuring to find similar climatological distribution to the widely adopted WOA18 since our method of mapping is fundamentally different from that of WOA. **Figure 10AB** shows vertically averaged, annual mean O_2 climatology from the WOA18 and the ensemble average of algorithm-based reconstruction.

(Figure 10 here)

Figure 10. Comparison of annual mean climatology between this study and World Ocean Atlas 2018 (WOA18). (A) Vertically averaged WOA18 O_2 climatology from 0-1,000m. (B) Same as (A) but for the ensemble mean of both algorithms including Exp1-6. (C) Difference between (A) and (B). (D) Area-weighted mean vertical profile of O_2. Red line is WOA18. Blue lines are ship-

only (Exp 1-3) algorithms, and magenta lines are ship+Argo (Exp 4-6) algorithms. (E) Differences between the algorithms in (D) and WOA18.

Figure 10C shows a slight, widespread negative bias in the open ocean with some localized overestimation of O_2 along the eastern boundary of ocean basins and along the Antarctic coasts. The horizontally (area-weighted) averaged vertical profile of O_2 is displayed in **Figure 10D**. It shows general similarity between the WOA18 and the reconstructed O_2, and the difference between them (**Figure 10E**) reveals the negative bias of 2-4 μmol/kg overall. The reconstructed O_2 climatologies with ML approaches are slightly lower than WOA18. The inclusion of Argo-O_2 data does not significantly impact on the negative bias of the climatological O_2 profile. Factors contributing to the negative mean bias may include differences in the period represented by the WOA18 and this study. The period represented by the ML-based climatology may reflect the time windows over which the training data were collected. The representations of the temporal trends are further examined in **Section 4**. Comparing the reconstructions with potential density and/or stratification as additional input variables (Exp 2,3,5,6), the addition of these variables did not significantly change the climatology. Based on the comparison with WOA18, both RF and NN algorithm performed well for reproducing the annual mean climatology.

3.4 Feature importances to explain relative contributions

In the RF algorithm, feature importances measure the relative importance between each of the predictor variables in estimating O_2. It is calculated by randomly removing a feature from the dataset during training and measuring how much each feature decreases the algorithm's overall accuracy. The larger the decrease in performance, the more important the feature is

deemed to be. **Figure 11** shows the feature importances determined from the Exp 1-6 with the best performing hyperparameter sets for each basin. The feature importances significantly vary across basins. In the Atlantic and Indian Ocean, latitude was considered the most influential variable in making O_2 estimation. In the Pacific and Southern Ocean, pressure was the most influential variable. Other variables, such as salinity, temperature, longitude, and potential density, all played some roles when they are included with relatively small influences. Inclusion of Argo data made little impact as the pairs of Exp 1&4, 2&5, and 3&6 show very similar feature importances.

(Figure 11 here)

Figure 11. Feature importances of the Random Forest algorithm for each basin. The relative importance of each feature variables is shown for Exp 1 through 6. "Pden" indicates potential density. Latitude (Lat*) and longitude (Lon*) are transformed to polar stereographic coordinates for the Arctic and Southern Ocean.

Feature importances offer insights into which factors contribute most significantly to the estimation of O_2. Climatological O_2 significantly varies latitudinally and in depth (pressure), likely making them two of the most important factors. Variability of T/S on isopycnal surfaces can indicate water mass shifts and circulation variability, thus these variables can play some important roles in estimating O_2 variability. Comparing Exp 1 and 2 (and Exp 4 and 5), the addition of potential density, in some cases, reduced the relative importance of T and S. Similarly, comparing Exp 2 and 3 (and Exp 5 and 6), the further addition of N^2 does not

significantly change the importance of T/S/σ_θ, indicating some roles played by the stratification. It is important to note that feature importances are calculated for the specific configuration of RF algorithms used in this study, and they may not indicate causal relationships.

4. Assessment of deoxygenation trends

Based on the comparison with the test data and annual mean climatology, we consider both NN and RF to provide reasonable reconstructions of the O_2 distribution, forming 12 ensemble members (NN 1-6 and RF 1-6) where numbers after NN and RF indicates the experiment number in **Table 1**. This ensemble includes 6 algorithms trained with shipboard data only, and another 6 with shipboard and Argo data. The following analysis aims to evaluate the impact of the Argo data on the reconstructions of deoxygenation trends.

The top panel in **Figure 12** shows the 12-month running mean of the O_2 inventory time series integrated over 0-1,000m. **Figure 12A** shows results from all algorithms grouped by (blue) ship-only and (red) ship and Argo ensemble members with the mean and range of 6 reconstructions respectively. **Figure 12BC** show separately the results of RF and NN algorithms including the mean and range of 3 reconstructions for each algorithm type. In general, all ensemble members show a moderate decrease from 1965 to around 1990, followed by stronger decline after 1990. They share similar climatological O_2 inventory but the deoxygenation trend is generally stronger in the NN algorithm. Before 2000s, the ship-only and ship+Argo reconstructions show similar O_2 inventory, and they diverge after mid-2000s regardless of the algorithm type. This coincides with the introduction of Argo O_2 data after 2005.

The ship+Argo reconstructions have much stronger deoxygenation rates during the 2010s. To highlight this point, deoxygenation trends of O_2 inventory (0-1,000m) are calculated

for 12 ensemble members over the 40-year period between 1970-2010. To avoid the end-point effect, linear regression is not used and the trend is calculated by taking the difference of 10-year means between (1965-1975) and (2005-2015). The ensemble mean trend is -229 $\pm$ 33 Tmol/decade for the ship only reconstructions, and the uncertainty is estimated based on the ensemble range. The magnitude of this trend is stronger than the recent deoxygenation estimates based on optimal interpolation of shipboard observation by Ito et al., (2024) of -175 $\pm$ 24 Tmol/decade. The optimal interpolation of Ito et al., (2024) likely underestimated the deoxygenation trend in data-sparse regions. For the ship+Argo reconstructions, the ensemble mean trend is -358 $\pm$ 93 Tmol/decade, which is approximately 57% stronger than the ship-only reconstruction, and caused by the stronger O_2 decrease after the mid-2000s.

(Figure 12 here)

Figure 12. Oxygen inventory in the units of Pmol (10^{15} mol) and its ensemble spread. (A) All 12 ensembles including RF and NN algorithms. (B) RF only and (C) NN only. 12 month running mean is applied to remove mean seasonal cycle. The shaded region is the ensemble spread calculated as the difference between yearly maximum and minimum.

Figure 12B shows the O_2 inventory time series based on the RF algorithm for the ship-only and ship+Argo groups, which has two important implications. First, the inclusion of Argo data significantly changes the recent (2005-) trajectory of O_2 inventory with a significantly stronger deoxygenation rate. Secondly, the additional input of potential density and/or stratification (N^2)

had practically no impact for the RF algorithms. However, these variables made significant impacts on the NN algorithm.

Figure 12C shows the O_2 inventory time series from the NN algorithms. Similar to the RF, the inclusion of Argo increases the deoxygenation rate. The NN algorithms also shows more spread across the ensemble with additional input of potential density and/or stratification. The ranges of reconstructed O_2 inventories are different between ship-only and ship+Argo groups, primarily coming from the NN algorithm. The range is calculated by the difference between maximum and minimum O_2 inventories (as illustrated by the blue and red shaded regions in **Figure 12**) which primarily comes from the NN algorithms. On average, the range of ship-only algorithms is 0.77 Pmol. The inclusion of Argo data significantly reduced the range to 0.47 Pmol, which is approximately 40% less than the ship-only case. This implies that the inclusion of Argo data not only increases the magnitude of deoxygenation trends but also reduces the spread between ensemble members with different configuration of input variables.

Spatial patterns of O_2 changes are examined as difference between the two decadal averages centered at 1970 and 2010. **Figure 13A** shows the horizontally (area-weighted) averaged vertical profiles of O_2, O_2 solubility and (-1) x AOU. The concentration of O_2 at saturation (O_{2sol}) is calculated with solubility coefficients derived from the data of Benson and Krause (1984) as fitted by Garcia and Gordon (1992). AOU stands for apparent oxygen utilization, and it is defined as the difference between O_2 solubility and O_2, $AOU = O_{2sol}(S,T) - O_2$. Near the surface, the O_2 decline is relatively moderate, and it sharply increases through the upper 200m. The largest O_2 decline occurs approximately at 150m depth, and it becomes relatively constant below 300m. The breakdown of the O_2 decrease is approximately equal split

between solubility and AOU at the surface, but the importance of AOU increases with depth, and approximately 85-90% of O_2 decline is explained by AOU below 300m.

Figure 13BCD shows the spatial patterns of the 40-year change as vertically (thickness-weighted) averages of O_2, O_{2sol} and (-1) x AOU over 0-1,000m. There are regions of strong O_2 decline including North Pacific, Southern Ocean, equatorial oceans in all basins, and along the northeastern coastline of North American continent. This pattern is in good agreement with the previously published study by Oschlies et al., (2018, Figure 3a). Comparing the patterns between the two components, AOU clearly dominates the overall O_2 decline, and in some regions, the solubility contributes significantly, including the northeastern coastline of North America and the frontal regions in the Southern Ocean. Subtropical south Pacific and south Indian Oceans are regions of moderate O_2 increase, and these features are also consistent with the previous work (Oschlies et al., 2018). In summary, the main driver of O_2 decline is AOU in the upper 1,000m, and our results are in qualitative agreement with previous works in terms of spatial patterns (e.g. Schmidtko et al., 2017; Oschlies et al., 2018; Ito et al., 2017).

(Figure 13 here)

Figure 13. 40-year ensemble mean change of O_2, AOU and O_2 solubility for (A) area-weighted horizontal averages and (BCD) thickness-weighted vertical averages from 0-1,000m. All plotted values are concentrations in the units of μmol/kg. The ensemble mean is calculated for both algorithm type trained with shipboard and Argo data. The 40-year change is estimated as the difference between the two 10-year means between (2005-2014) minus (1965-1974).

Figure 14 shows the regional breakdown of the O_2 inventory with an emphasis on comparing (blue) ship-only and (red) ship+Argo reconstructions. There are regional differences in the evolution of O_2 inventory. While there are significant overlaps between ship-only and ship+Argo cases, the ship+Argo reconstructions exhibit stronger decrease of O_2 inventory after 2000s including North/Equatorial Atlantic, North/Equatorial Pacific, Indian and Southern Oceans. Inclusion of Argo data significantly reduced the ensemble spread in the Equatorial/South Pacific Ocean. Detailed results from inventory trend calculations are displayed in the supplementary Table 5.

(Figure 14 here)

Figure 14. Basin-scale O_2 inventory trend. Global ocean is divided to 10 basins. Blue lines and shading show ensemble mean and ensemble range for ship-only reconstructions, and red lines and shading are for ship+Argo reconstructions. Arctic is northward of 60°N, and the Southern Ocean is southward of 50°S. The division between equatorial and North Atlantic/Pacific basin is set to 15°N, and the division between equatorial and South Atlantic/Pacific/Indian basin is set to 15°S.

The inclusion of Argo data has different impact on the estimated deoxygenation trend. Globally, the 40-year trend (1970-2010) increased by 56% when Argo data is included. The strongest effects are in the Equatorial Pacific (+94%), Equatorial Indian (+73%), and South Indian (+66%), where these regions have been relatively under-sampled by historical shipboard observations (**Figure 1**). In terms of the contributions to the global deoxygenation trend, the

three major regions are North Pacific (23%), Equatorial Pacific (20%), and Southern Ocean (18%). These three regions together explain more than 60% of the global deoxygenation trend, and this is predominantly (>85%) driven by the increasing AOU. Thus, the effect of warming-induced solubility loss is unlikely the major mechanism for ocean deoxygenation, and it must be primarily driven by the circulation and biochemical changes as expressed by the AOU component.

5. Uncertainty analysis

There are 3 types of uncertainty including measurement error, sampling error and mapping (interpolation) error, and for each type, there can be random errors and biases. Assuming that measurement (ΔO_{2meas}), sampling (ΔO_{2sampl}) and interpolation ($\Delta O_{2interp}$) errors are independent and uncorrelated, the combined median uncertainty can be calculated as:

$$\Delta O_2 = \left\{\Delta O_{2meas}^2 + \Delta O_{2sampl}^2 + \Delta O_{2interp}^2\right\}^{1/2} \qquad (1)$$

Measurement errors depend on specific techniques and instrumentation for making measurements. Bottle O_2 can include random errors of 1 μmol/kg or smaller with Winkler titration (Carpenter, 1965). CTD-O_2 sensors calibrated to Winkler O_2 data is expected to have similar errors. Delayed-mode adjusted Argo-O_2 has overall errors of about 3 μmol/kg (Maurer et al., 2021). In the oxycline region, there can be a larger error of approximately 10 μmol/kg for Argo-O_2 data due to uncorrected sensor response time, potentially including random and systemic bias components. Response time correction has not yet been applied to the delayed mode adjusted data used in this study. For simplicity, uniform constant measurement error is

assumed including ΔO_{2meas} = 1 μmol/kg for bottle and CTD-O_2 data, and ΔO_{2meas} = 3 μmol/kg for the Argo-O_2 data.

When multiple profiles are available within the monthly bin, the standard deviation can be used to estimate the magnitude of the sampling error. The variance of the binned data is averaged over time and depth for the shipboard (bottle/CTD-O_2) and Argo data separately. They are combined with the measurement errors according to Eq. (1). **Figure 15AB** shows the non-uniform distribution of this uncertainty. The global mean value of the combined measurement and sampling error is 4.8 μmol/kg for the shipboard (bottle/CTD-O_2) data, and is 5.5 μmol/kg for Argo data. However, there is significant spatial variability for the sampling errors likely due to the regional variability of the background O_2 gradient and wave/eddy activities. Other potential factors can include the sampling density. It can exceed 20 μmol/kg in regions such as Scotia and Newfoundland shelves and Kuroshio/Oyashio region.

(Figure 15 here)

Figure 15. An estimate of measurement, sampling and mapping errors based on the standard deviation and vertically averaged over 0-1,000m including (A) shipboard measurement and sampling errors, (B) Argo measurement and sampling errors, and (C) mapping errors from the test data including Exp 4-6 for both algorithms. The units are in μmol/kg.

Mapping uncertainties can be estimated by the comparison with the O_2 data withheld from the training as documented in section 3.2. The estimated O_2 values had the mean bias of less than 2 μmol/kg and RMSE of 13-18 μmol/kg globally. Its spatial structure can be calculated

by the misfit between the reconstructed O_2 and test data not used in the training of algorithm. The misfit can include mean bias and random error. To calculate the random component of the mapping error ($\Delta O_{2interp}^2$), the square of average misfit (mean bias) is subtracted from the averaged misfit squared, and this calculation is performed for each longitude-latitude grid cell to determine pattern of mapping uncertainty (**Figure 15C).** The global mean value of the interpolation (mapping) error ($\Delta O_{2interp}$) is 12.3 μmol/kg. Similar to the sampling errors, the magnitude of the mapping error is elevated nearby the eastern and western boundary current systems, tropical oxygen minimum zones, and the Southern Ocean. These regions contain elevated levels of horizontal and vertical gradients of O_2. This error estimates are comparable but somewhat greater than the magnitude of "algorithm errors" for the GOBAI-O_2 dataset of Sharp et al., (2023). Based on the typical magnitudes of these errors as discussed above, the combined uncertainty is approximately 13.5 μmol/kg globally, which is primarily dominated by the interpolation errors and secondary by the sampling error. The uncertainty is regionally elevated near the edge of oxygen minimum zones and strong ocean currents close to the western and eastern boundaries of ocean basins.

6. Conclusion

Since the mid-2000s, Argo floats equipped with O_2 sensors have been deployed in different parts of the global oceans, and the development of in-situ calibration methods reduced the measurement uncertainties of the Argo-O_2 sensors to approximately 3 μmol/kg. Coincidentally the number of shipboard observations has decreased in recent decades, and as a result, it is difficult to estimate the basin-scale deoxygenation trends based on shipboard observation only. Recently, a gridded, time-varying O_2 product has been developed using ML

approaches (Sharp et al., 2023), reconstructing the global O_2 distribution since 2004. This study and GOBAI-O_2 are similar in methodology, but there are some differences. Both this study and GOBAI-O_2 used delayed mode Argo data, but we further limited to the O_2 profiles calibrated with two established methods including in-air pO_2 measurement (Johnson et al., 2015; Bushinsky and Emerson 2015) and climatological air-sea disequilibrium (Takeshita et al., 2013). GOBAI-O_2 further applied a bias correction of -1.18 μmol/kg based on the match-up profiles (Sharp et al., 2013, Appendix D). The GOBAI-O_2 product is an average of two ML-based datasets with two-layer NN and RF. In this study, we trained a larger number of algorithms with varying sets of input data and hyperparameters and selected 12 algorithms to form an ensemble of O_2 estimates. Results from each of the ensemble members with and without Argo-O_2 data are available in public domain from zenodo (Ito and Cervania, 2024).

Contrasting algorithms trained with ship only and ship+Argo O_2 profiles was the main theme of this paper. The historical observations since 1965 included quality-controlled bottle and CTD-O_2 data, but the number of shipboard profiles has been declining since 1980s (see Figure 2). The inclusion of Argo data made two major impacts on the representation of global ocean deoxygenation. First, the inclusion of Argo data increased the magnitude of the global deoxygenation significantly. The 40-year (1970-2010) trend of 0-1,000m O_2 inventory is -229 $\pm$ 33 Tmol/decade for the ship only reconstructions, but it is -358 $\pm$ 93 Tmol/decade for the case of ship+Argo, which is approximately 56% stronger. This implies that recent increase in data coverage by BGC Argo array had impact on reconstruction of global-scale O_2 changes. Increased data coverage has contributed to capture recent declines of O_2 in regions where no shipboard observation is available. Secondly, the inclusion of Argo data significantly decreased the range of the global O_2 inventory estimates among the ensemble members. The range of ship-only

algorithms is 0.77 Pmol on average, and the inclusion of Argo data significantly reduced the range to 0.47 Pmol, which is approximately a 40% reduction. Thus, the inclusion of Argo data not only increases the magnitude of deoxygenation trends but also narrows the uncertainty.

Our uncertainty analysis considered three sources of errors including measurement, sampling, and interpolation errors. Of these, interpolation errors are likely the largest source of the errors with the global mean magnitude of 13.5 μmol/kg averaged over the ensemble members. Regionally it can be twice as high as this global mean value. These regions exhibit strong variability and strong background O_2 gradients, which can generate large uncertainties than the global mean. Additional potential source of errors includes the sensor calibration bias for Argo-O_2 observations due to finite response time of optode sensors, which may cause systemic bias in the oxycline regions (Bittig et al., 2014; 2018).

Due to the results of anthropogenic carbon dioxide and other greenhouse gas emissions, the ocean is warming, losing oxygen and being acidified. While these ecosystem stressors are projected to intensify for coming decades, our understandings of their impacts on marine ecosystems remains limited. While this study at 1°x1° resolution focused on improving the method of filling data gaps for basin-scale O_2 distribution, this resolution is too low for coastal studies. It remains to be tested how well ML approaches can be used to map biogeochemical properties at higher resolution in the coastal waters.

Acknowledgement

This project is supported by National Science Foundation (OCE-2123546). We acknowledge high-performance computing support from Cheyenne/Casper (doi:10.5065/D6RX99HX)

provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

Open Research

The gridded dissolved oxygen dataset generated from this study are available at zenodo (Ito and Cervania, 2024) via https://doi.org/10.5281/zenodo.11129715 with the Creative Commons Attribution 4.0. The source codes are available from github via https://github.com/takaito1/ML4O2 with the MIT license.

References

Barnes, S. L. (1964). A Technique for Maximizing Details in Numerical Weather Map Analysis. *Journal of Applied Meteorology and Climatology, 3*(4), 396-409.

Benson, B. B., & Krause Jr, D. (1984). The concentration and isotopic fractionation of oxygen dissolved in freshwater and seawater in equilibrium with the atmosphere1. *Limnology and Oceanography, 29*(3), 620-632. https://doi.org/10.4319/lo.1984.29.3.0620

Bindoff, N. L., Cheung, W. W. L., Kairo, J. G., Arístegui, J., Guinder, V. A., Hallberg, R., et al. (2019). *Changing Ocean, Marine Ecosystems, and Dependent Communities.* Retrieved from Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157964.007

Bittig, H. C., Fiedler, B., Scholz, et al. (2014). Time response of oxygen optodes on profiling platforms and its dependence on flow speed and temperature. Limnology and Oceanography: Methods, 12(8), 617-636.

Bittig, H. C., & Körtzinger, A. (2017). Technical note: Update on response times, in-air measurements, and in situ drift for oxygen optodes on profiling platforms. *Ocean Sci., 13*(1), 1-11. https://os.copernicus.org/articles/13/1/2017/

Bittig, H. C., Körtzinger, A., Neill, C., et al. (2018). Oxygen optode sensors: principle, characterization, calibration, and application in the ocean. Frontiers in Marine Science, 4, 429.

Bittig, H. C., Maurer, T. L., Plant, J. N., Schmechtig, C., Wong, A. P. S., Claustre, H., et al. (2019). A BGC-Argo Guide: Planning, Deployment, Data Handling and Usage. *Frontiers in Marine Science, 6*. Review. https://www.frontiersin.org/articles/10.3389/fmars.2019.00502

Boyer, T. P., Baranova, O. K., Coleman, C., Garcia, H. E., Grodsky, A., Locarnini, R. A., et al. (2018). *World Ocean Database 2018*., Ed. Mishonov, A.V., NOAA Atlas NESDIS, Silver Spring, MD.

Broullón, D., Pérez, F. F., Velo, A., Hoppema, M., Olsen, A., Takahashi, T., et al. (2019). A global monthly climatology of total alkalinity: a neural network approach. *EARTH SYSTEM SCIENCE DATA, 11*(3), 1109-1127.

Brunton, S. L., & Kutz, J. N. (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge: Cambridge University Press.

Bushinsky, S. M., & Emerson, S. (2015). Marine biological production from in situ oxygen measurements on a profiling float in the subarctic Pacific Ocean. *Global Biogeochemical Cycles, 29*(12), 2050-2060.

Carpenter, J. H. (1965). THE ACCURACY OF THE WINKLER METHOD FOR DISSOLVED OXYGEN ANALYSIS1. *Limnology and Oceanography, 10*(1), 135-140. https://doi.org/10.4319/lo.1965.10.1.0135

Chen, B. Z., Liu, H. B., Xiao, W. P., Wang, L., & Huang, B. Q. (2020). A machine-learning approach to modeling picophytoplankton abundances in the South China Sea. *PROGRESS IN OCEANOGRAPHY, 189*.

Chen, S. L., Hu, C. M., Barnes, B. B., Wanninkhof, R., Cai, W. J., Barbero, L., & Pierrot, D. (2019). A machine learning approach to estimate surface ocean pCO2 from satellite measurements. *REMOTE SENSING OF ENVIRONMENT, 228*, 203-226.

CISL. (2019). Cheyenne: HPE/SGI ICE XA System (University Community Computing). In. Boulder, CO: Computational and Information Systems Laboratory, National Center for Atmospheric Research.

Culberson, C.H., G. Knapp, M.C. Stalcup, R.T. Williams, and F. Zemlyak (1991). A comparison of methods for the determination of dissolved oxygen in seawater. Report No. WHPO 91-2, WOCE Hydrographic Program Office, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts., U.S.A.

Friedland, K. D., Morse, R. E., Shackell, N., Tam, J. C., Morano, J. L., Moisan, J. R., & Brady, D. C. (2020). Changing Physical Conditions and Lower and Upper Trophic Level Responses on the US Northeast Shelf. *Frontiers in Marine Science, 7*, 18.

Garcia, H. E., & Gordon, L. I. (1992). OXYGEN SOLUBILITY IN SEAWATER - BETTER FITTING EQUATIONS. *Limnology and Oceanography, 37*(6), 1307-1312. Note.

878 Garcia, H. E., Weathers, K., Paver, C. R., Smolyar, I., Boyer, T. P., Locarnini, R. A., et al.
879 (2018). *World Ocean Atlas 2018, Volume 3: Dissolved Oxygen, Apparent Oxygen*
880 *Utilization, and Oxygen Saturation*., NOAA Atlas NESDIS, Silver Springs, MD.

881 Gloege, L., McKinley, G. A., Landschützer, P., Fay, A. R., Frölicher, T. L., Fyfe, J. C., et al.
882 (2021). Quantifying Errors in Observationally Based Estimates of Ocean Carbon Sink
883 Variability. *Global Biogeochemical Cycles, 35*(4).

884 Good, S. A., Martin, M. J., & Rayner, N. A. (2013). EN4: Quality controlled ocean temperature
885 and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal*
886 *of Geophysical Research-Oceans, 118*(12), 6704-6716.

887 Gregor, L., Lebehot, A. D., Kok, S., & Scheel Monteiro, P. M. (2019). A comparative
888 assessment of the uncertainties of global surface ocean CO2 estimates using a machine-
889 learning ensemble (CSIR-ML6 version 2019a) – have we hit the wall? *Geosci. Model*
890 *Dev., 12*(12), 5113-5136. https://gmd.copernicus.org/articles/12/5113/2019/

891 Gruber, N., Boyd, P. W., Frölicher, T. L., & Vogt, M. (2021). Biogeochemical extremes and
892 compound events in the ocean. *Nature, 600*(7889), 395-407.
893 https://doi.org/10.1038/s41586-021-03981-7

894 Huang, Y. B., Tagliabue, A., & Cassar, N. (2022). Data-Driven Modeling of Dissolved Iron in
895 the Global Ocean. *Frontiers in Marine Science, 9*.

896 Ho, T. K. (1995). Random Decision Forests. Proceedings of the 3rd International Conference on
897 Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

Ito, T., & Cervania, A. (2024). Machine Learning for O2 Project (ML4O2) [Data set]. In Journal of Geophysical Research Machine Learning and Computation. Zenodo. https://doi.org/10.5281/zenodo.11129715.

Ito, T., Garcia, H. E., Wang, Z., Minobe, S., Long, M. C., Cebrian, J., et al. (2023). Underestimation of global O2 loss in optimally interpolated historical ocean observations. *Biogeosciences Discuss., 2023*, 1-22. https://bg.copernicus.org/preprints/bg-2023-72

Johnson, K. S., Coletti, L. J., Jannasch, H. W., Sakamoto, C. M., Swift, D. D., & Riser, S. C. (2013). Long-Term Nitrate Measurements in the Ocean Using the in situ Ultraviolet Spectrophotometer: Sensor Integration into the APEX Profiling Float. *JOURNAL OF ATMOSPHERIC AND OCEANIC TECHNOLOGY, 30*(8), 1854-1866.

Johnson, K. S., Plant, J. N., Riser, S. C., & Gilbert, D. (2015). Air Oxygen Calibration of Oxygen Optodes on a Profiling Float Array. *JOURNAL OF ATMOSPHERIC AND OCEANIC TECHNOLOGY, 32*(11), 2160-2172.

Kleinberg E (2000). "On the Algorithmic Implementation of Stochastic Discrimination". IEEE Transactions on Pattern Analysis and Machine Intelligence. 22 (5): 473–490.

Landschützer, P., Gruber, N., Bakker, D. C. E., Schuster, U., Nakaoka, S., Payne, M. R., et al. (2013). A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink. *Biogeosciences, 10*(11), 7793-7815.

LeCun. Y, L. Bottou, Y. Bengio and P. Haffner, (1998), Gradient-based learning applied to document recognition, in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

Maurer, T. L., Plant, J. N., & Johnson, K. S. (2021). Delayed-Mode Quality Control of Oxygen, Nitrate, and pH Data on SOCCOM Biogeochemical Profiling Floats. *Frontiers in Marine Science, 8*. Methods. https://www.frontiersin.org/articles/10.3389/fmars.2021.683207

Moussa, H., Benallal, M. A., Goyet, C., & Lefèvre, N. (2016). Satellite-derived CO₂ fugacity in surface seawater of the tropical Atlantic Ocean using a feedforward neural network. *INTERNATIONAL JOURNAL OF REMOTE SENSING, 37*(3), 580-598.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Pershing, A. J., Alexander, M. A., Hernandez, C. M., Kerr, L. A., Le Bris, A., Mills, K. E., et al. (2015). Slow adaptation in the face of rapid warming leads to collapse of the Gulf of Maine cod fishery. *Science, 350*(6262), 809-812.

Probst P, Wright MN, Boulesteix A-L. (2019) Hyperparameters and tuning strategies for random forest. WIREs Data Mining Knowl Discov. 9:e1301. https://doi.org/10.1002/widm.1301

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature, 566*(7743), 195-204.

Roemmich, D., Alford, M. H., Claustre, H., Johnson, K., King, B., Moum, J., et al. (2019). On the Future of Argo: A Global, Full-Depth, Multi-Disciplinary Array. *Frontiers in Marine Science, 6*. Review. https://www.frontiersin.org/articles/10.3389/fmars.2019.00439

Sarmiento, J. L., Johnson, K. S., Arteaga, et al. (2023). The Southern Ocean carbon and climate observations and modeling (SOCCOM) project: A review. Progress in Oceanography, 103130.

Sauzède, R., Bittig, H. C., Claustre, H., de Fommervault, O. P., Gattuso, J. P., Legendre, L., & Johnson, K. S. (2017). 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*.

Seidov, D., Mishonov, A., Reagan, J., Baranova, O., Cross, S., & Parsons, R. (2018). REGIONAL CLIMATOLOGY OF THE NORTHWEST ATLANTIC OCEAN High-Resolution Mapping of Ocean Structure and Change. *Bulletin of the American Meteorological Society, 99*(10), 2129-2138.

Sharp, J. D., Fassbender, A. J., Carter, B. R., Johnson, G. C., Schultz, C., & Dunne, J. P. (2023). GOBAI-O2: temporally and spatially resolved fields of ocean interior dissolved oxygen over nearly 2 decades. *Earth Syst. Sci. Data, 15*(10), 4481-4518. https://essd.copernicus.org/articles/15/4481/2023/

Sharp, J. D., Fassbender, A. J., Carter, B. R., Lavin, P. D., & Sutton, A. J. (2022). A monthly surface pCO_2 product for the California Current Large Marine Ecosystem. *EARTH SYSTEM SCIENCE DATA, 14*(4), 2081-2108.

959 Stramma, L., Johnson, G. C., Sprintall, J., & Mohrholz, V. (2008). Expanding Oxygen-Minimum
960 Zones in the Tropical Oceans. *Science, 320*(5876), 655-658.
961 https://doi.org/10.1126/science.1153847

962 Takeshita, Y., Martz, T. R., Johnson, K. S., Plant, J. N., Gilbert, D., Riser, S. C., et al. (2013). A
963 climatology-based quality control procedure for profiling float oxygen data. *Journal of*
964 *Geophysical Research-Oceans, 118*(10), 5640-5650.

965 Wunsch, C. (1996). *The Ocean Circulation Inverse Problem*. Cambridge: Cambridge University
966 Press.

967 Zeng, J. Y., Nojiri, Y., Nakaoka, S., Nakajima, H., & Shirai, T. (2015). Surface ocean CO_2 in
968 1990-2011 modelled using a feed-forward neural network. *GEOSCIENCE DATA*
969 *JOURNAL, 2*(1), 47-51.

970