Title: Improved estimates of North Atlantic deoxygenation trends by combining shipboard and Argo observations using machine learning algorithms

Authors: Takamitsu Ito and Ahron Cervania

Affiliation: School of Earth and Atmospheric Sciences, Georgia Institute of Technology

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 is uncertain due to its sparse and irregular sampling. A refined estimate of deoxygenation rate is developed for the North Atlantic basin using machine learning techniques and biogeochemical Argo array. The source data includes 159 thousand historical shipboard (bottle and CTD-O_2) profiles from 1965 to 2020 and 17 thousand Argo O_2 profiles after 2005. Neural network and random forest algorithms were trained using 80% of this data using different hyperparameters and predictor variable sets. From a total of 240 trained algorithms, 12 high performing algorithms were selected based on their ability to accurately predict the 20% of oxygen data withheld from training. The final product includes gridded monthly O_2 ensembles with similar skills (mean bias < 1mol/kg and $R^2 > 0.95$). The reconstruction of basin-scale oxygen inventory shows a moderate increase before 1980 and steep decline after 1990 in agreement with a previous estimate using an optimal interpolation method. However, significant differences exist between reconstructions trained with only shipboard data and with both shipboard and Argo data. The gridded oxygen datasets using only shipboard measurements resulted in a wide spread of deoxygenation trends (0.8-2.7% per decade) during 1990-2010. When both shipboard and Argo were used, the resulting deoxygenation trends converged within

a smaller spread (1.4-2.0% per decade). This study demonstrates the importance of new biogeochemical Argo arrays in combination with applications of machine learning techniques.

Plain language summary

Oxygen is an essential molecule existing in the seawater. But its concentrations are declining in many parts of the oceans. Its causes 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 oceans. 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 improved the estimates of oxygen contained in the North Atlantic Ocean by applying machine learning techniques, with the specific goals to synthesize 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. Future work remains to apply this method beyond North Atlantic to the global oceans including the coastal waters.

Key points

- A new ensemble dataset of oxygen is developed for the North Atlantic basin 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 reduced the ensemble spread in the deoxygenation rate by approximately a factor of 4.

1. Introduction

Historical observations from past decades have shown growing influences of anthropogenic perturbations on marine ecosystem and biogeochemistry (Friedland et al., 2020; Gruber et al., 2021; Pershing et al., 2015; Seidov et al., 2018). 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 ± 0.35 % per decade, for the upper 1,000m (Bindoff et al., 2019). 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 sources of O_2 data 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 1μmol/kg. Most modern oxygen chemical titration measurements are based on Carpenter's whole bottle titration method and an amperometric or photometric end-detection with a precision of about 0.5-1μmol/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.

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 (Henry C. Bittig et al., 2019; Kenneth S. Johnson et al., 2013). Chemical sensors for measuring biogeochemical data require post-deployment quality control and calibration (Maurer et al., 2021). There are realtime, realtime 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. **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 1965 to 2020. WOD18 is an international collaboration among national data centers, oceanographic research institutions and investigators to provide a comprehensive dataset of quality-controlled oceanographic variables. Fewer profiles are taken in the open ocean away from the coasts, especially in the central subtropical regions. The number of profiles taken each year/month also fluctuates significantly. Prior to 1990, most O_2 profiles are taken by ship-based bottle measurements. After the 1990s, CTD-O_2 profiles increased and became the major O_2 data source. Since the mid-2000s, the number of Argo-O_2 profiles has steadily increased. Including all three platforms, the total of 176,049 profiles are taken in the North Atlantic basin from Equator to 65°N including 61% bottle, 29% CTD-O_2 and 10% Argo-O_2 measurements from 1965 to 2020. Focusing on the later period after January 2000, the total of 52,903 O_2 profiles are taken including 12% bottle, 56% CTD-O_2, and 32% Argo-O_2 measurements.

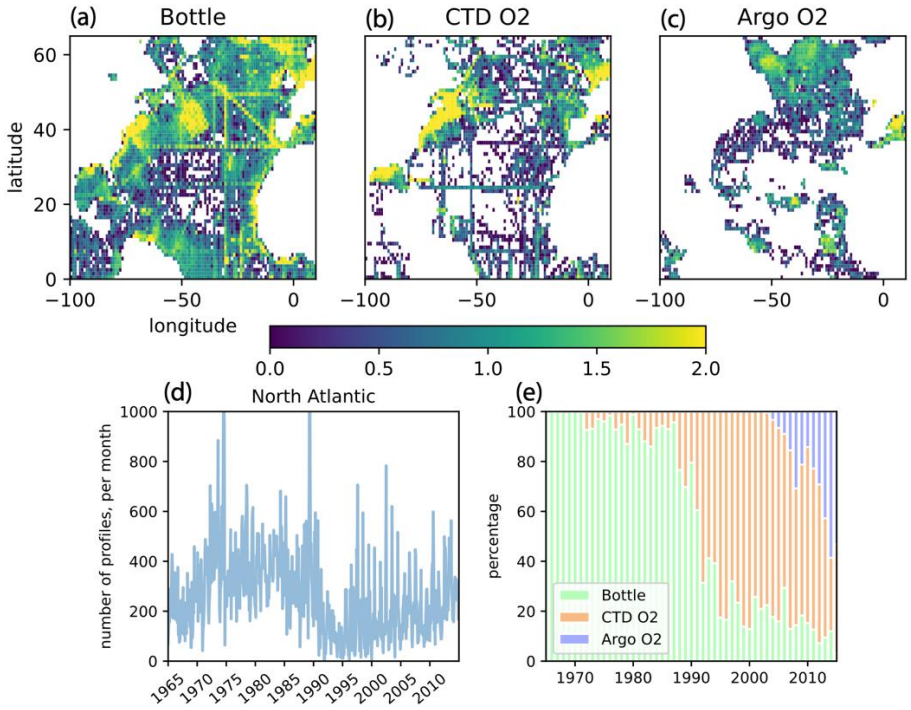


Figure 1. Sampling density (a-c) Logarithm (base 10) of the cumulative profile count within each 1°x1° longitude-latitude cell for oxygen (O_2) based on the World Ocean Database 2018 (Boyer et al., 2018) downloaded in October 2023. The color saturates at 2 (more than 100 profiles) per cell since 1965. (d) The number of O_2 profiles per month in the North Atlantic basin. (e) The breakdown among the three data types between bottle, CTD-O_2 and Argo-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-fill 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 (S. L. 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 (S. L. 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 (B. Z. 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 with one another owing to large uncertainties.

In the North Atlantic basin, approximately one-third of all O_2 profiles are measured by biogeochemical Argo floats after January 2000, and its share is increasing (see **Figure 1**). The calibration of Argo-O_2 data is still under development, especially for the response time of optode sensors in the upper ocean oxycline (H. C. Bittig & Körtzinger, 2017). Despite these potential biases and uncertainties, there can still 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) in that a longer time period is covered from January 1965 to December 2020 using the combination of Argo-O_2 and historical shipboard observations. This study will form 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 potential reduction of uncertainties by the inclusion of Argo-O_2 data. Separate sets of ML-based O_2 ensembles are formed based on the algorithms trained with the shipboard data only and with the shipboard and Argo-O_2 data. The comparison of deoxygenation trends and the ensemble spread quantifies the potential uncertainty reduction in the deoxygenation trends.

2. Methods

This method 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

The shipboard O_2 measurements are obtained from WOD18. The preprocessing of the data includes a check for data quality using the WOD18 quality control (QC) flags. 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 the upper 1,000m of water column. The North Atlantic grid cells are selected according to the basin mask of the

World Ocean Atlas 2018 (H.E. Garcia et al., 2018). The target analysis period is after 1965 when the modern oxygen titration method is established by Carpenter as referenced above. Over the North Atlantic 108,643 bottle O_2 profiles and 50,223 CTD-O_2 profiles are obtained from WOD18 after questionable profiles are removed. 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. The bottle profiles are averaged within the 1°x1° bins monthly from 1965 to 1986. After 1987, the bottle and CTD-O_2 profiles are averaged within the 1°x1° bins weighted by the profile counts within the same month.

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 are searched for ones containing delayed-mode O_2 data using two standard methods of bias correction including in-air pO_2 measurement with atmospheric reanalysis data (Bushinsky & Emerson, 2015; K. S. Johnson et al., 2015) and climatological air-sea disequilibrium of surface O_2 (Takeshita et al., 2013). There are 276 BGC-Argo floats that satisfy this condition in the Atlantic basin. The Argo-O_2 data points with acceptable QC flag (indicated as 1, 2 or 8) are then placed into monthly bins which are the 1°x1° longitude-latitude grid cells.

This study aims to extract regional relationships that allow filling data gaps in O_2 using surrogate (predictor) variables such as temperature (T), salinity (S), and pressure using machine learning approaches. As a basis for the surrogate variables, optimally interpolated monthly gridded T/S fields are obtained from the Hadley Centre EN version 4 dataset (hereafter, EN4, Good et al., 2013). It is a global gridded dataset from 1900 to present at the horizontal resolution

of 1°x1° in longitude-latitude grid and with 42 vertical depth levels (20 levels within the 0-1,000m).

2.2 Machine learning algorithms

In supervised learning, a computer program is designed to learn the relationship between a large number of paired input-output examples. In this study, the output (predictand) variable is the O_2 concentration, and the input (predictor) variable can include physical variables and coordinates. The potential predictor variables include absolute salinity, conservative temperature, pressure, potential density, Brunt-Väisälä frequency, longitude, latitude, time, and month. Some of these variables are coordinates and others are derived from the EN4 dataset. It is not clear whether including all above variables will improve the estimation of O_2. There is no one-size-fits-all solution in ML. 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 (NN) are composed of interconnected nodes (neurons) arranged in layers (input, hidden, and output layers). NN is capable of representing complex, nonlinear relationships and capture intricate patterns, but require a large amount of training data. In contrast, Random Forest (RF) is an ensemble learning method that combines multiple decision trees to make predictions. RF can capture complex relationships, but it may struggle with very

193 subtle patterns. RF can handle missing data effectively by using surrogate splits, which means it
194 may outperform NN in data-poor regions. In addition, RF can provide feature importance which
195 can help interpret the results.

196 In this study, we will employ the Scikit-Learn version 1.3 (Pedregosa et al., 2011) for
197 their python implementation of NN and RF regression models. For each type of algorithms, there
198 are several free parameters (hyperparameters) that cannot be learned from the data and must be
199 selected before training. These parameters govern the learning process and influence how the
200 model learns the relationship between the predictor and predictand variables. In practice, it's hard
201 to know in advance which algorithm/hyperparameter set works better for a particular problem,
202 and it requires testing multiple algorithms to make a good model choice by experimentation.
203 Examples of hyperparameters include the number of nodes for each hidden layer in neural
204 networks, the regularization parameter in regression models, or the depth of a decision tree.
205 Hyperparameter tuning involves selecting the best combinations of these settings to achieve the
206 best performance.

207 In oceanographic data, observations always contain some level of noises. Overfitting
208 occurs when an algorithm fits the noises in the training data rather than capturing the signal, and
209 as a result, it negatively impacts its ability to generalize to new, unseen data. Overfitting could
210 occur when a model is too complex relative to the size of the training data and the noise level. To
211 avoid overfitting, 80% of the observed O_2 profiles are used to train the algorithms, and the
212 remaining 20% are withheld as test data to measure how well the trained algorithms can
213 reconstruct the profiles that are not used during the training.

214 K-fold cross validation is used for hyperparameter turning, which is a resampling
215 procedure that helps in estimating how well an algorithm will perform on unseen data. The

training data (80% of oxygen profiles) are randomly split into K 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 group that is withheld from the training in terms of R^2 score. 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.

2.3 Experimental design

Considering various factors discussed in the Section 2.2, a workflow is developed to develop 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 8. All experiments use shipboard O_2 as the predictand variable, and Argo-O_2 is also included in Exp 6 through 8. All experiments also include conservative temperature (T), absolute salinity (S), longitude, latitude, and time as predictor variables. Time is counted as the number of months since January 1965. Exp 2 additionally includes pressure (P), and Exp 3 includes P and month of year (mon) with January being 1 and December being 12. Exp 4 further includes potential density (σ_θ) and Exp 5 additionally include the strength of stratification as the square of Brunt-Väisälä frequency (N^2). 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 the seasonal cycle can be important for O_2 especially in the near-surface layer for biological O_2 production, and because isopycnal surfaces and water column stratification can be important indicators of O_2 ventilation. Comparing Exp 2-5 can inform the importance of

including these additional factors. Exp 6-8 are repetition of Exp 3-5 but with the inclusion of Argo-O_2 data as additional predictand variable.

	T	S	long	lat	time	P	mon	σ_θ	N^2	Argo
Exp 1										
Exp 2										
Exp 3										
Exp 4										
Exp 5										
Exp 6										
Exp 7										
Exp 8										

Table 1. Different combinations of input/output variables. "T" is conservative temperature (°C). "S" is absolute salinity (g/kg). "long" is longitude and "lat" is latitude, both in degrees. "P" is pressure (dbar). "σ_θ" is potential density (kg/m^3), and "N" is Brunt-Väisälä frequency (s^{-1}). "time" is measured as the number of month since January 1965. "mon" is the month of year.

Two types of algorithms, NN and RF are trained for each experiment (Exp1-8). For each algorithm, a suite of hyperparameters sets is considered (12 sets for NN and 18 sets for RF), thus a total of 240 algorithms are trained for different combinations of algorithm type, hyperparameter sets, and input/output parameter choices. For NN algorithm, the number of nodes in hidden layers and the regularization parameter are systematically changed (see **Table 2**). Four sets of hidden layers are considered including 5-5-5-5, 10-10-10-10, 20-20-20-20, and 40-40-40-40, and three different regularization parameters are considered including 0.001, 0.01 and 0.1. Increasing the number of nodes allows more complexity whereas increasing the

regularization parameter prevents the model from becoming too complex. The combination of hyperparameters results in 12 different configurations of the NN algorithm.

	regularization	hidden layers
HP set 1	0.001	5-5-5-5
HP set 2	0.001	10-10-10-10
HP set 3	0.001	20-20-20-20
HP set 4	0.001	40-40-40-40
HP set 5	0.01	5-5-5-5
HP set 6	0.01	10-10-10-10
HP set 7	0.01	20-20-20-20
HP set 8	0.01	40-40-40-40
HP set 9	0.1	5-5-5-5
HP set 10	0.1	10-10-10-10
HP set 11	0.1	20-20-20-20
HP set 12	0.1	40-40-40-40

Table 2. A list of hyperparameters for Neural Network algorithm.

For RF algorithm, different configurations are explored (see **Table 3**) for the number of trees (number of estimators), the minimum number of samples required for a leaf node (minimum samples leaf), and the maximum number of features for split at each tree node (max features). Greater number of trees avoids overfitting and stabilizes the algorithm, and it is varied from 100 to 200 to 500. Increasing minimum samples leaf controls the growth of trees and prevents overfitting, and it is varied from 1 to 6 to 12. Limiting maximum features decorrelates trees and helps to prevent overfitting, and it is varied from 1 to 3. The combination of these hyperparameters results in 18 different configurations of the RF algorithm.

	max features	min sample leaf	n of estimators
HP set 1	1	1	100
HP set 2	1	1	200
HP set 3	1	1	500
HP set 4	1	6	100
HP set 5	1	6	200
HP set 6	1	6	500
HP set 7	1	12	100
HP set 8	1	12	200
HP set 9	1	12	500
HP set 10	3	1	100
HP set 11	3	1	200
HP set 12	3	1	500
HP set 13	3	6	100
HP set 14	3	6	200
HP set 15	3	6	500
HP set 16	3	12	100
HP set 17	3	12	200
HP set 18	3	12	500

Table 3. A list of hyperparameters for Random Forest algorithm

The best performing algorithm is selected after training all possible combination of hyperparameters for each combination of input/output variables and algorithm type using R^2 value as the performance metric. 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 highest performing algorithms are identified, and the gridded O_2 datasets are generated by projection of gridded predictor variables for further validation and analysis.

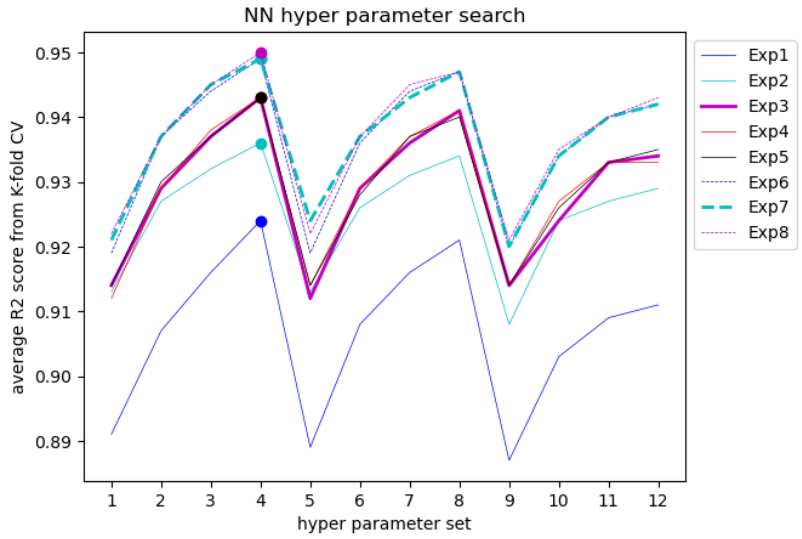
3. Results: hyperparameter tuning and performance evaluation

A total of 240 ML algorithms is trained including 96 NN and 144 RF regression models based on different combinations of input/output variables and hyperparameter sets. Each of the 240 algorithms is trained 5 times using K-fold cross validation approach, thus the total of 1,200 trainings were performed. These calculations were computationally demanding but it can be efficiently carried out in the parallel computing platform with large memory using Cheyenne/Casper 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 K-fold cross validation approach (K=5), and the mean R^2 scores are recorded. **Figure 2** shows that the variation of the mean score for the NN algorithm with the hyperparameter sets listed in **Table 2**. Overall, the NN algorithms with adequate input data (Exp 3-8) were capable of reproducing O_2 observations withheld from the training with very high skills, and the inclusion of Argo-O_2 data further increased the skill. Each line comes from the same set of input/output variables (Exp 1-8 in **Table 1**) and the variation of the R^2 scores is consistent among all experiments with some constant offset.

The sensitivity of algorithm performance to the choice of hyperparameter sets is largely independent of the specific choice in the input/output variables, but the overall performance itself significantly depend on the choice of input/output variables. The peak performances consistently occurred for the 4^{th} hyperparameter set with the smallest regularization and highest complexity (number of nodes).



306

307 **Figure 2.** Mean R^2 scores from K-fold cross validation (K=5) using Neural Network algorithm.
308 Each line color represents different combinations of input/output variables from Table 1. Results
309 from Exp6-8 includes Argo-O_2 data are shown in dashed lines. The dots indicate the best
310 performing hyperparameter for each input/output variable set.

311

312 Exp 1 includes the least number of input variables (T, S, long, lat, and time) and showed
313 the lowest skill with the highest score of $R^2 \sim 0.92$. Even though Exp 1 is the weakest case, the
314 algorithm was still able to reproduce 92% of variance in the data withheld from the training,
315 which is encouraging. Exp 2 includes additional input data of pressure, and it increased the
316 performance to $R^2 \sim 0.93$. Exp 3-5 additionally includes input variable of month (Exp 3), month
317 and σ_θ (Exp 4), and month, σ_θ and N^2 (Exp 5). These cases shared essentially the identical
318 performance score of $R^2 \sim 0.94$. The additional inputs of potential density (σ_θ) and stratification
319 (N^2) apparently did not increase the R^2 score. Exp 6-8 additionally included the Argo-O_2 data for
320 the predictand while mirroring the same input parameter sets for Exp 3-5. The R^2 score of Exp 6-

8 are essentially identical, and showed the highest scores of $R^2 \sim 0.95$, indicating the benefit of additional training data from the Argo-O_2.

Figure 3 shows the R^2 score of RF algorithms with the hyperparameter sets listed in **Table 3**. Similar to NN, the RF algorithms with fewer input data (Exp 1-2) performed relatively poorly. Cases with adequate input data (Exp 3-8) demonstrate improved performance in reproducing the O_2 observations withheld from the training ($R^2 \sim 0.97$). Similar to the NN algorithms, the inclusion of Argo-O_2 data improved the skill (dashed lines in **Figure 3**). Overall, the R^2 scores are generally higher than the NN algorithms. Better performances were found with the minimum samples leaf of 1. In particular, the best score was achieved with maximum features of 3, minimum samples leaf of 1, and number of estimators of 500. This parameter choice involves a trade-off between model complexity and overfitting. The best performing algorithms in RF algorithm group was Exp6 with relatively fewer input variables (T, S, lon, lat, time, pressure and month). As with the NN algorithms, additional variables such as potential density or stratification did not improve the skill.

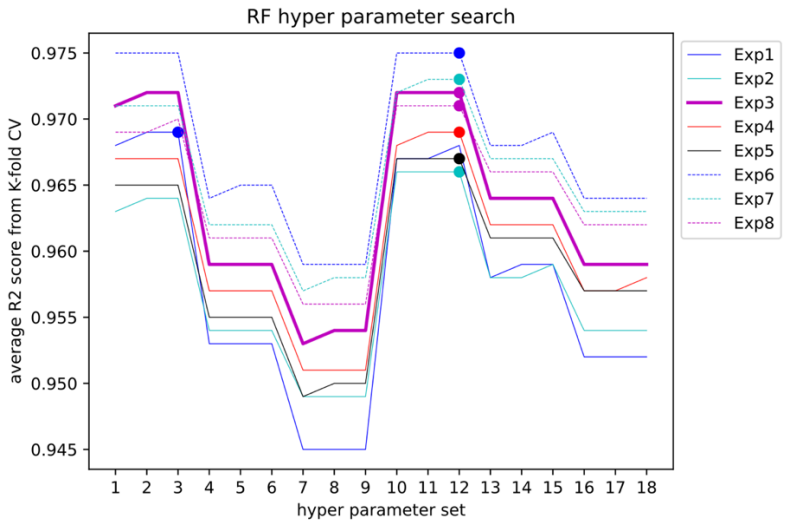


Figure 3. Same as Figure 1 but for the Random Forest algorithm. The dots indicate the best performing hyperparameter.

338

339 **3.2 Validation and quantification of uncertainties using the test data**

340 The test data consists of 20% of all input/output data that are set aside and unused for
341 training algorithms, including approximately 230,000 data points. These test data are used to
342 further evaluate the algorithms and to quantify the uncertainties. We selected the best performing
343 hyperparameter sets for NN and RF algorithms for each of the experiments and examined the
344 performance using three metrics including mean bias, root-mean-square error (RMSE) and
345 correlation coefficient (R) and the results are listed in **Table 4**.

346

	Neural network			Random Forest		
	Bias (umol/kg)	RMSE (umol/kg)	R	Bias (umol/kg)	RMSE (umol/kg)	R
Exp1	-0.49	16.21	0.96	0.10	9.57	0.99
Exp2	1.72	14.72	0.97	0.01	9.92	0.98
Exp3	-0.99	14.08	0.97	0.01	9.05	0.99
Exp4	-0.11	13.92	0.97	0.04	9.52	0.99
Exp5	0.40	14.03	0.97	0.02	9.82	0.98
Exp6	-0.97	13.42	0.97	0.04	8.93	0.99
Exp7	0.40	13.28	0.97	0.02	9.43	0.99
Exp8	0.86	13.26	0.97	0.03	9.66	0.99

347

348 **Table 4.** Uncertainty estimation of 16 algorithms for each experiment listed in **Table 1**. For
349 each experiment, mean bias, RMSE and R values are reported for NN and RF algorithms with
350 the best performing hyperparameter sets.

351

352 For each set of input/output variables (experiments), RF algorithms showed lower mean
353 bias, lower RMSE, and higher R value, indicating somewhat better skill. Comparing the

algorithms trained with shipboard only (Exp 3-5) and shipboard and Argo-O_2 data (Exp 6-8), there is no clear difference in terms of bias, RMSE or R values. The magnitude of the mean bias from the NN algorithms is less than 2μmol/kg, and that of RF algorithms is less than 0.1μmol/kg. The R values are about 0.96-0.97 for the NN algorithms and that of RF algorithms are about 0.98-0.99. The values of RMSE are useful estimates of the uncertainties due to gap filling using these algorithms. RMSE of the NN algorithms are in the range of 13 to 16μmol/kg, and that of the RF algorithms is less than 10 μmol/kg. Similar to results from the previous section, Exp1 and 2 shows slightly weaker performances relative to other experiments.

In comparison to a recently developed global dataset, GOBAI-O_2 (Sharp et al., 2023), they found the global-scale RMSE of 8.8μmol/kg which is similar but slightly less than our RF algorithm. GOBAI-O_2 employs similar neural network and random forest algorithms under different configurations, data sources mainly based on Argo-O_2 (with additional GLODAPv2 profiles), more recent period (2004-present), and importantly, their analysis covers the global domain. Thus, we do not expect the same uncertainties, but our results are indeed on the same magnitudes.

3.3 Evaluation of climatological O_2 distribution

Using the algorithms developed and tested in Section 2.2, we projected O_2 distributions using the gridded EN4 data for the North Atlantic from 1965 to 2010, and we further analyze the results in comparison to the well-established climatological distribution using World Ocean Atlas 2018 (WOA18). **Figure 4** shows the summary of comparison for annual mean climatology at five depth levels including 10m, 100m, 200m, 400m and 700m. This is not a validation since the shipboard data used to assemble World Ocean Atlas were also used in the training of the

algorithms, however, it is reassuring to find similar climatological distribution to the widely adopted WOA18. This comparison show that the Exp 1 has a qualitatively different representation of climatology than all other cases with the largest discrepancies with WOA18.

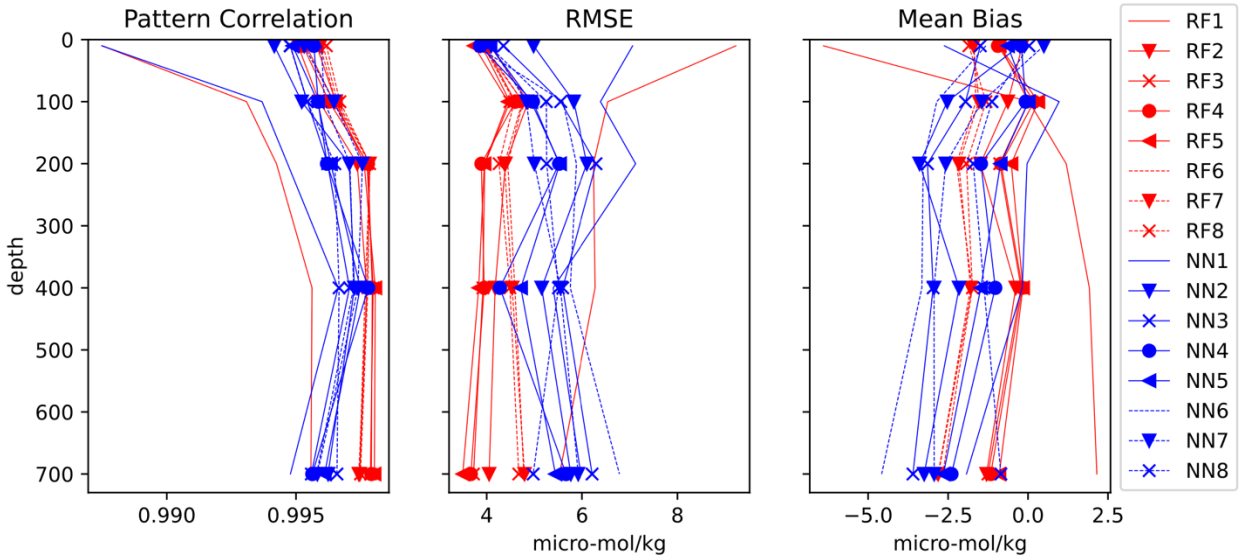


Figure 4. Pattern correlation (R), RMSE and mean bias of the annual mean climatology. Blue lines are NN algorithms and red lines are RF algorithms. Each line and dots are indicating experiments with different input/output variable sets. Dash line indicates experiments including both shipboard and Argo-O_2 data.

Comparing the NN and RF algorithms for Exp 2-8, the RF algorithms (red lines in **Figure 4**) performs slightly better than the NN (blue lines in **Figure 4**), where RF shows higher pattern correlation (>0.995) and smaller RMSE (<5μmol/kg). The results from NN are more variable and show slightly lower pattern correlation and higher RMSE. The mean bias of climatological distributions are generally negative with the exception of Exp 1, indicating that reconstructed O_2 climatologies with ML approaches are slightly lower than WOA18. The inclusion of Argo-O_2 data further enhances the negative bias of the climatological O_2 profile.

393 Factors contributing to the negative mean bias may include differences in time period
394 represented by the shipboard observations and Argo-O_2 datasets. There are greater number of
395 shipboard O_2 profiles during 1980s than later periods, and WOA18 is based on bottle data. The
396 period represented by the ML-based climatology may reflect the time windows over which the
397 training data were collected. The inclusion of Argo-O_2 data mostly sampled after 2010 could
398 result in different climatology than that trained by the bottle observations centered around 1980s.
399 The representations of the temporal trends are further examined in **Section 4**.
400 Seasonal O_2 amplitudes are important indicators of thermally-induced solubility changes
401 as well as the biological O_2 production, and are examined among the 16 O_2 data products (NN
402 and RF for each of Exp 1-8) as the difference between mean JJA and mean DJF climatologies.
403 **Figure 5** and **6** shows the seasonal amplitude of O_2 at four different depths from the RF and NN
404 algorithm respectively. At the surface (10m), there is a strong negative anomaly in the central
405 subtropics and at mid-latitudes according to WOA18. There is also a weak positive anomaly in
406 the subpolar region in WOA18. At 100m depth, the subtropics shows a positive anomaly, and the
407 subpolar region shows a negative anomaly. These patterns are reasonably well captured in the
408 Exp3-8. The algorithm may underestimate the amplitude of subsurface (400-700m) seasonality,
409 while WOA18 also shows significant noises there.
410

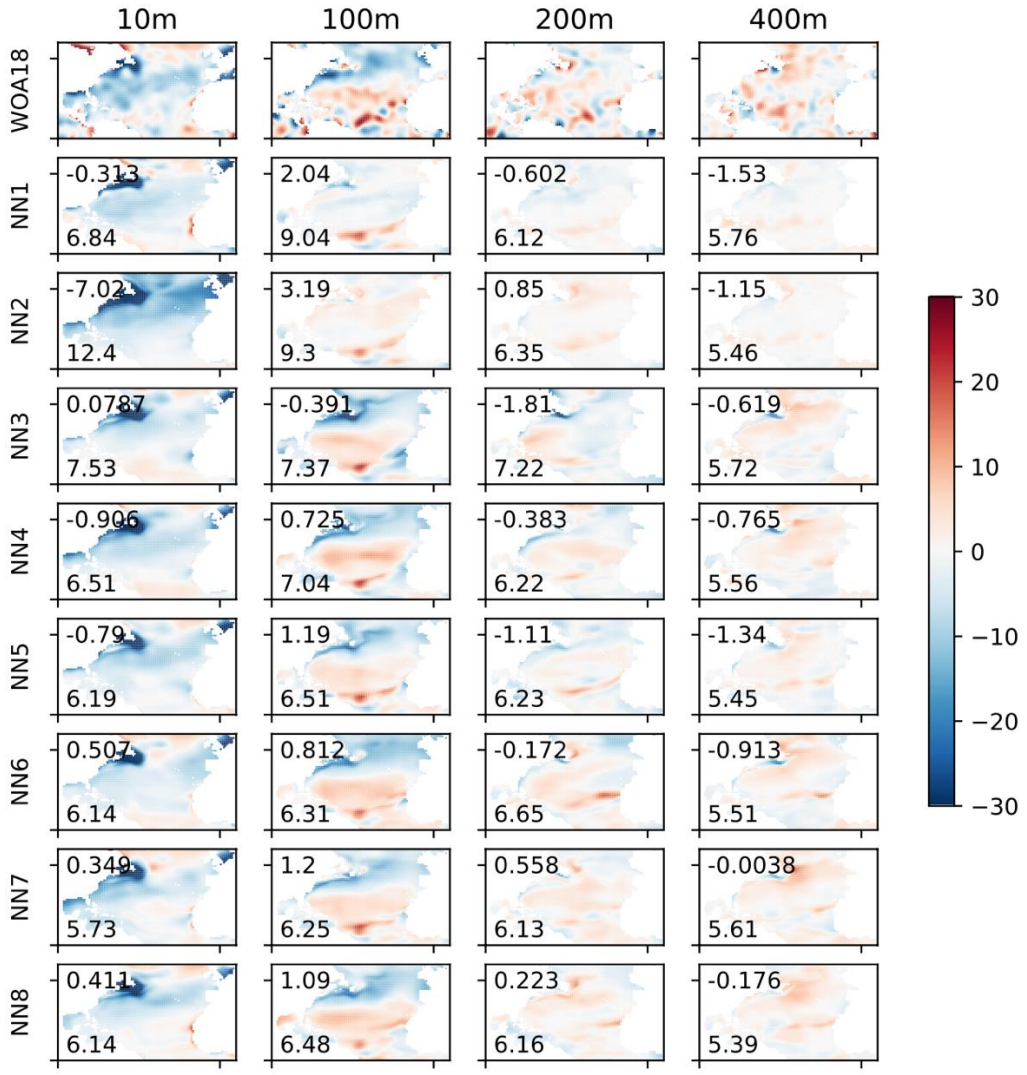


Figure 5. Summer (JJA) minus Winter (DJF) climatological O_2 plotted at 10m, 100m, 200m and 400m depth. The top row is WOA18, and the second row and below are from the 8 experiments with NN algorithm. Positive value means the summertime O_2 level is higher than the wintertime values. Upper left corner shows the mean bias and the lower left is the RMSE, both in μmol/kg.

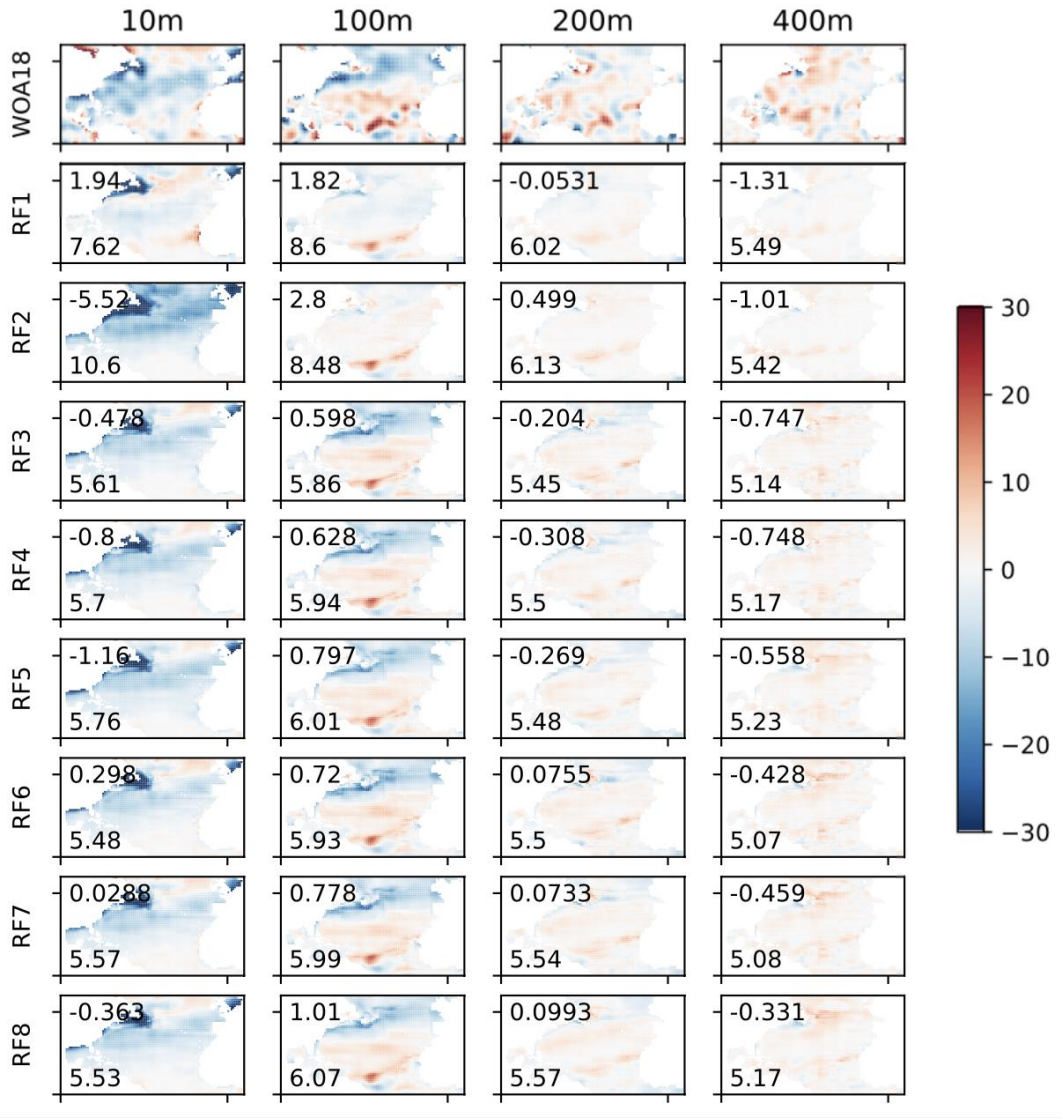


Figure 6. Same as Figure 6 but for RF algorithm.

Both NN and RF algorithms with Exp 1-2 performed differently from other cases with significantly weaker seasonal variability in the subsurface water, and greater magnitudes of mean bias and RMSE. These two cases lack pressure and month from the predictor variables, which are likely important factors for the O_2 seasonal cycle. While it was difficult to detect this bias from the validation with the test data, we conclude that Exp 1-2 performed significantly poorer than Exp 3-8 in terms of representing the mean seasonal cycle correctly, thus the inclusion of

pressure and month in predictor variables is important. Beyond this, there was no clear differences in terms of performance with different predictor variable choices. The addition of potential density and/or stratification did not significantly improve the performance. Based on the comparison with WOA18, both RF and NN algorithm with Exp 3-8 performed well for reproducing the annual mean climatology as well as the contrast between the summer and winter months.

3.4 Feature importances

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 7** shows the feature importances determined from the Exp 1-8 with best performing hyperparameter sets. Across all the cases, latitude was considered the most influential variable in making O_2 estimation. Following the latitude, temperature and salinity are also important factors for all cases. When pressure is included, it played significant role throughout, sharing similar weight as salinity. Other variables, such as potential density, stratification, time/month all played some roles when they are included as input variables with relatively small influences.

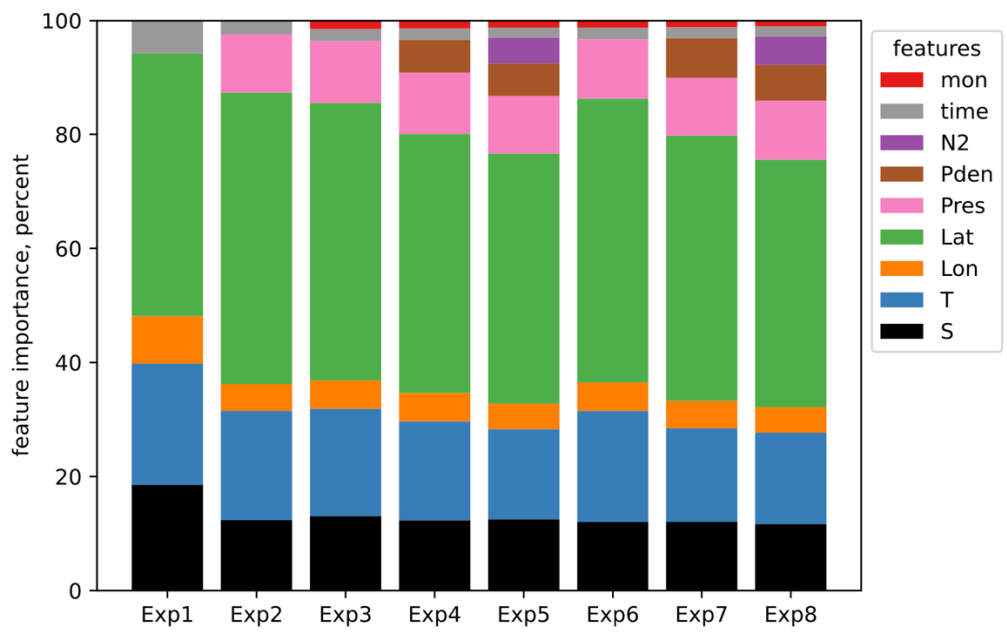


Figure 7. Feature importances of the Random Forest algorithm. The relative importance of each feature variables are shown for each experiment.

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. Temperature and salinity are both important factors. Comparing Exp 3 and 4 (and Exp 6 and 7), the addition of potential density did not necessarily reduce the relative importance of T/S. Rather the algorithm mainly reduced the importance of latitude. Variability of T/S on isopycnal surfaces can indicate water mass shifts and circulation variability, thus these variables can play some roles in estimating O_2 variability. Similarly, comparing Exp 4 and 5 (and Exp 7 and 8), the further addition of N^2 does not significantly reduce the importance of $T/S/\sigma_\theta$, indicating some roles played by the stratification and its variability. It is important to note that feature importances are calculated for the specific

459 configuration of RF algorithms used in this study, and they may not indicate causal relationships
460 and the interpretation requires caution.
461
462 **4. Results: deoxygenation trends**
463 Based on the comparison with the annual mean and seasonal climatology, we consider
464 both NN and RF algorithms with the input datasets from Exp 3-8 to provide reasonable
465 reconstructions of the O_2 distribution, forming 12 ensemble members (NN 3-8 and RF 3-8)
466 where numbers after NN and RF indicates the experiment number in **Table 1**. In other words,
467 results from Exp 1-2 are excluded due to their relatively weak performances in reproducing the
468 annual mean climatology and the climatological seasonal cycle. Top panel in **Figure 8** shows the
469 deseasoned O_2 inventory time series integrated over 0-1,000m as anomalies from the ensemble
470 mean climatological O_2 inventory of 5.93×10^{15} mol. Results from all algorithms show a
471 moderate increase from 1965 to around 1990, followed by a significant decline after 1990. The
472 O_2 inventories calculated by the NN algorithms show more diverse trajectories relative to that of
473 RF algorithms after 1990. In general, the O_2 inventories from NN algorithms decline more
474 strongly than the RF algorithms after 1990. The range of O_2 inventories estimated from the
475 shipboard data only are grouped together in green, and that from the shipboard and Argo-O_2 data
476 are in blue. The envelope is the range bounded by the maximum and minimum values of the 6
477 ensembles for each case (NN 3-5/RF 3-5 in green, and NN 6-8/RF 6-8 in blue). The black and
478 magenta lines are independent estimates of O_2 inventory anomalies based on optimal
479 interpolation of WOD18 shipboard profiles (bottle and CTD-O_2, Ito et al., 2023) and GOBAI-O_2
480 dataset (Sharp et al., 2023).

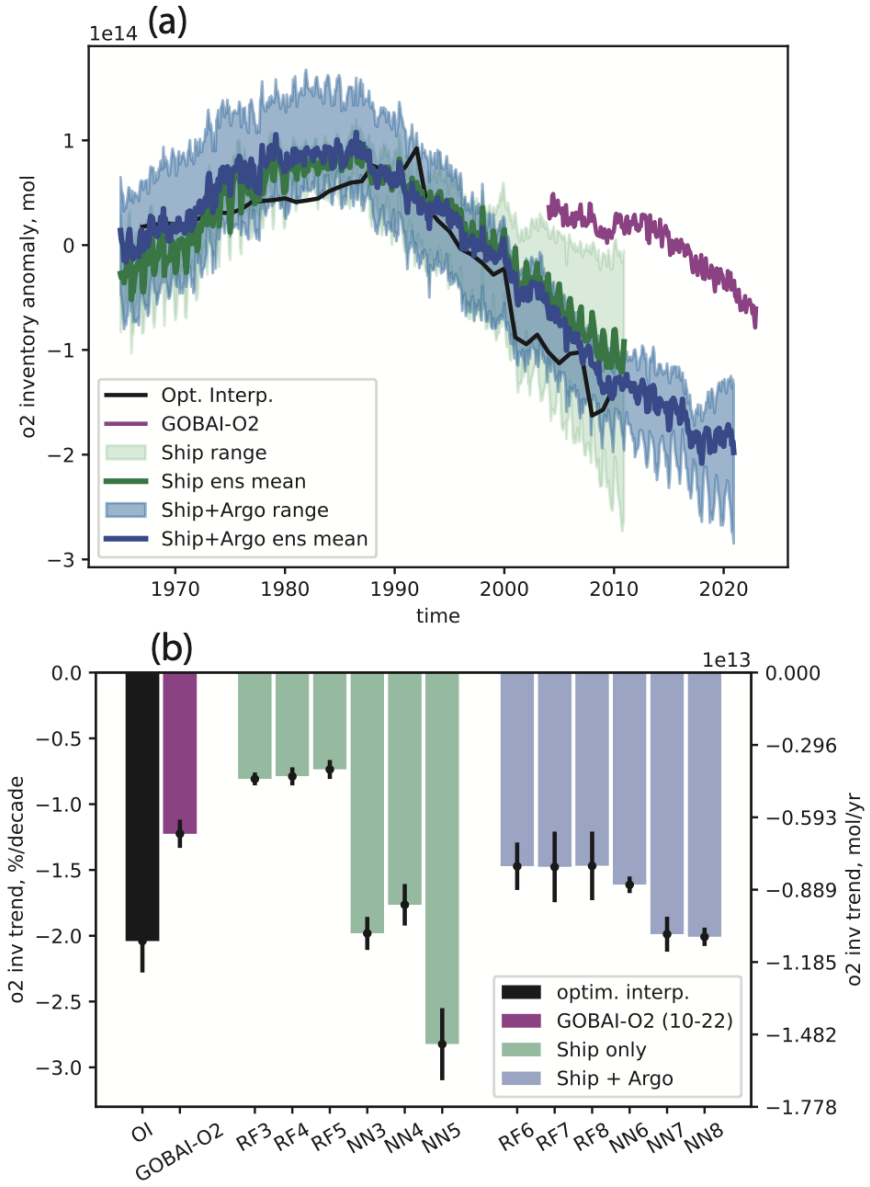


Figure 8. (a) Oxygen inventory anomalies in the units of mol. Two algorithms trained with (green) shipboard data only and (blue) shipboard and Argo-O_2 data. The black, solid line is based on optimal interpolation of WOD profiles (Ito et al., 2023), and magenta line is based on GOBAI-O_2 (Sharp et al., 2023). (b) The magnitudes of linear trend from 1990 to 2010 from the 12 ensembles with 95% confidence interval. Blue bars are based on the shipboard data only and orange bars are based on the shipboard and Argo-O_2 data.

The solid lines are the respective ensemble means. The ensemble means are generally similar and in general agreement with the optimal interpolation. The GOBAI-O_2 is primarily based on Argo data for the period after 2004, thus its climatological mean O_2 is different from all other datasets. Vertical position of the magenta line in **Figure 8a** is referenced to its own climatology for 2004-2022, and we focus on its temporal variation not the mean value. An important difference between the two groups with and without the Argo-O_2 is their respective range. Estimates based on the shipboard and Argo-O_2 data (blue envelope) maintains similar range throughout the period from 1990 to 2020. However, the estimates from the shipboard data only (green envelope) are diverging after 2000, likely due to the lack of constraints near the end of the time series.

Figure 8b shows the linear trends of O_2 inventories from optimal interpolation, GOBAI-O_2, and from the two groups. The period of the trend analysis is from 1990 to 2010 except for GOBAI-O_2 with the period of 2010 to 2022. The optimal interpolation estimated the deoxygenation of 2.0% per decade, and GOBAI-O_2 estimated 1.2% per decade. Estimates based on the shipboard data only are in green, and those with shipboard and Argo-O_2 data are in blue. The estimated deoxygenation rates vary from 0.7 to 2.8% per decade based on the shipboard data only for the same period. However, when the Argo-O_2 data is included, the estimated range of deoxygenation rates are constrained in the range of 1.5 to 2.0% per decade. Ensemble mean deoxygenation rates did not show significant difference between the two cases, but the inclusion of the Argo-O_2 data narrowed the range of estimated deoxygenation rate by a factor of 4.2, which is a remarkable improvement. As shown in **Figure 8b**, the RF algorithms estimated the weakest deoxygenation rates and the NN estimated the strongest trends when they are trained with the

511 shipboard data only. When Argo-O_2 data are included, the deoxygenation rates from RF became
512 stronger and that from NN became weaker, converging towards a much narrower range.

513 **Figure 9** shows the zonal mean O_2 trends in the upper 1,000m of the North Atlantic basin
514 from 1990 to 2010. The panel (a,c,e) shows the zonal mean trend of O_2, (-1) x *AOU*, and O_2
515 solubility. O_2 solubility is a function of salinity and temperature where the solubility coefficients
516 are derived from the data of Benson and Krause (1984) as fitted by Garcia and Gordon (1992).
517 *AOU* stands for apparent oxygen utilization, and is defined as the difference between O_2
518 solubility and O_2. **Figure 9b** shows the climatological annual mean O_2, showing the high O_2
519 water column around 60°N and the ventilated subtropical thermocline in the subtropics, and
520 oxygen minimum zone (OMZ) in the tropical thermocline. The oxygen loss occurs in several hot
521 spots. At subpolar latitudes around 60°N, a strong O_2 decline occurs in the upper water column
522 due to the decline of solubility (**Figure 9ae**). **Figure 9d** displays the R^2 value of the linear trend,
523 which measures the fraction of O_2 variance explained by the linear trend. In this figure, the high
524 R^2 value means that the temporal variability is dominated by the trend. At subtropical and low
525 latitudes, O_2 trends are primarily driven by (-1) x *AOU*, at the base of the ventilated thermocline
526 and the boundary between subtropics and tropical OMZ where the expansion of the tropical
527 OMZ has been documented and discussed extensively (Stramma et al., 2008). Examination of
528 zonal mean trends for individual ensemble members are generally similar, while there are some
529 disagreements in the detailed spatial structure. Most importantly, the overall magnitude of the
530 trends is weaker in the RF algorithm that that of NN.

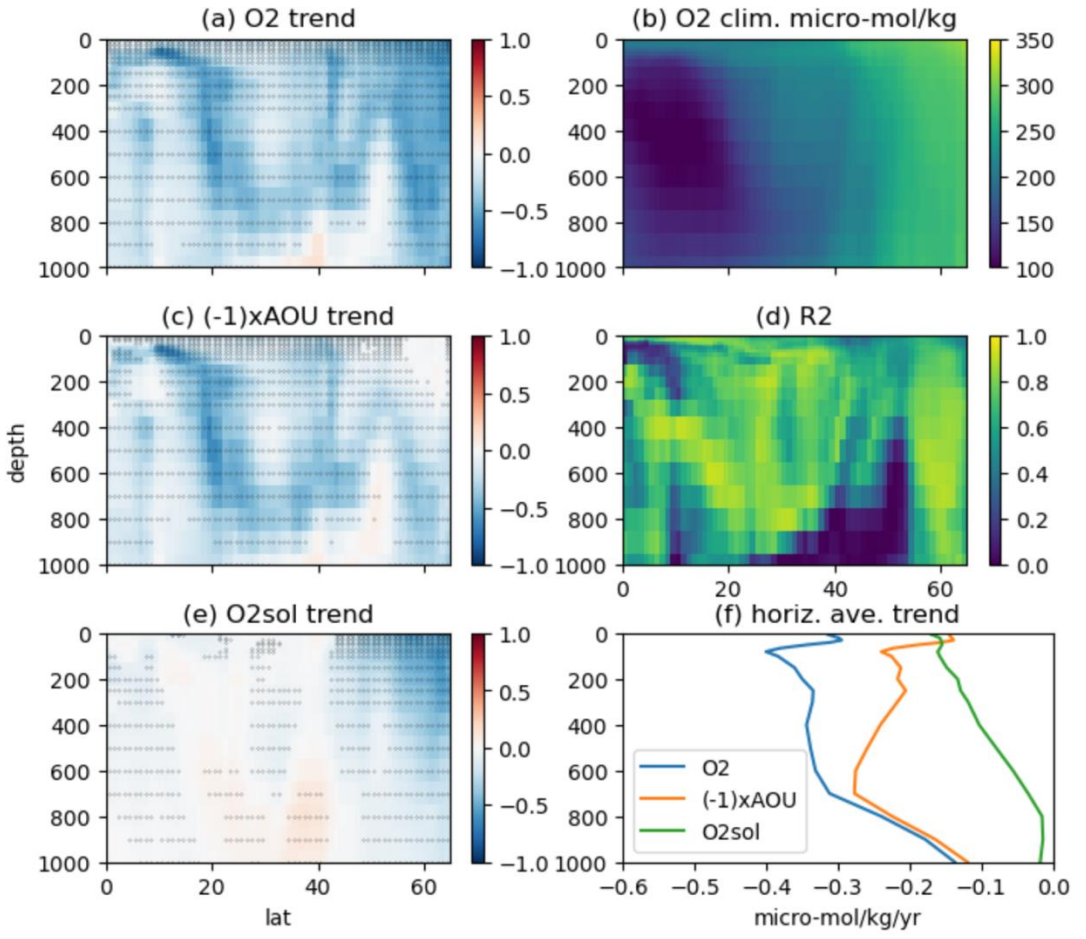


Figure 9. Ensemble mean, zonal mean trend of O_2, *AOU* and O_2 solubility. (a,c,e) Zonal mean trends in the units of μmol/kg/year. Dots indicate statistical significant trend at 95% confidence interval. (b) O_2 climatology, (d) R^2 value of O_2 trend, and (f) horizontally averaged trend.

In the low latitude deoxygenation, there is no strong temperature increase (nor solubility decline) and these low-latitude trends are predominantly caused by *AOU* changes though circulation, water mass shifts, mixing and/or biochemical processes as shown in **Figure 9ce**. Finally, **Figure 9f** shows the horizontally averaged trend. As expected, the surface trend is primarily driven by the O_2 solubility and the *AOU* trend increases its importance in the subsurface waters, and it becomes the dominant mode of O_2 loss in the main thermocline.

5. Results: uncertainty analysis

There are 3 levels of uncertainty including measurement error, sampling error and mapping (interpolation) error, and for each level, there can be random errors and biases. Measurement errors depend on specific techniques and instrumentation for making measurements. For example, bottle O_2 can include random errors of 1 μmol/kg with Winkler titration, whereas delayed-mode Argo-O_2 has errors of about 3 μmol/kg. In the oxycline region, there can be a larger error O(10 μmol/kg) for Argo-O_2 data due to uncorrected sensor response time, potentially including random and systemic bias components.

Sampling errors can be estimated by the standard deviation of monthly binned data. **Figure 10** shows the non-uniform distribution of this uncertainty. The mean value of the standard deviation of monthly binned data is 4.5 μmol/kg for the whole basin but its value can exceed 20μmol/kg in regions such as Scotia and Newfoundland shelves. 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.

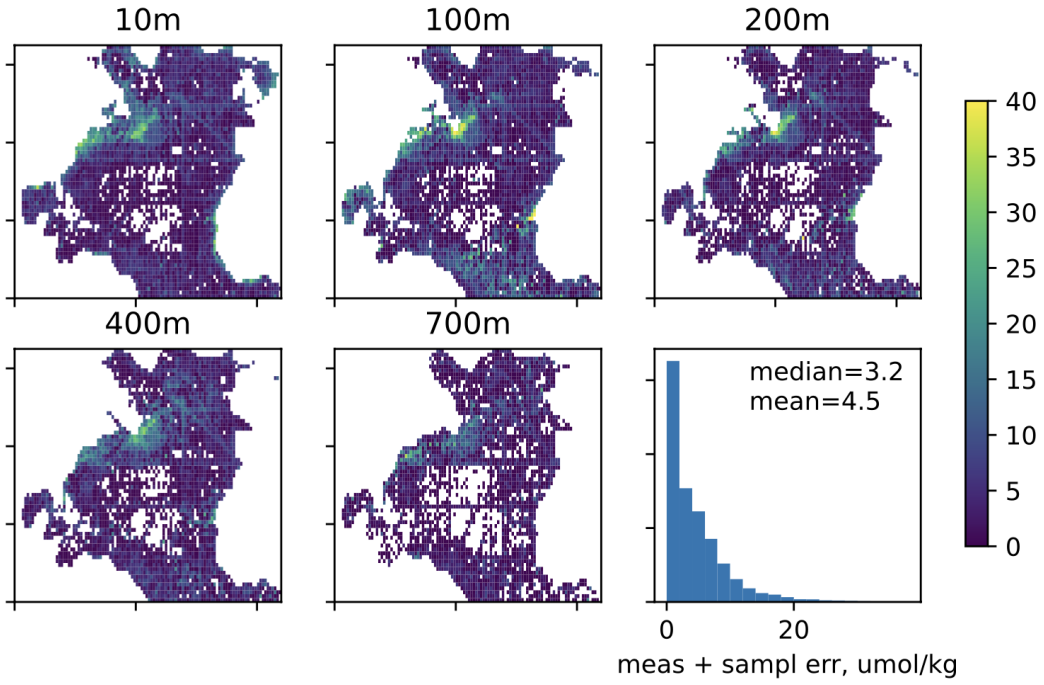


Figure 10. An estimate of the sum of measurement and sampling errors based on the standard deviation of binned data for 1°x1° monthly grid cells at 10m, 100m, 200m, 400m and 700m. The units are in µmol/kg.

Mapping uncertainties can be estimated by the comparison with the O_2 data withheld from the training as documented in section 3.2. The O_2 values estimated from NN algorithms had the RMSE of 13.3 to 14.1 µmol/kg and that of RF algorithms are in the range of 8.9 to 9.8 µmol/kg. These are overall estimates of the mapping/interpolation errors in this study. These error estimates are comparable to "algorithm errors" for the GOBAI-O_2 dataset of Sharp et al., (2023).

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:

572 $$\Delta O_2 = \left\{ \Delta O_{2meas}^2 + \Delta O_{2sampl}^2 + \Delta O_{2interp}^2 \right\}^{1/2}$$

573

574 Based on the typical magnitudes of these errors as discussed above, the combined uncertainty is
575 15 μmol/kg for NN and 10 μmol/kg for RF algorithm, primarily dominated by the
576 mapping/interpolation error. In the Scotia and Newfoundland shelves, the combined uncertainty
577 can be significantly higher.

578

579 **4. Discussion and conclusion**

580 Since the mid-2000s, Argo floats equipped with O_2 sensors have been deployed in
581 different parts of the global oceans, and the development of in-situ calibration methods reduced
582 the measurement uncertainties of the Argo-O_2 sensors to approximately 3 μmol/kg.
583 Coincidentally the number of shipboard observations has decreased in the recent decades, and as
584 a result, it is difficult to estimate the basin-scale deoxygenation trends for recent periods based
585 on shipboard observation only. Recently, a gridded, time-varying O_2 product has been developed
586 using ML approaches (Sharp et al., 2023), reconstructing the global O_2 distribution since 2004.
587 There are a few notable similarities and differences between GOBAI-O_2 and this regional study.
588 A unique feature in this study was to make a contrast between the O_2 datasets based on shipboard
589 data only versus the synthesis of historical shipboard measurements and the new Argo-O_2 data.
590 Thus, we included historical observation from an earlier period since 1965, allowing to evaluate
591 deoxygenation trends over a longer period. Results from each of the ensemble members with and
592 without Argo-O2 data are available in public domain from zenodo (Ito and Cervania, 2023).
593 These gridded data products will be helpful for validating computational biogeochemistry

models as only a few temporally and spatially varying O_2 datasets are currently available in the public domain.

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 only, but we further limited to the O_2 profiles calibrated with two well-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 due to the offset of -1.18 μmol/kg based on 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 models. In this study, we trained a large number of algorithms (240 cases) with varying sets of input data and hyperparameters and selected 12 algorithms with high skills to form an ensemble of O_2 estimates. Despite these differences, the resulting O_2 inventory anomalies shared generally similar trend for 2010s (see **Figure 8**). This study focused on the North Atlantic basin which was sampled most densely and frequently in the historical observations, and ML algorithms were trained using relatively abundant sample numbers. The ML approach remains to be tested in other, less frequently sampled basins using the combination of historical and Argo-O_2 observations. The success of GOBAI-O_2 in generating a global dataset is indeed encouraging that synthesis of shipboard and Argo-O_2 data is indeed possible for other basins as well.

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 for the most part of the North Atlantic with the overall magnitude of 10-14 μmol/kg. There is an exception with the relatively high sampling error near the western boundary regions

such as Gulf Stream, Scotia and Newfoundland shelves. These regions exhibit strong natural variability that can generate similar or even larger uncertainties than the interpolation errors.

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, especially in the coastal waters. 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 at much higher resolution.

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. The results from this study are available from zenodo (Ito and Cervania, 2023, doi:10.5281/zenodo.10430869). Optimal interpolation of WOD18 O_2 data is available from zenodo (Ito, 2023, doi:10.5281/zenodo.10367379).

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., & 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., 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.

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

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

681 Garcia, H. E., Weathers, K., Paver, C. R., Smolyar, I., Boyer, T. P., Locarnini, R. A., et al.
682 (2018). *World Ocean Atlas 2018, Volume 3: Dissolved Oxygen, Apparent Oxygen*
683 *Utilization, and Oxygen Saturation*., NOAA Atlas NESDIS, Silver Springs, MD.

684 Gloege, L., McKinley, G. A., Landschützer, P., Fay, A. R., Frölicher, T. L., Fyfe, J. C., et al.
685 (2021). Quantifying Errors in Observationally Based Estimates of Ocean Carbon Sink
686 Variability. *Global Biogeochemical Cycles, 35*(4).

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

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

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

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

Ito, T. (2023). Optimally interpolated dissolved oxygen based on the World Ocean Database 2018 and CMIP6 models [Data set]. In Biogeosciences. Zenodo. https://doi.org/10.5281/zenodo.10367379.

Ito, T. and A. Cervania (2023). Machine-Learning for O2 (ML4O2) Project, North Atlantic ensembles [Data set]. In Journal of Geophysical Research Oceans, Zenodo. https://doi.org/10.5281/zenodo.10430869.

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.

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.

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

721 Moussa, H., Benallal, M. A., Goyet, C., & Lefèvre, N. (2016). Satellite-derived
722 CO₂ fugacity in surface seawater of the tropical Atlantic Ocean using a
723 feedforward neural network. *INTERNATIONAL JOURNAL OF REMOTE SENSING*,
724 *37*(3), 580-598.

725 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011).
726 Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*,
727 2825-2830.

728 Pershing, A. J., Alexander, M. A., Hernandez, C. M., Kerr, L. A., Le Bris, A., Mills, K. E., et al.
729 (2015). Slow adaptation in the face of rapid warming leads to collapse of the Gulf of
730 Maine cod fishery. *Science, 350*(6262), 809-812.

731 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat.
732 (2019). Deep learning and process understanding for data-driven Earth system science.
733 *Nature, 566*(7743), 195-204.

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

737 Sauzède, R., Bittig, H. C., Claustre, H., de Fommervault, O. P., Gattuso, J. P., Legendre, L., &
738 Johnson, K. S. (2017). Estimates of Water-Column Nutrient Concentrations and

739 Carbonate System Parameters in the Global Ocean: A Novel Approach Based on Neural
740 Networks. *Frontiers in Marine Science, 4*.

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

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

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

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

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

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

760 Zeng, J. Y., Nojiri, Y., Nakaoka, S., Nakajima, H., & Shirai, T. (2015). Surface ocean CO_2 in
761 1990-2011 modelled using a feed-forward neural network. *GEOSCIENCE DATA*
762 *JOURNAL, 2*(1), 47-51.

763

764