

# Journal of Geophysical Research: Oceans

## RESEARCH ARTICLE

10.1029/2017JC013404

### Special Section:

The Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) Project: Technologies, Methods, and Early Results

#### Key Points:

- Modern statistical and machine learning tools can be used to map  $O_2$  where T/S, and sufficient  $O_2$  data are available
- The Southern Ocean State Estimate and the World Ocean Atlas climatology may overestimate the annual mean  $O_2$  at 150 m in the Southern Ocean
- The Southern Ocean State Estimate may underestimate the annual cycle of oxygen at 150 m in the Southern Ocean

#### Supporting Information:

- Supporting Information S1

#### Correspondence to:

D. Giglio,  
dgiglio@ucsd.edu

#### Citation:

Giglio, D., Lyubchich, V., & Mazloff, M. R. (2018). Estimating oxygen in the Southern Ocean using Argo temperature and salinity. *Journal of Geophysical Research: Oceans*, 123, 4280–4297. <https://doi.org/10.1029/2017JC013404>

Received 31 AUG 2017

Accepted 5 APR 2018

Accepted article online 30 APR 2018

Published online 30 JUN 2018

## Estimating Oxygen in the Southern Ocean Using Argo Temperature and Salinity

D. Giglio<sup>1</sup> , V. Lyubchich<sup>2</sup> , and M. R. Mazloff<sup>1</sup> 

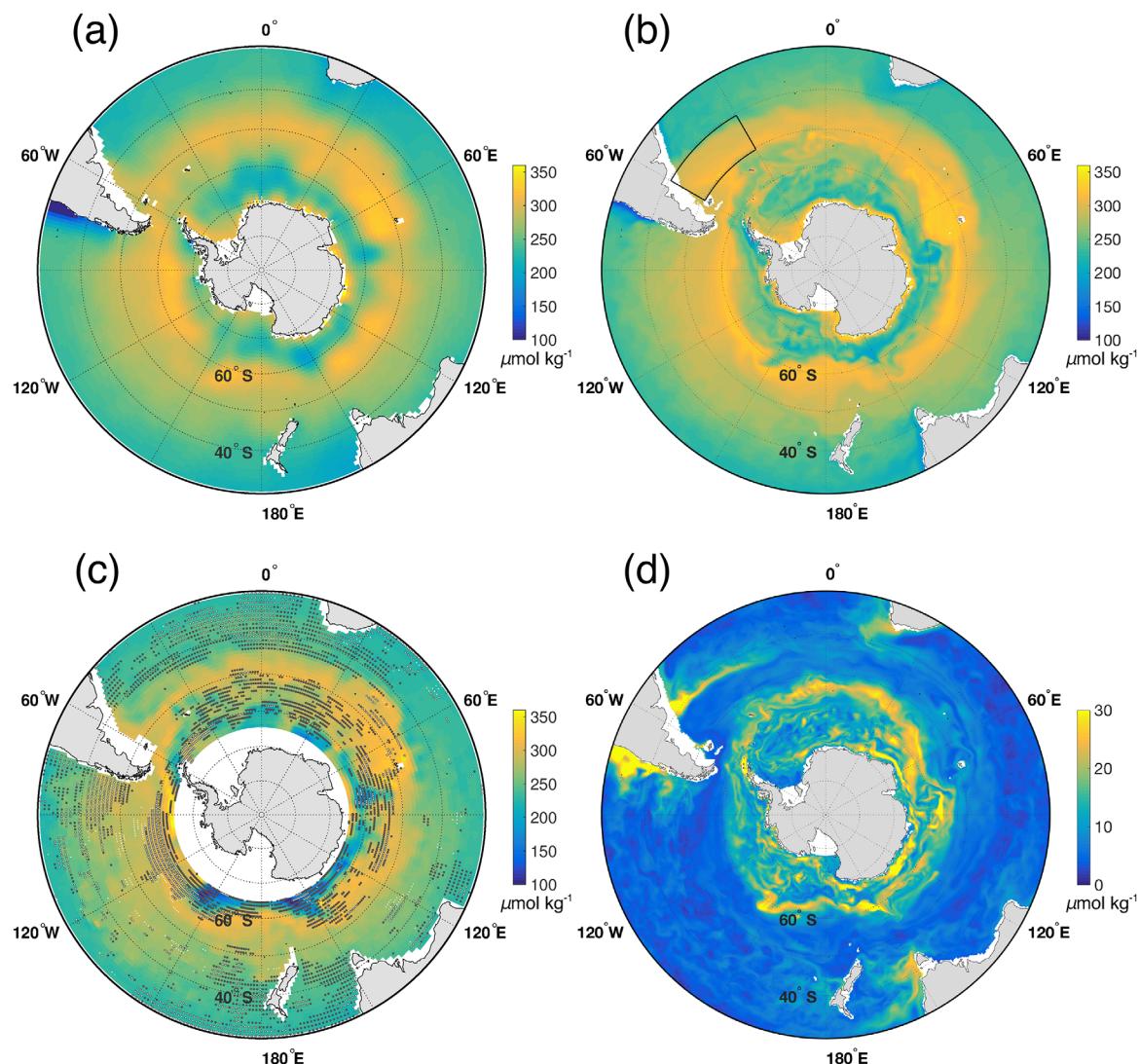
<sup>1</sup>Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA, <sup>2</sup>Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD, USA

**Abstract** An Argo-based estimate of oxygen ( $O_2$ ) at 150 m is presented for the Southern Ocean (SO) from temperature (T), salinity (S), and  $O_2$  Argo profiles collected during 2008–2012. The method is based on a supervised machine learning algorithm known as random forest (RF) regression and provides an estimate for  $O_2$  given T, S, location, and time information. The method is validated by attempting to reproduce the Southern Ocean State Estimate (SOSE)  $O_2$  field using synthetic data sampled from SOSE. The RF mapping shows skill in the majority of the domain but is problematic in some of the boundary regions. Maps of  $O_2$  at 150 m derived from observed profiles suggest that SOSE and the World Ocean Atlas 2013 climatology may overestimate annual mean  $O_2$  in the SO, both on a global and basin scale. A large regional bias is found east of Argentina, where high  $O_2$  values in the Argo-based estimate are confined closer to the coast compared to other products. SOSE may also underestimate the annual cycle of  $O_2$ . Evaluation of the RF-based method demonstrates its potential to improve understanding of  $O_2$  annual mean fields and variability from sparse  $O_2$  measurements. This implies the algorithm will also be effective for mapping other biogeochemical variables (e.g., nutrients and carbon). Furthermore, our RF evaluation results can be used to inform the design of future enhancements to the current array of  $O_2$  profiling floats.

## 1. Introduction

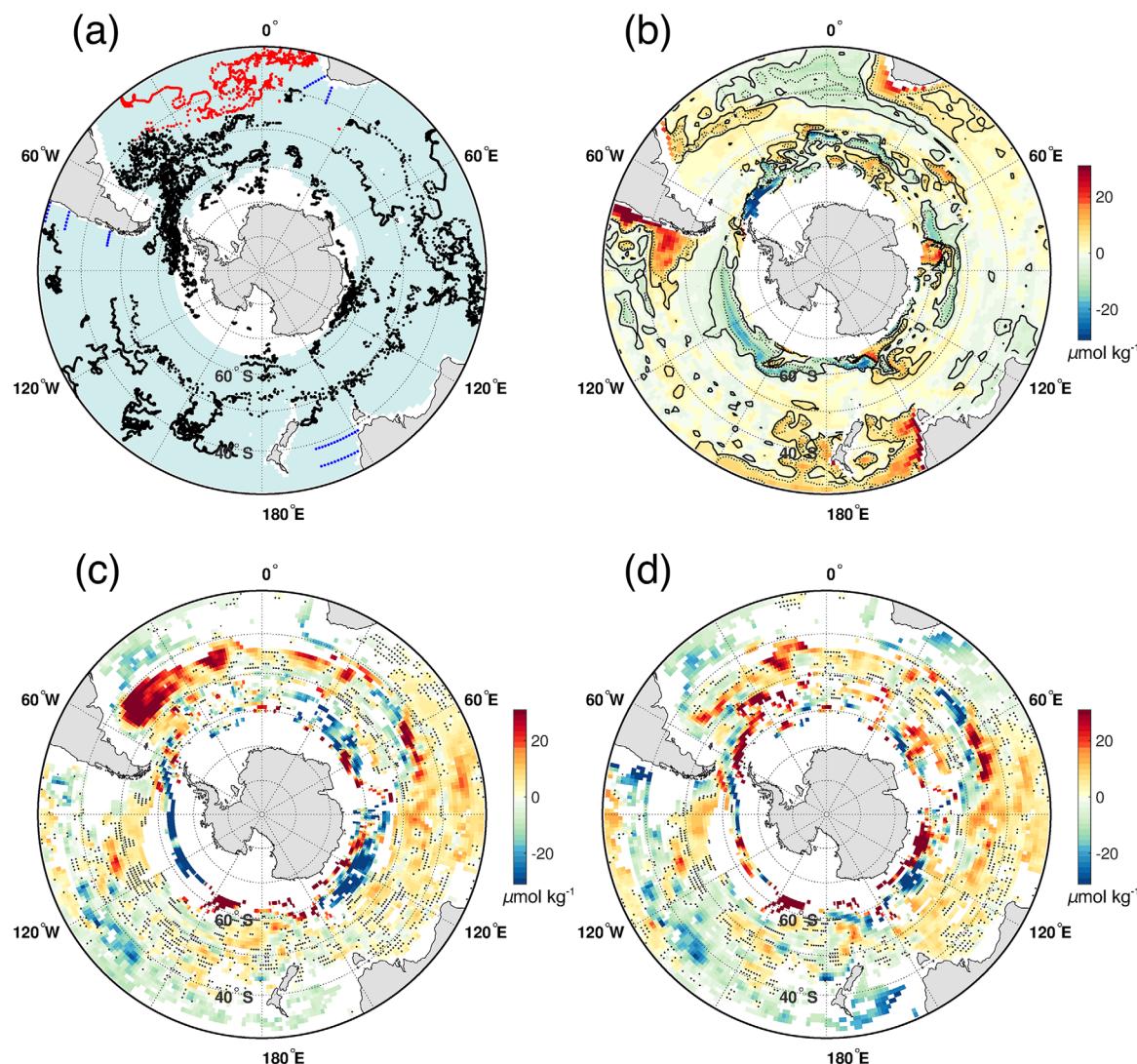
The oxygen ( $O_2$ ) distribution in the world oceans regulates survival of benthic organisms and the nitrogen cycle (Altabet, 2007; Diaz & Rosenberg, 2008) and informs on carbonate system parameters (Bushinsky & Emerson, 2015; Carter et al., 2016; Williams et al., 2016, 2017). The few (sparse) accurate long-term measurements of  $O_2$  that exist in the ocean interior indicate that  $O_2$  concentrations have declined over the past 50 years (Emerson & Bushinsky, 2014). Yet it remains a challenge to verify the inferred sensitivity of the carbon and oxygen cycle to changes in ecosystem dynamics, nutrient availability, and ocean circulation that are predicted by global climate models (Emerson & Bushinsky, 2014). An important step in this direction was mounting  $O_2$  sensors on Argo profiling floats (i.e.,  $\approx 200$  Argo floats with calibrated  $O_2$  data were deployed globally during 2002–2015; Drucker & Riser, 2016). The Argo array, currently including  $\approx 3,800$  floats, has been measuring ocean temperature and salinity (T and S) on a global scale for over a decade, with unprecedented resolution and no seasonal bias (supporting information Figures S1g–S1i), and provided a great opportunity to sample  $O_2$ . Former accuracy of  $\pm 3\%$  for  $O_2$  Argo observations (Takeshita et al., 2013) has been improved by Drucker and Riser (2016), who recalibrate the phase correction determined in the lab for each specific sensor (Optode) after a multi-Optode foil calibration, using the best available in situ reference data. The improved  $O_2$  data set by Drucker and Riser (2016) is still sparse. Yet  $O_2$  profiles have great potential to further our understanding of the  $O_2$  large-scale distribution, especially in the Southern Ocean (SO), where most of the observations previous to Argo (included in recent mapping efforts, e.g., Figure 1a) have been collected during summer (supporting information Figures S1b and S1c).

The concentration of  $O_2$  in the ocean is controlled by ocean-atmosphere interactions, biological activity, ocean circulation, and T/S (e.g., warmer water holds less  $O_2$ ). Modern machine learning algorithms with high computational speeds and less assumptions about the data (e.g., assumptions about the structure of spatiotemporal covariance and forms of relationships between variables) compared to traditional statistical techniques have a great potential for modeling highly nonlinear relationships between the variables and for efficient estimation of  $O_2$  from sparse observations. Nevertheless, there have been few studies implementing machine learning algorithms for such tasks, including neural networks (Chapman & Charantonis,



**Figure 1.** (a–c) Annual mean  $O_2$ ,  $\mu\text{mol kg}^{-1}$ . (a) WOA13 (i.e., data since the early 1900s), (b) SOSE (2008–2012), and (c) <sup>RF</sup>Argo<sub>TS</sub>  $O_2$  (2008–2012). The black box in Figure 1b indicates a region East of Argentina ( $60^\circ\text{W}$ – $30^\circ\text{W}$ ,  $50.5^\circ\text{S}$ – $40.5^\circ\text{S}$ ). Dark gray dots in Figure 1c indicate locations where  $\Delta_{RF_g} \geq 1.5\%$ , i.e., where the method bias is expected to be 1.5% the actual value or larger; white dots (at times overlaying dark gray dots) indicate where  $|\Delta_{RF_g}^*|$  is larger than  $O_2$  variability (i.e., than the  $O_2$  standard deviation in SOSE, in Figure 1d). (d)  $O_2$  standard deviation in SOSE,  $\mu\text{mol kg}^{-1}$ .

2017; Sauzède et al., 2015, 2016, 2017), and decision trees (Kamikawaji et al., 2016), and none of them target the estimation of  $O_2$ . Here we investigate the utility of a machine learning algorithm (random forest regression; Breiman, 2001) to map  $O_2$  using Argo observations of  $O_2$  (Figure 2a, supporting information Figures S1d–S1f) and T/S (supporting information Figures S1g–S1i). Random forest has a number of competitive advantages over other machine learning techniques, including high performance and low number of tuning parameters. Our proof of concept focuses on  $O_2$  (at one depth level, section 3) during the 5-year period between 2008 and 2012, when the Biogeochemical Southern Ocean State Estimate (B-SOSE, further referred to as SOSE) is available. SOSE (Verdy & Mazloff, 2017) provides a high-resolution dynamically consistent estimate of the 3-D ocean (including biogeochemistry) that is informed by (i.e., assimilates) observations and can be used to test our method. This analysis provides a framework to improve our understanding of the 3-D  $O_2$  distribution in the SO and of the regional 12-month anomaly from the annual mean, highlighting possible biases of other available products (e.g., based on mapping sparse  $O_2$  measurements or combining observations and models in ocean state estimates). The framework described here can also inform float deployment strategies to maximize the information we can learn about the large-scale  $O_2$  distribution from



**Figure 2.** (a)  $O_2$  data locations during 2008–2012 (black dots). Other dots are fictitious data locations in (blue) scenario 1 and (red and blue) scenario 2. Light blue shade indicates the area of the Southern Ocean considered in this analysis. (b–d) Maps of time mean  $O_2$  differences ( $\Delta^*$ ,  $\mu\text{mol kg}^{-1}$ ) between products, after removing the area-weighted spatial average from each. (b)  $^{RF_g}\text{SOSE}_{TS}$   $O_2$  minus  $SOSE O_2$  (*i.e.*,  $\Delta_{RF_g}^*$ ), (c)  $SOSE O_2$  minus  $^{RF}\text{Argo}_{TS}$   $O_2$ , and (d) WOA13  $O_2$  minus  $^{RF}\text{Argo}_{TS}$   $O_2$ . The time mean is over the period 2008–2012, except for WOA13, which includes observations since the early 1900s. Black contours in Figure 2b indicate  $\Delta_{\%RF_g}$  equal (solid line) 1.5% and (dots) 3%. In Figures 2c and 2d, regions where  $\Delta_{\%RF_g} \leq 1.5\%$  are not shown and black dots indicate where  $\Delta^* \leq \Delta_{RF_g}^*$ , as described in section 3.2.

future  $O_2$  profiles, e.g., from the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project floats (Johnson et al., 2017).

The rest of the paper is organized as follows. Sections 2.1, 2.2, and 3 describe observations, model, and method. Results are in section 4, with a summary of the method's performance and diagnostics in section 4.1, our estimate of the SO  $O_2$  at 150 m from Argo (5-year mean and 12-month anomaly) in section 4.2, and assessment of different sampling scenarios in section 4.3. Section 5 includes a summary and conclusions.

## 2. Data and Products

### 2.1. Observations

The Argo array has provided T/S observations in the upper 2,000 dbar of the global ocean for over a decade. Other sensors have also been tested on a number of floats for collecting information related to the biogeochemistry of the oceans (e.g.,  $O_2$ , nitrate, and chlorophyll  $a$ ) as part of the BGC-Argo program. Here we use

O<sub>2</sub> observations from Argo profiling floats for the period 2008–2012 (Figure 2a, black dots; supporting information Figure S1d), after bias correction described in Drucker and Riser (2016). We also use Argo T/S for the same period (including those floats that do not measure O<sub>2</sub>; supporting information Figure S1g).

We compare our Argo T/S-based estimate of O<sub>2</sub> in the SO with the World Ocean Atlas 2013 O<sub>2</sub> climatology (WOA13, Garcia et al., 2014). WOA13 is an objectively analyzed product at standard depth levels, which includes available O<sub>2</sub> observations (e.g., profiling floats and ship-based measurements) since the early 1900s.

## 2.2. The Biogeochemical Southern Ocean State Estimate

The Southern Ocean State Estimate is produced by performing a least squares fit optimization of a modern general circulation model (the MITgcm) to all available ocean observations (Mazloff et al., 2010). This is accomplished iteratively through the adjoint method and results in a high-resolution dynamically consistent estimate of the 3-D ocean T/S, pressure, and velocity fields. More recently, biogeochemistry has been included in SOSE (Verdy & Mazloff, 2017) and SOSE O<sub>2</sub> fields are used here (along with T/S) for comparison with our Argo T/S-based estimate of O<sub>2</sub> and to test our method. SOSE monthly and 3-day fields are considered, along with O<sub>2</sub>, T/S sampled (in SOSE) at the time and locations of Argo floats for the period 2008–2012 (SOSE profiles).

## 3. Method

All the results and discussion in this analysis are for O<sub>2</sub> at 150 m. The 150 m depth level is chosen since it is characterized by significant annual variability in O<sub>2</sub>, has reasonable sample size, and biological processes including phytoplankton blooms are less influential there (than closer to the surface), making T/S better predictors of O<sub>2</sub>.

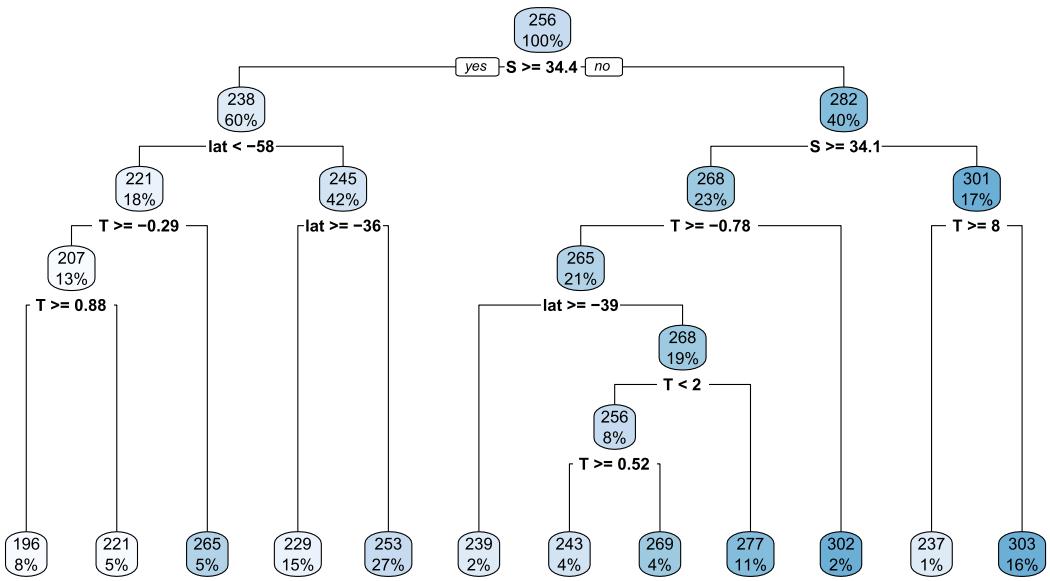
The method presented here is based on random forest (RF; Breiman, 2001) regression and is applied separately to two data sets: Argo observed profiles and SOSE estimated profiles (i.e., SOSE O<sub>2</sub> at the time and location of Argo O<sub>2</sub> observations considered in this analysis). Our Argo-based O<sub>2</sub> estimate is compared with WOA13 and SOSE products to inform biases they may have, while the analogous estimate from SOSE profiles can be compared to SOSE (actual) fields to verify how the method performs. Using the model fields, we also simulate scenarios where different numbers of O<sub>2</sub> profiles are available, and estimate how the RF-based results may improve if more floats are deployed.

Compared with other machine learning methods (such as gradient boosting, neural, and deep neural networks), RF is known for a great balance of predictive performance, number of tuning parameters, and computational speed. In particular, predictive accuracy of RF is often comparable to deep machine learning and better than other techniques; RF has fewer tuning parameters than many other methods, especially those for deep learning; RF's speed is often in between deep learning approaches (slower) and other machine learning methods (faster), such as stand-alone regression trees (e.g., see Ahmad et al., 2017; Futoma et al., 2015; Kutkina & Feuerriegel, 2016). R package *randomForest* (Liaw & Wiener, 2002) was used to apply the RF method in this study.

### 3.1. Random Forest Regression

The basic element of a random forest regression is a regression tree (Breiman et al., 1984) similar to the one in Figure 3. Following from top to bottom, the data set is split by the values of predictors (e.g., T, S, latitude, longitude, year, and month) that yield subsets most homogeneous based on the response variable (O<sub>2</sub>). The bottom (terminal) nodes, which are not getting split anymore, provide O<sub>2</sub> predictions for a given combination of predictors. Notice that some predictors are not represented in the tree (Figure 3) if they cannot split the data better than other predictors.

Let  $\Theta(\mathbf{Z})$  be a regression tree grown on training data  $\mathbf{Z}$ , where  $\mathbf{Z}$  can be represented as a  $N \times (p+1)$  matrix with  $N$  observations of  $p$  predictor variables and a response variable. Tree  $\Theta(\mathbf{Z})$  is constructed by a recursive search over all possible values of predictors to find best splits that yield the most homogeneous subsets of the response variable. Subsequent splits can be made using the same or a different predictor. Thus, nonlinear relationships between response and predictors and (if subsequent splits are based on different predictors) interaction effects between variables can be accommodated in a tree (Berk, 2016). Forecast or fitted



**Figure 3.** A regression tree obtained from the observed Argo data. Each node shows average  $O_2$  ( $\mu\text{mol kg}^{-1}$ ) and node size as percentage of the total sample size  $N = 8,704$ .

value for some combination of predictors  $x$ ,  $\Theta(x; \mathbf{Z})$ , is an average value of the response variable in the corresponding terminal node of the tree. As with any other regression technique, better results are achieved when no extrapolation is done, i.e., when  $x$  values are within the range of the predictors in  $\mathbf{Z}$ .

Bootstrap aggregation (bagging) works well to reduce variance of tree-based predictions. In bagging, predictions from a large number ( $B$ ) of trees are aggregated, where each tree is grown on a sample with replacement (bootstrap sample)  $\mathbf{Z}_b^*$  of size  $N$  from the original data  $\mathbf{Z}$ . The reduction of variance is higher the smaller is the correlation between the pairs of trees. The technique of random forest (RF, Breiman, 2001) aims to further reduce correlation between trees by using only  $m \leq p$  predictors randomly selected at each tree split. We use the recommended  $m$  of  $\lfloor p/3 \rfloor$ , and the minimum node size of five (Hastie et al., 2009). RF predictions can be written as

$$RF\hat{f}(x) = B^{-1} \sum_{b=1}^B \Theta_b(x, \mathbf{Z}_b^*), \quad (1)$$

where  $\Theta_b(x, \mathbf{Z}_b^*)$  are predictions from the  $b$ th random forest tree grown on the  $b$ th bootstrap sample  $\mathbf{Z}_b^*$ .

Since each  $\mathbf{Z}_b^*$  is a sample with replacement, about  $N/3$  of the original data are left out each time and not used to construct a tree (so-called out-of-bag, OOB, data). In the Breiman's RF approach, OOB error can be calculated by (i) using the OOB  $x_{OOB}$  to forecast the response variable using the tree, then (ii) comparing the obtained predictions with the actual response (let  $RFMSE$  be the obtained OOB mean squared error). At the next step,  $i$ th predictor's values in  $x_{OOB}$  are randomly permuted and error  $RFMSE(i)$  is estimated ( $i=1, \dots, p$ ). The more important the predictor, the bigger the difference between the two errors. The average difference for  $i$ th predictor over  $B$  trees is

$$I_i = B^{-1} \sum_{b=1}^B (RFMSE(i)_b - RFMSE_b), \quad (2)$$

where  $RFMSE(i)$  is OOB MSE with the values of  $i$ th predictor in OOB data permuted. In this permutation-based approach, the largest  $I_i$  correspond to the most important predictors.

With a few extra steps (see section 5.7 by Berk, 2016), we assess relationships between each predictor and response in RF. Suppose that  $i$ th predictor has  $m$  unique values  $u_{i,1}, \dots, u_{i,m}$  in  $\mathbf{Z}$ . The partial dependence for  $i$ th predictor is then represented by  $m$  average predictions from the original RF when  $x$  in (1) is replaced with a data set  $\mathbf{Z}$  modified in such way that all values of the  $i$ th predictor are equal to one of its  $m$  unique

values (the rest of the predictors remain unmodified). In other words, for each  $j=1, \dots, m$ , we (i) replace all (i.e.,  $N$ ) values of  $i$ th predictor in  $\mathbf{Z}$  with  $u_{ij}$ ; (ii) use the modified  $\mathbf{Z}$  to obtain  $N$  predictions from the RF, then (iii) calculate an average of the  $N$  predictions.

### 3.2. Applying RF to Profile Data and Evaluating Results

In this section, we describe how the RF method is applied to profile data to produce and evaluate  $O_2$  estimates presented in section 4.

The RFs are trained separately on Argo and SOSE profiles during 2008–2012 (i.e., when data both from Argo and SOSE are available) using  $O_2$  as the response variable and T/S, latitude, longitude, month, and year as predictors. An alternate version of the RFs is also considered which includes two additional predictors: surface chlorophyll (November–February climatology, when cloud coverage is less likely to cause gaps in the record) and time-varying sea surface height, both derived from satellite data. If not specified otherwise, RF estimates discussed in the manuscript are based on six predictors. Although biology is not explicitly accounted for as one of the predictors in the main set of RFs, differences in biological processes in time and space are partially accommodated using the spatiotemporal variables already included in the RF. As we run the RF, we assess whether the mean squared error stabilizes with the number of trees in RF increasing up to  $B$ . If not,  $B$  should be increased.

We then obtain RF predictions of  $O_2$  on Argo T/S profiles collected in 2008–2012 and grid these estimates onto a  $1^\circ \times 1^\circ$  (latitude/longitude) grid using a weighted least squares fit (LSF) method. Similarly to Roemmich and Gilson (2009), the LSF algorithm fits to the  $K$  data points nearest to a grid point a three-dimensional surface defined by linear and quadratic terms in latitude and longitude, along with a time mean, the annual cycle and its 2, 3, 4, and 6 months harmonics. Different from Roemmich and Gilson (2009), weights here are based on the actual distance between the data and the grid point; also, only data within an ellipse centered on the grid point are considered (with axes of the ellipse measuring 4,000 and 2,000 km in the zonal and meridional direction, respectively); finally, data are excluded if land is present between the data and the grid point. If available within the ellipse, the 100 data points (nearest to the grid point) per each of the 12 months are included in the fit, i.e.,  $K_{\max}=100 \times 12 = 1,200$ .

We refer to the resulting Argo-based gridded  $O_2$  fields as  ${}^{\text{RF}}\text{Argo}_{\text{TS}} O_2$ , while  ${}^{\text{RF}}\text{SOSE}_{\text{TS}} O_2$  is an analogous product but using SOSE  $O_2$ , T/S profiles.  ${}^{\text{RF}}\text{Argo}_{\text{TS}} O_2$  has the potential to inform biases in other available products (e.g., WOA13 and SOSE) in regions where the RF method is determined to perform well.

To determine these skillful regions, we apply the RF approach to SOSE profiles and estimate  $O_2$  directly on the SOSE T/S grid. We refer to this estimate as  ${}^{\text{RF}_g}\text{SOSE}_{\text{TS}} O_2$ , with  $g$  indicating the  $O_2$  prediction is based on the T/S SOSE (output) grid.  ${}^{\text{RF}_g}\text{SOSE}_{\text{TS}} O_2$  is used to assess the RF method performance rather than  ${}^{\text{RF}}\text{SOSE}_{\text{TS}} O_2$  since we are first interested in how the RF method performs, not the gridding algorithm. For the 5-year mean fields, the difference  $\Delta_{\text{RF}_g}^*$  between  ${}^{\text{RF}_g}\text{SOSE}_{\text{TS}} O_2$  and SOSE (after removing the area-weighted spatial average from each product, Figure 5c) is taken as measure of local RF performance. We trust the RF approach in regions where  $\Delta_{\text{RF}_g}^* < \gamma_{1.5}$ , with  $\Delta_{\text{RF}_g}^* = \frac{|\Delta_{\text{RF}_g}|}{O_2^{\text{SOSE}}} \times 100$ ,  $O_2^{\text{SOSE}}$  the actual SOSE field, and  $\gamma_{1.5}=1.5\%$ . In the trusted regions, differences  $|\Delta^*| > |\Delta_{\text{RF}_g}|$  between  ${}^{\text{RF}}\text{Argo}_{\text{TS}} O_2$  and other products (i.e., SOSE or WOA13; removing from each the area-weighted spatial average) inform biases in the product of interest. Percentages of the domain area (shaded in Figure 2a) where described criteria apply for different values of  $\gamma$  are summarized in Table 2.

The 12-month anomaly (from the annual mean) is compared across products for regional averages, with differences between  ${}^{\text{RF}}\text{SOSE}_{\text{TS}} O_2$  and SOSE  $O_2$  informing on the bias of the RF-based estimate of the  $O_2$  annual cycle.

Finally, as an indication of how the LSF gridding method performs,  $\text{SOSE}_{\text{LSF}} O_2$  is also shown.  $\text{SOSE}_{\text{LSF}} O_2$  results from applying the LSF gridding to SOSE  $O_2$  at the Argo T/S locations. A summary of the different  $O_2$  estimates discussed in this analysis can be found in Table 1. The domain of interest for the comparison across products is shaded in Figure 2a and includes regions where Argo T/S data are available (supporting information Figure S1g).

**Table 1***Summary Table for Different  $O_2$  Estimates in This Analysis*

$O_2$ RF predictions on SOSE T/S output grid, after training RF with $mp_i$ model profiles	$O_2$ at T/S profile data locations are gridded using the LSF method		
$mp_i$ at obs locations	$mp_i$ at test locations*	$O_2$ from RF predictions	SOSE $O_2$
$RF_g$ SOSE <sub>TS</sub> $O_2$ R1 in Table 3	$RF_{g1}$ SOSE <sub>TS</sub> $O_2$ *Figure 2a, black/blue dots	Argo profiles	Model profiles
	$RF_{g2}$ SOSE <sub>TS</sub> $O_2$ *Figure 2a, black/blue/red	$RF_{Argo_{TS}}$ $O_2$	$RF_{SOSE_{TS}}$ $O_2$
	$RF_{gTS}$ SOSE <sub>TS</sub> $O_2$ *Argo T/S locations		SOSE <sub>LSF</sub> $O_2$
	A1, A2 in Table 3		

### 3.3. Simulating Different Sampling Scenarios

As described in section 3.2, we assess the method's performance based on how the  $RF_g$  SOSE<sub>TS</sub>  $O_2$  estimate compares to the actual SOSE  $O_2$  fields. We also investigate (in SOSE) the sensitivity of the RF regression to using different numbers of  $O_2$  profiles. Scenarios are tested using profiles at available locations plus additional data, which result in these hypothetical  $O_2$  estimates:  $RF_{g1}$  SOSE<sub>TS</sub>  $O_2$ , from additional  $O_2$  data near boundary regions (black and blue dots in Figure 2a);  $RF_{g2}$  SOSE<sub>TS</sub>  $O_2$ , as  $RF_{g1}$  SOSE<sub>TS</sub>  $O_2$  but with additional  $O_2$  data also in the South Atlantic (black, blue, and red dots in Figure 2a);  $RF_{gTS}$  SOSE<sub>TS</sub>  $O_2$ , from  $O_2$  data at Argo T/S locations during 2008–2012 (see supporting information Figure S1g for the number of profiles in  $1^\circ \times 1^\circ$  latitude/longitude bins).

Additional experiments are performed for the month of October during 2008–2012, training the RF with only October data. In this case, scenarios with 2 and 4 times the number ( $n = 545$ ) of actual Argo  $O_2$  October profiles are evaluated, along with two different distributions:

1. The number of data is doubled (quadrupled) randomly within the domain (with additional profiles distributed at least 300 km apart from each other and from real profiles).
2. The number of data is doubled (quadrupled) within each zonal sector, with additional profiles  $\approx 200$  km or more apart.

For each scenario, an ensemble of 10 members is considered, and sample distributions are shown in supporting information Figure S7.

As a measure of bias for the climatological October  $O_2$  distribution resulting from each scenario ( $i$ ), we consider two contributions to the difference  $\Delta_i$  between RF estimates  $RF_{O_{2,i}}$  and SOSE  $O_2$ , i.e.,  $\Delta_{i,1}$  and  $\Delta_{i,2}$ . Specifically, the difference between RF estimates and actual SOSE values:

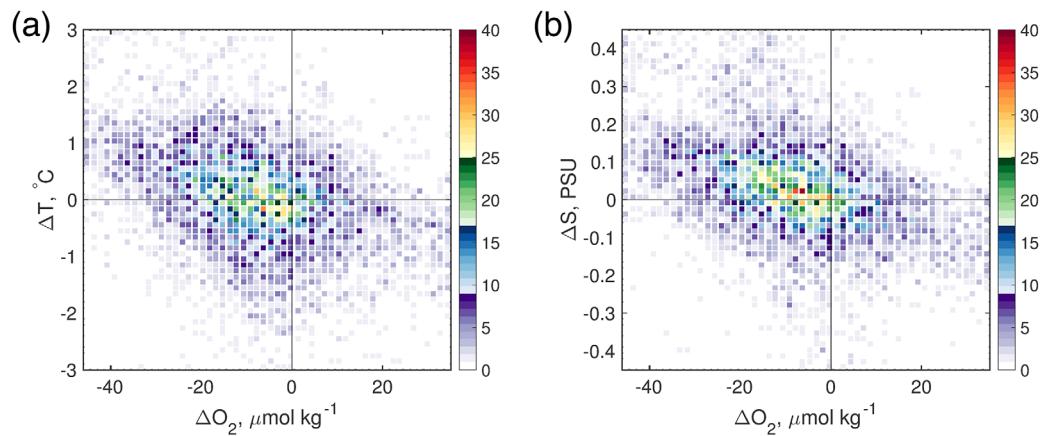
$$\Delta_{i,1} = \langle RF_{O_{2,i}} \rangle - \langle O_2 \rangle,$$

where  $\langle \cdot \rangle$  indicates an area-weighted average. In turn,  $\Delta_{i,2} = \sqrt{\langle SE \rangle}$ , where

$$SE = [ (RF_{O_{2,i}} - \langle RF_{O_{2,i}} \rangle) - (O_2 - \langle O_2 \rangle) ]^2.$$

We also consider the ratio of  $\Delta_{i,2}$  and the square root of the area-weighted variance of SOSE 12-month climatology, i.e.,  $\sigma_{SOSE} = \sqrt{\langle \sigma^2 \rangle} = 5.52 \mu\text{mol kg}^{-1}$ . The average and range of  $\Delta_{i,1}$ ,  $\Delta_{i,2}$ , and  $\Delta_{i,2}/\sigma_{SOSE}$  across 10 Monte Carlo simulations are compared with a similar estimate obtained with RF using the real distribution of October profiles (scenario R1).

We note that  $\sqrt{SE}$  is the same as  $\Delta_{RF_g}^*$  described in section 3.2, except now it is for the October mean over 2008–2012 rather than the 5-year mean.



**Figure 4.** Histograms of differences in  $O_2$  ( $\Delta O_2$ ,  $\mu\text{mol kg}^{-1}$ ), potential temperature ( $\Delta T$ ,  $^{\circ}\text{C}$ ), and salinity ( $\Delta S$ , PSU) between Argo profiles and SOSE. Color indicates the number of points in (a)  $\Delta O_2$ ,  $\Delta T$  and (b)  $\Delta O_2$ ,  $\Delta S$  bins.

## 4. Results

We use RF regression separately on colocated Argo and SOSE profiles and estimate  ${}^{RF}\text{Argo}_{TS} O_2$  and  ${}^{RF}\text{SOSE}_{TS} O_2$ , respectively. These estimates are expected to differ, since the observed profiles are saltier, and characterized by lower  $O_2$  than SOSE (Figure 4). In the following, section 4.1 presents a summary of the method performance from model profiles (section 4.1.1) and diagnostics from RF regression (section 4.1.2), section 4.2 describes the Argo-based  $O_2$  estimate  ${}^{RF}\text{Argo}_{TS} O_2$ , and section 4.3 shows experiments training RF with data sets of different sizes and distributions (based on SOSE fields).

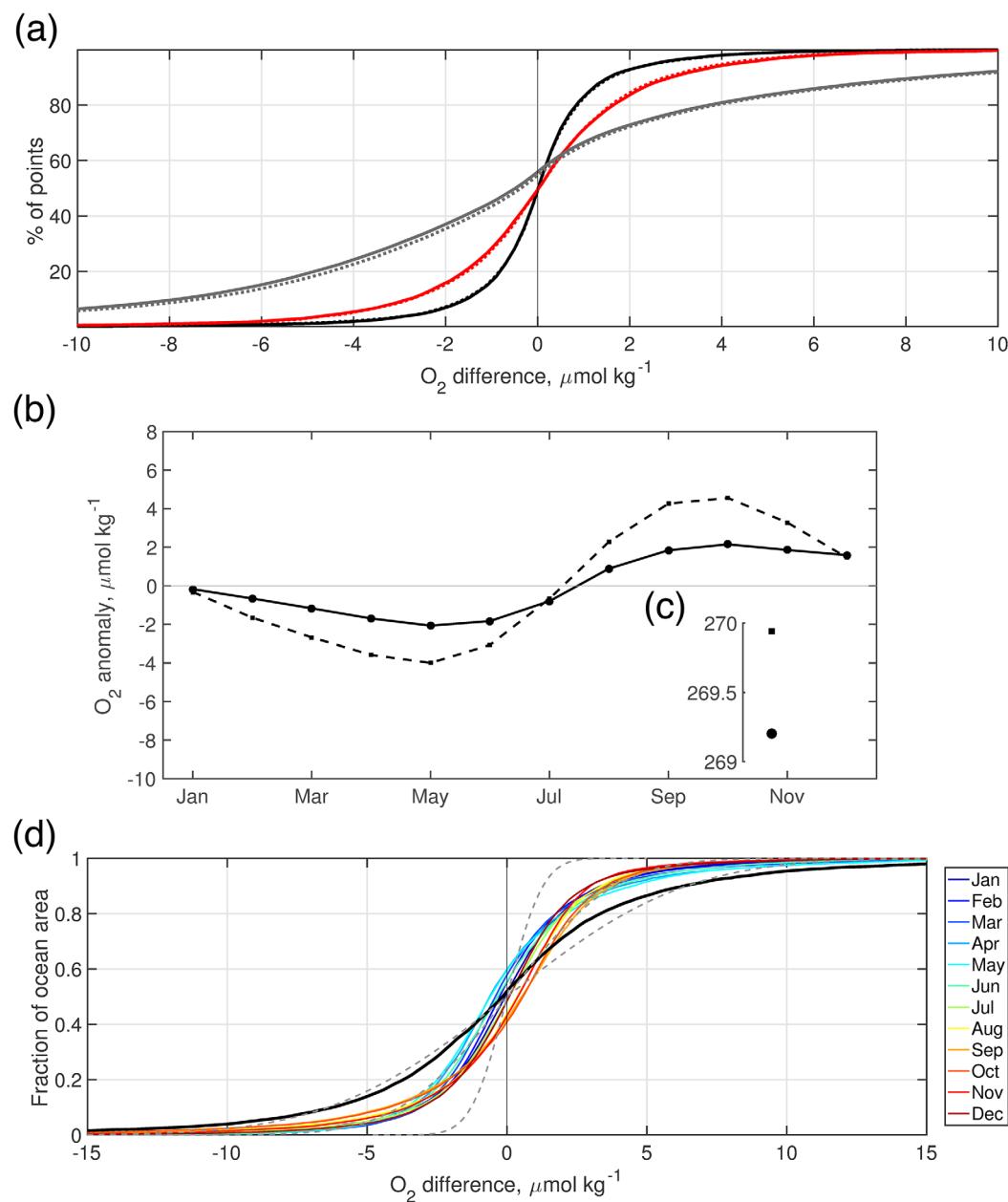
### 4.1. Random Forest Regression

#### 4.1.1. RF Performance

Figure 5 shows a summary of how the method performs for model profiles, comparing both  $O_2$  predictions at Argo floats locations (Figure 5a) and  ${}^{RF}\text{SOSE}_{TS} O_2$  (Figures 5b–5d) to actual SOSE fields. Predictions at locations where  $O_2$  is available (i.e., locations of data used to train the RF) are close to the actual  $O_2$  values, with a better performance for model profiles compared to (noisier) observations (black versus red line in Figure 5a). Predictions at the locations of Argo T/S profiles show differences smaller than  $4 \mu\text{mol kg}^{-1}$  around 60% of the times. These results are the same for RF with six or eight predictors (dotted lines in Figure 5a). Yet percentages in Figure 5a may not be very informative, since the distribution of T/S locations is not uniform in space (supporting information Figures S1g–S1i) and regions with larger/smaller estimate bias may be accounted for differently.

Differences between  ${}^{RF}\text{SOSE}_{TS} O_2$  and SOSE  $O_2$  are, instead, helpful to assess the RF method performance.  $\Delta_1$  (as defined in section 3.3 but now for  $O_2$  5-year mean over all the months) is small (Figure 5c) and indicates the estimated field is slightly biased low. The area-weighted cumulative distribution function for differences between 5-year means of  ${}^{RF}\text{SOSE}_{TS} O_2$  and SOSE (i.e.,  $\sqrt{SE}$  as defined in section 3.3 but now for  $O_2$  5-year mean over all the months) has heavier tails than a Gaussian distribution with standard deviation equal to  $5 \mu\text{mol kg}^{-1}$  (Figure 5d). Yet in  $\approx 65\%$  of the domain, differences between the two products are less than 1.5% of the actual value (Table 2), hence we trust the RF method here.  $\Delta_2$  (as defined in section 3.3 but now for  $O_2$  5-year mean over all the months) is  $6.79 \mu\text{mol kg}^{-1}$  ( $\Delta_2/\langle \sigma \rangle = 1.23$ ) and may be dominated by differences e.g., at eastern boundaries (i.e., small spatial scales associated with upwelling) and in the southern part of the basin (Figure 2b). These large differences likely arise from lack of data (e.g., no  $O_2$  Argo observations east of Australia, and west of Africa, and South America), and sparse data south of  $\approx 58^{\circ}\text{S}$  for both  $O_2$  and T/S (Figure 2a, supporting information Figures S1d–S1i). Moreover, these are also regions characterized by high variability (Figure 1d). Nevertheless, the bias is less than 3% of the actual SOSE field in 86% of the domain of interest (shaded in Figure 2a), Table 2.

Twelve-month anomalies (from the 5-year mean) show a larger  $\Delta_1$  compared to the 5-year mean (Figure 5b). Yet area-weighted cumulative distribution functions for  $\sqrt{SE}$  (of 12-month anomaly fields) are closer to



**Figure 5.** Summary of the comparison between RF  $O_2$  estimates and actual  $O_2$  fields. (a) Percentage of profiles included, (sequentially) accounting for differences, from negative to positive, between: (red) RF  $O_2$  estimates based on Argo T/S and actual Argo  $O_2$ ; (black) RF  $O_2$  estimates based on SOSE T/S (at locations where both Argo T/S and  $O_2$  are available) and actual SOSE  $O_2$ ; (gray) RF  $O_2$  estimates based on SOSE T/S (at locations where Argo T/S are available) and actual SOSE  $O_2$ . Dotted lines are for RF  $O_2$  estimates based on 8 predictors. (b, c) Area-weighted spatial average  $O_2$  between 30°S and 60°S for (b) the 12-month anomaly (from the annual mean) and for (c) the annual mean,  $\mu\text{mol kg}^{-1}$ . Solid line and dots for  $^{RF_g}SOSE_{TS}$   $O_2$ , dashed line and squares for SOSE. (d) Fraction of ocean area covered, (sequentially) accounting for differences, from negative to positive, between  $^{RF_g}SOSE_{TS}$   $O_2$  and SOSE, once the area-weighted average has been removed from each product (i.e.,  $\sqrt{SE}$ ). Thin solid lines are for 12-month anomalies (month in legend), the black thick line is for the 5-year mean. Gray dashed lines are for Gaussian distributions with standard deviations 1, 3, and 5  $\mu\text{mol kg}^{-1}$ .

a Gaussian distribution with standard deviation of 3 (rather than 5)  $\mu\text{mol kg}^{-1}$ , with most of the domain characterized by differences between  $\pm 2.5 \mu\text{mol kg}^{-1}$  (Figure 5d).

#### 4.1.2. RF Diagnostics

Mean squared errors for the ensemble of random forest trees decline and stabilize at about 100 trees for both RFs on Argo and SOSE profiles (Figure 6). Thus,  $B = 500$  used in this research is sufficient. RF explained

**Table 2**

Percentage of the Total Ocean Area of Interest (i.e., Shaded Area in Figure 2a) Where (Top)  $\Delta\% < \gamma$  and (Bottom)  $\Delta\%_{RF_g} < \gamma$  and  $|\Delta^*| > |\Delta_{RF_g}^*|$ , i.e., Where Differences  $\Delta^*$  Between  $d'$  and  $d'_c$  Are Significant

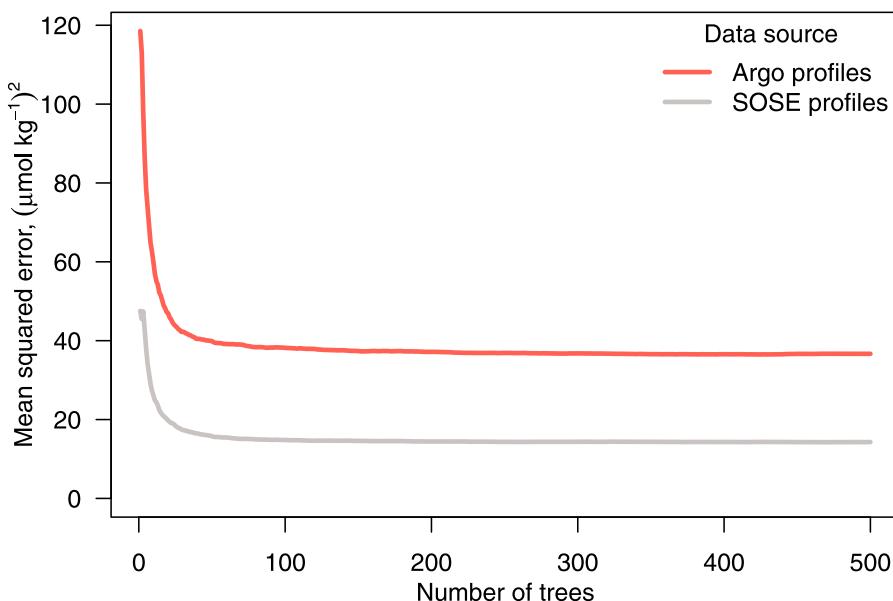
$d$	$d_c$	Percentage of area where $\Delta\% < \gamma$			
		$\gamma=1\%$	$\gamma=1.5\%$	$\gamma=2\%$	$\gamma=3\%$
$RF_g$ SOSE <sub>TS</sub>	SOSE	50	65	74	86
$RF_{g1}$ SOSE <sub>TS</sub>	SOSE	46	62	72	83
$RF_{g2}$ SOSE <sub>TS</sub>	SOSE	52	68	77	88
$RF_{gTS}$ SOSE <sub>TS</sub>	SOSE	80	89	94	98
Percentage of area where $\Delta\%_{RF_g} < \gamma$ and $\Delta^* > \Delta_{RF_g}^*$					
SOSE	$RF_{ArgoTS}$	43	54	60	67
WOA13	$RF_{ArgoTS}$	44	55	62	69

Note.  $\Delta^* = d' - d'_c$  and  $\Delta\% = \frac{|\Delta^*|}{d'}$ , with  $d$ ,  $d_c$ , O<sub>2</sub> estimates on a grid, and the prime symbol indicating the (area-weighted) spatial average has been removed from the product.  $\Delta\%_{RF_g}$  is  $\Delta\%$  when  $d$  is  $RF_g$  SOSE<sub>TS</sub> and  $d_c$  is SOSE, and similarly for  $\Delta_{RF_g}^*$ . Results for RFs with eight predictors are comparable and not shown.

96.9% of O<sub>2</sub> variance in Argo profiles, and 98.5% in SOSE profiles. Figure 7 informs on the relative importance of the predictors used to estimate O<sub>2</sub> in the RFs, whereas Figure 8 contains partial dependence plots when values of one predictor change while other predictors remain fixed.

T/S emerge as the two most important predictors, both when starting from Argo observations and model profiles (Figure 7), consistent with water mass dynamics in the SO (i.e., ventilation) and with T/S regulating solubility of dissolved O<sub>2</sub> in water (i.e., cold water can hold more O<sub>2</sub>). Antarctic Intermediate Water (AAIW) is characterized by high-oxygen and low-salinity and begins to subduct north of the Antarctic Circumpolar Current (ACC) Polar front (Giglio & Johnson, 2016, 2017; Orsi et al., 1995), resulting in a (circumpolar) low-salinity region that aligns with peak O<sub>2</sub> values at 150 m (Figures 1a–1c). O<sub>2</sub> decreases as salinity increases (Figure 8a). Despite O<sub>2</sub> solubility decreases, O<sub>2</sub> increases noticeably as temperature increases in the range  $\approx 2.5\text{--}5.1^\circ\text{C}$  (Figure 8b). In this temperature interval, less profiles are available (compared to other ranges; see bars on the x axis of Figure 8b), yet  $\approx 2.5\text{--}5.1^\circ\text{C}$  waters are in the low-salinity region (not shown) and the observed behavior adequately reflects the AAIW ventilation.

Consistent with the described water mass properties, latitude is also a key predictor for O<sub>2</sub> (Figure 7), with the O<sub>2</sub> meridional peak showing prominently in the partial dependence diagnostics (Figure 8c). The overall importance of longitude is much lower (Figure 7), yet O<sub>2</sub> zonal gradients are present across bathymetric



**Figure 6.** Mean squared error for different number of trees in the random forests.

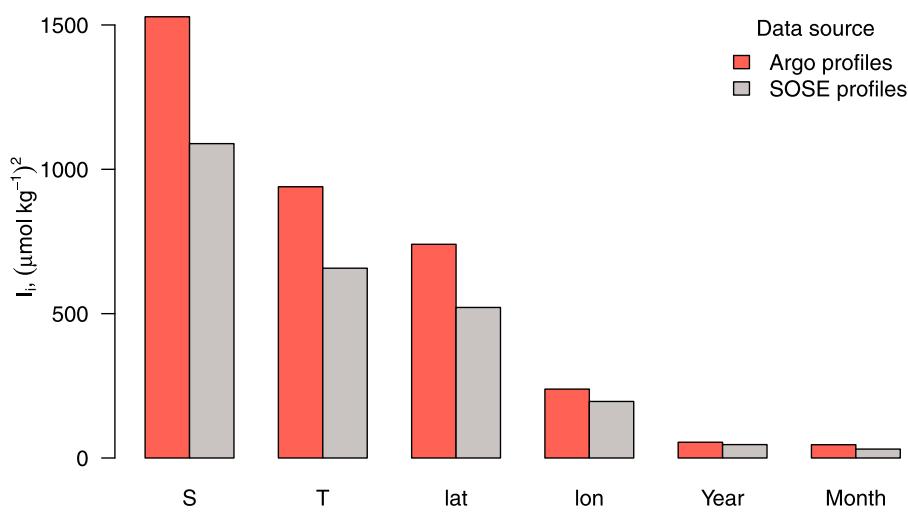


Figure 7. Permutation-based variable importance.

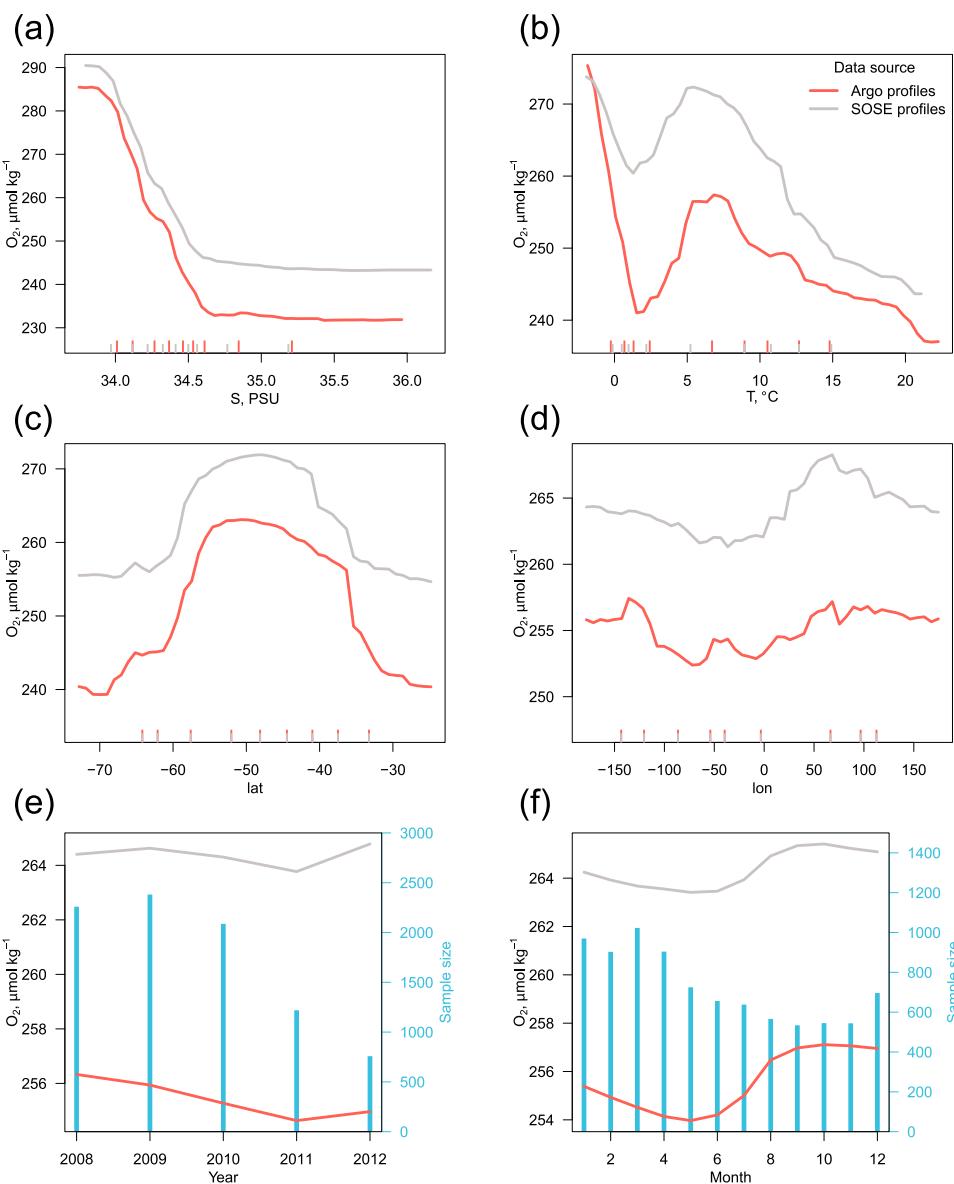
features that constrain ocean circulation. As an example, Figure 8d shows an increase of  $O_2$  in longitude in correspondence of the Kerguelen Plateau ( $\approx 70^\circ E$ ). Sharp  $O_2$  changes are also observed in the  $140^\circ W$ – $105^\circ W$  sector (Figure 8d), corresponding to the Eltanin Fracture Zone, Menard Fracture Zone, and East Pacific Rise. Gradients are generally stronger for the observations, as model fields are smoother by construction (not only in longitude). Year and month are similarly less important compared to other predictors (Figure 7), with the partial dependence on month showing the annual cycle of  $O_2$  (Figure 8f), larger for Argo than SOSE profiles. The partial dependence on year shows  $O_2$  decreasing from 2008 to 2012 (Figure 8e). Yet the sample size varies largely across years (blue bars in Figure 8e) and observed interannual variability may be not significant.

Overall, four predictors (latitude, longitude, month, and year) had the same values across the Argo and SOSE profiles, whereas the response variable ( $O_2$ ) and two most important predictors (T/S) had different values (Figure 4). RF estimates of  $O_2$  based on SOSE profiles are generally higher than those based on Argo data (Figure 8). Differences in individual patterns observed in the partial dependence plots (Figure 8) may imply differences in relationships (between  $O_2$  and other variables) existing in the Argo data and represented in SOSE. For example, such differences are observed for T/S relationships with  $O_2$  (Figures 8a and 8b). However, one needs to keep in mind that partial dependence of a variable is assessed while keeping other variables fixed, which means that T (S) would be fixed at different levels for Argo and SOSE when checking the partial dependence on S (T), consistent with differences in T/S diagrams between Argo and SOSE (not shown). Thus, the comparison of patterns is not straightforward in this case.

Results for the RF with eight predictors (supporting information Figures S2–S3) are consistent with those (above) for six predictors. Sea surface height aligns with circulation pathways and is a more important predictor than latitude when using Argo profiles, and as important for model profiles (supporting information Figure S2, bottom). We find that chlorophyll is not a very important predictor.

#### 4.2. Argo T/S-Based $O_2$ Estimate from RF Regression

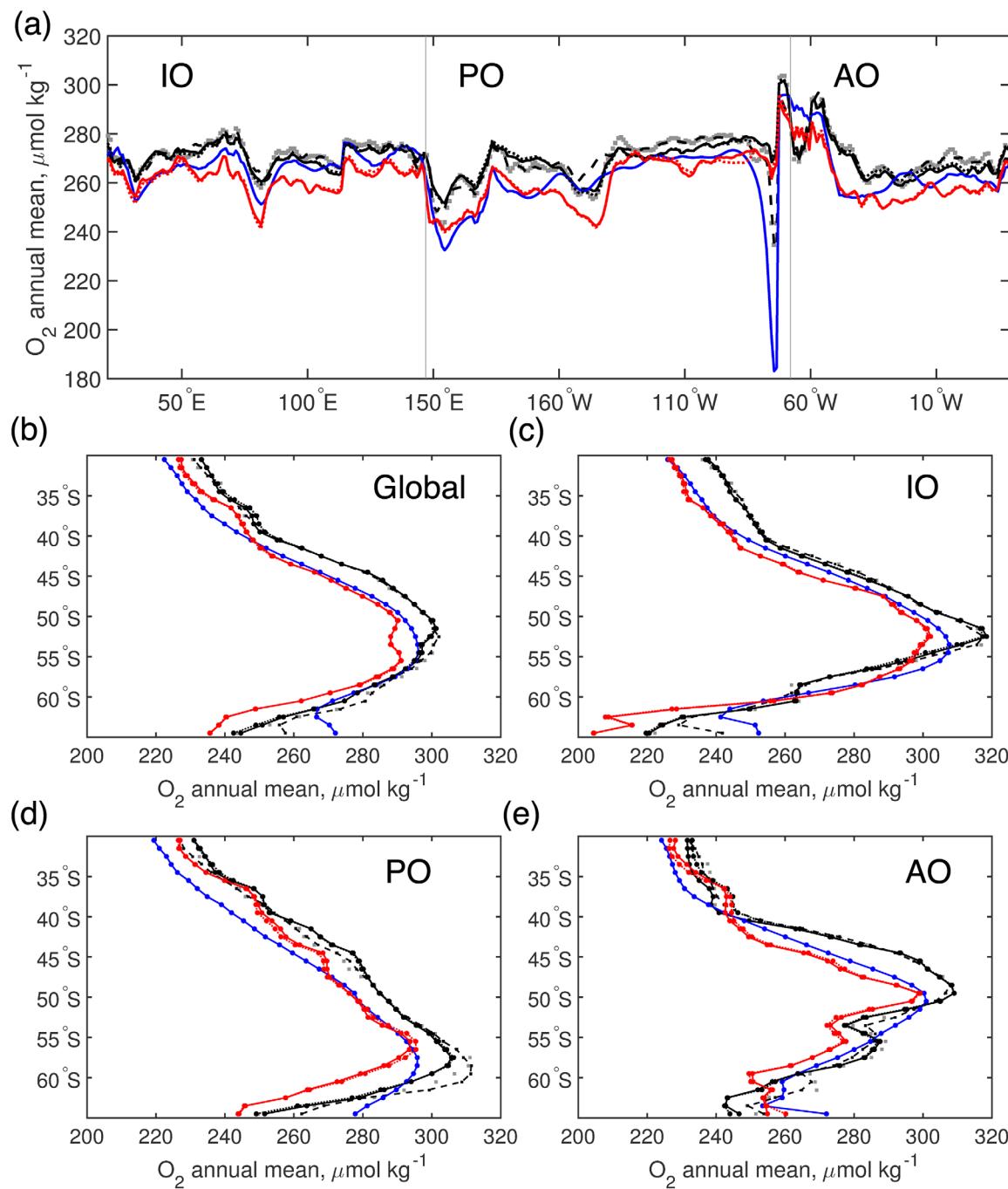
The general structure of 5-year mean  ${}^{RF} \text{Argo}_{TS} O_2$  (Figure 1c) is consistent with other products (Figures 1a, 1b, and 9), with a (circumpolar) meridional peak in  $O_2$  along the ACC. As described in section 4.1, the southern boundary of the peak  $O_2$  region aligns with the ACC Polar Front (see the location of S-PF in Chapman (2017), Figure 10), where AAIW starts subducting.  ${}^{RF} \text{Argo}_{TS} O_2$  shows overall lower values than WOA13 and SOSE (unmasked regions with red shades in Figures 2c, 2d, 9, and 10b), and is closer to WOA13 in the region east of Argentina. Figures 9 and 10b show how  ${}^{RF} \text{SOSE}_{TS} O_2$  is slightly biased, maybe due to limited sampling by available profiles (black dots in Figure 2a). Yet this slight RF method bias does not account for the difference between  ${}^{RF} \text{Argo}_{TS} O_2$  and SOSE in Figures 2c, 9, and 10b, and our results suggest SOSE  $O_2$  is biased high compared to observations. Also, gray squares in Figures 9 and 10 (i.e.,  $\text{SOSE}_{LSF} O_2$ ) align well with SOSE, indicating that the LSF gridding is not causing a bias. Exceptions to 5-year mean  ${}^{RF} \text{Argo}_{TS} O_2$  comparing more closely to WOA13 than SOSE in the SO on a basin scale (Figure 9), are latitudes north of  $\approx 40^\circ S$  in the Pacific and Atlantic



**Figure 8.** Partial dependence of  $O_2$  from (red) observed and (gray) model profiles. For continuous variables, inner ticks on the x axes denote deciles of the observed and model data, respectively. For discrete variables (year and month), blue bars report sample size on the right y axes.

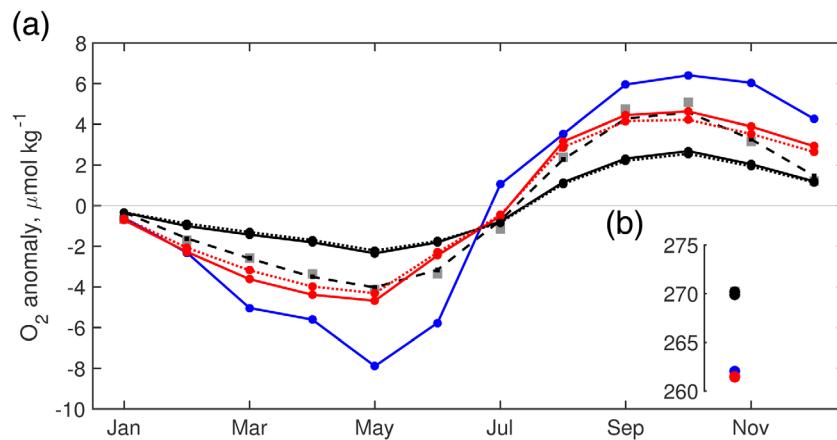
sectors of the SO (maybe due to  ${}^{RF}\text{Argo}_{TS}$  positive bias near continental boundaries), and latitudes south of  $\approx 62^{\circ}\text{S}$  globally, and in the Indian and Pacific sectors (Figures 9b–9d). Figure 9b is consistent with a comparison of SOSE and WOA13 with Argo profiles in Verdy and Mazloff (2017, their Figure 8a, for  $O_2$  at 200 m). A large regional difference between Argo-based  $O_2$  and other products is reported east of Argentina, as  ${}^{RF}\text{Argo}_{TS}$  high  $O_2$  is confined to the coast (Figures 1c, 2c, and 2d). Sharp differences also characterize (upwelling) regions close to eastern boundaries, where the RF regression method (from model profiles) fails to estimate the “true” field (i.e., SOSE; as described in section 4.1.1), hence  ${}^{RF}\text{Argo}_{TS}$   $O_2$  may too. A 5-year mean  $O_2$  reconstruction from binning RF  $O_2$  estimated (from observations) at the locations of Argo T/S profiles is not as smooth as  ${}^{RF}\text{Argo}_{TS}$   $O_2$  in Figure 1c and is sparser, due to bins with no floats especially in the southern part of the basin. Yet it is consistent with  ${}^{RF}\text{Argo}_{TS}$   $O_2$  and is not showed here.

The area-weighted average  ${}^{RF}\text{Argo}_{TS}$   $O_2$  anomaly (from the annual mean) shows a phasing of the annual cycle that is consistent with other products, with largest  $O_2$  in September–October (Figure 10), when the mixed layer depth is at 150 m or deeper along the ACC (Holte et al., 2017).



**Figure 9.** Annual mean  $\text{O}_2$  (as in Figures 1a–1c),  $\mu\text{mol kg}^{-1}$ . (a) Meridional average between  $64.5^{\circ}\text{S}$  and  $30.5^{\circ}\text{S}$ , with vertical gray lines indicating different basins. (b–e) Zonal average (b) globally, (c) in the Indian Ocean (IO), (d) in the Pacific Ocean (PO), and (e) in the Atlantic Ocean (AO). Different lines for (Red)  ${}^{\text{RF}}\text{Argo}_{\text{TS}}$   $\text{O}_2$ , (blue) WOA13, (solid black)  ${}^{\text{RF}}\text{SOSE}_{\text{TS}}$   $\text{O}_2$ , (dashed black) SOSE, (gray squares)  ${}^{\text{RF}}\text{SOSE}_{\text{LSF}}$   $\text{O}_2$ , i.e., applying the LSF mapping to actual SOSE  $\text{O}_2$  at the locations of Argo T/S profiles. Red and black dotted lines are for (red)  ${}^{\text{RF}}\text{Argo}_{\text{TS}}$   $\text{O}_2$  and (black)  ${}^{\text{RF}}\text{SOSE}_{\text{TS}}$   $\text{O}_2$  based on eight predictors.

${}^{\text{RF}}\text{SOSE}_{\text{TS}}$   $\text{O}_2$  shows an amplitude that is about half that of SOSE (Figure 10a). If this difference is an indication of the method bias (when applied to the available profiles), the actual amplitude of the annual cycle in the region may be larger than what  ${}^{\text{RF}}\text{Argo}_{\text{TS}}$   $\text{O}_2$  suggests (hence larger than SOSE), and closer to WOA13. Yet the spatial average in WOA13 may also be affected by poor spatial coverage and seasonal bias of the database used in the mapping (supporting information Figures S1a–S1c). Finally, gray squares in Figure 10a



**Figure 10.** Comparison across products for  $O_2$  large-scale averages in the SO ( $60^{\circ}\text{S}$ – $30^{\circ}\text{S}$ , over the domain shaded in Figure 2a). Area-weighted average  $O_2$  (a) 12-month anomaly and (b) annual mean,  $\mu\text{mol kg}^{-1}$ . Different lines/markers for (red, dots)  ${}^{\text{RF}}\text{Argo}_{\text{TS}}$   $O_2$ , (blue, dots) WOA13, (solid black, dots)  ${}^{\text{RF}}\text{SOSE}_{\text{TS}}$   $O_2$ , (dashed black, squares) SOSE, (gray squares)  ${}^{\text{RF}}\text{SOSE}_{\text{LSF}}$   $O_2$ , i.e., applying the LSF mapping to actual SOSE  $O_2$  at the locations of Argo T/S profiles. Red and black dotted lines in Figure 10a are for (red)  ${}^{\text{RF}}\text{Argo}_{\text{TS}}$   $O_2$  and (black)  ${}^{\text{RF}}\text{SOSE}_{\text{TS}}$   $O_2$  based on eight predictors.

show that the LSF gridding is not causing a bias in the estimate of the annual cycle (i.e., the squares align with the black dashed line from SOSE).

Results for the RF with 8 predictors are also shown in Figures 9 and 10 (dotted lines) and are similar to those for the RF with 6 predictors. For the eight predictors case,  $\Delta_{\text{RF}_g}^*$  shows some differences from Figure 2b (supporting information Figure S4). Yet percentages of domain area in Table 2 are comparable (i.e., differences are within few percent).

#### 4.3. Results from Different Sampling Scenarios

As described in section 4.2, Figure 2b shows a large bias in regions close to continental boundaries. Yet adding to the data set dedicated sections near continents (blue dots in Figure 2a, visited twice a year) results in an estimate ( ${}^{\text{RF}_g}\text{SOSE}_{\text{TS}}$   $O_2$ ) with larger bias than  ${}^{\text{RF}}\text{SOSE}_{\text{TS}}$   $O_2$  (Table 2). This larger bias originates from including a relatively large amount of data in (narrow) low  $O_2$  regimes which are not representative of the

**Table 3**

Comparison of  ${}^{\text{RF}_g}\text{SOSE}_{\text{TS}}$   $O_2$  With the Actual SOSE for the Month of October in 2008–2012: Difference in Area-Weighted Spatial Average of  $O_2$  ( $\Delta_1$ ,  $\mu\text{mol kg}^{-1}$ ); Square Root of the Area-Weighted SE for Different Estimates After Removing the Area-Weighted Spatial Average From Each Field ( $\Delta_2$ ,  $\mu\text{mol kg}^{-1}$ ); Ratio of  $\Delta_2$  and  $\sigma_{\text{SOSE}} = 5.52 \mu\text{mol kg}^{-1}$ , and Percentage of the Total Area With Relative Absolute Errors in 5-year Mean Estimates (After Removing the Spatial Average) Below 1%, 1.5%, 2%, and 3% (Area, %)

Scenario	Parameter	Number of profiles in October		
		$n$	$2n$	$4n$
R1. Real distribution	$\Delta_1$	0.66		
	$\Delta_2$	10.32		
	$\Delta_2/\sigma_{\text{SOSE}}$	1.87		
	Area	30, 43, 55, 72		
A1. Additional profiles randomly distributed	$\Delta_1$		-0.28 (-0.91, 0.42)	-0.20 (-0.51, 0.15)
	$\Delta_2$		8.29 (8.10, 8.57)	7.21 (6.94, 7.56)
	$\Delta_2/\sigma_{\text{SOSE}}$		1.50 (1.47, 1.55)	1.31 (1.26, 1.37)
	Area		41, 58, 70, 82	48, 65, 76, 86
A2. Additional profiles in sectors	$\Delta_1$		-0.18 (-0.62, 0.24)	-0.15 (-0.48, 0.16)
	$\Delta_2$		8.20 (8.08, 8.65)	7.25 (6.95, 7.44)
	$\Delta_2/\sigma_{\text{SOSE}}$		1.49 (1.46, 1.57)	1.31 (1.26, 1.35)
	Area		41, 58, 69, 82	47, 64, 74, 85

Note. Range of values from 10 Monte Carlo simulations is in parentheses. Scenarios use only October data in input to the RF, with  $n = 545$  profiles.

large-scale  $O_2$  in the Southern Ocean. Adding to the sections more data in the Atlantic (red dots in Figure 2a, with same spatial and temporal sampling as in the Indian Ocean) improves the estimate (see  $^{RF_{92}}SOSE_{TS} O_2$  in Table 2). If  $O_2$  were available at all of the Argo T/S locations in the SO during 2008–2012, 94% of the domain would be characterized by a bias less than 2% in the 5-year mean (see  $^{RF_{9TS}}SOSE_{TS} O_2$  in Table 2).

We now describe how the RF method performs in terms of  $\Delta_1$  and  $\Delta_2$  for different sampling scenarios of model profiles, focusing on the month of October (see section 3.3). RF trained on October profiles available in 2008–2012 ( $n = 545$  profiles, scenario R1) results in  $\Delta_1 = 0.66 \mu\text{mol kg}^{-1}$  (Table 3). With a higher saturation of the spatial field at sample sizes  $2n$  and  $4n$ , biases reduce (up to four times) under both scenarios A1 and A2. RF predictions for October tend to underestimate the mean  $O_2$ , yet estimates from scenario A2 are closer to the actual SOSE field.

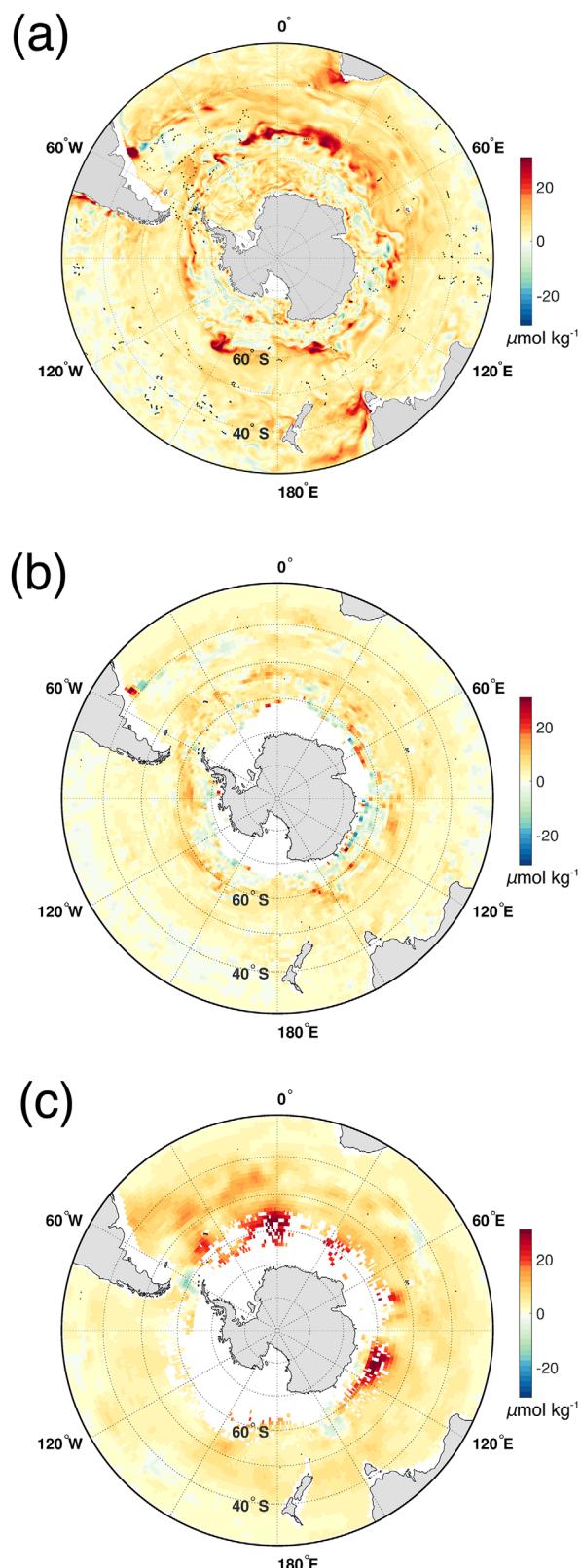
The uncertainty ( $\Delta_2$ ) in estimating the spatial structure of  $O_2$  (once the spatial mean has been removed) has similar reduction from scenario R1 to scenarios A1 and A2 with  $2n$ , and then with  $4n$ . However, the reductions are about 20% for the  $n$  to  $2n$  transition, and about 12% for the transition from  $2n$  to  $4n$  (Table 3). Scenario A2 shows comparable or better performance than scenario A1, consistent with T/S being the main predictors for  $O_2$  from the RF regression, and sampling in space and time being secondary to sampling more T/S pairs overall.

## 5. Discussion and Conclusions

Modern statistical and supervised machine learning tools can be used to map  $O_2$  where T/S, and sufficient  $O_2$  data are available. We estimate  $O_2$  at 150 m during 2008–2012, based on observed  $O_2$ , T/S Argo profiles and a RF regression method. The same RF-based estimate is also computed from model profiles (colocated with observations) to evaluate errors in the method by comparison with the actual model solution (SOSE). Method performance is poor in some of the boundary regions (e.g., regions of small spatial scales associated with upwelling and lack of data, Figure 2b) and, in general, where T/S profiles are sparse. Yet in  $\approx 65\%$  of the domain, differences in the 2008–2012 mean between the estimated  $^{RF_9}SOSE_{TS} O_2$  and SOSE  $O_2$  are less than 1.5% of the actual value, and  $^{RF}Argo_{TS} O_2$  can inform biases in SOSE and WOA13 in at least  $\approx 55\%$  of the domain (Table 2).

Overall,  $^{RF}Argo_{TS} O_2$  suggests that SOSE and WOA13 may overestimate annual mean  $O_2$  at 150 m in the SO, both on a global and basin scale (Figures 1a–1c, 2c, 2d, 9, and 10b).  $^{RF}Argo_{TS} O_2$  is closer to WOA13 than SOSE, except for a few regions. For example, south of  $\approx 62^\circ\text{S}$ , globally and in the Indian and Pacific sectors of the SO (Figures 9b–9d), where  $O_2$  observations are sparse. A large regional difference in annual mean between Argo-based  $O_2$  and other products is noticed east of Argentina, as  $^{RF}Argo_{TS} O_2$  high  $O_2$  is confined to the coast (Figures 1c, 2c, and 2d). Differences between  $^{RF}Argo_{TS} O_2$  and SOSE reflect differences in T/S (supporting information Figure S5, e.g., 2008–2012 time mean in SOSE is colder and fresher and characterized by high  $O_2$  east of Argentina compared to Argo). Also, when comparing Argo and WOA13, some of the differences are related to WOA13 including data for a much longer period (and related differences in T/S, supporting information Figure S6).

$^{RF}Argo_{TS} O_2$  also suggests that SOSE may underestimate the annual cycle of  $O_2$  at 150 m (Figure 10a). While  $O_2$  annual variability based on Argo T/S has similar amplitude to SOSE in the SO (north of  $60.5^\circ\text{S}$ , Figure 10a), the RF regression method may be biased low and the real signal may be stronger. The RF-based estimate from model profiles is, indeed, characterized by half the variability in actual SOSE (Figure 10a). This emerges also in maps of the 12-month  $O_2$  anomaly from the 5-year mean (e.g., Figures 11a and 11b).  $^{RF_9}SOSE_{TS} O_2$  represents well the spatial distribution of SOSE anomaly in most regions (except e.g., some of the boundary regions as discussed previously), yet the amplitude is smaller. The number of profiles available varies across months (Figure 8f), yet is limited for all (and especially for October) and characteristic localized regions with relatively larger annual variability may be undersampled (Figure 11a, sparse black dots). Except for some localized regions,  $^{RF}Argo_{TS} O_2$  October anomaly is larger than in SOSE (Figure 11c). In Figure 11c, locations south of  $55^\circ\text{S}$  are masked out if less than five Argo T/S observations are available in a  $1^\circ \times 1^\circ$  (latitude/longitude) bin centered at the grid point of interest, since  $O_2$  estimates here may reflect T/S biases.



**Figure 11.**  $\text{O}_2$  October anomaly from the 5-year mean,  $\mu\text{mol kg}^{-1}$ . (a) SOSE, (b)  ${}^{\text{RF}}\text{SOSE}_{\text{TS}}$   $\text{O}_2$ , and (c)  ${}^{\text{RF}}\text{Argo}_{\text{TS}}$   $\text{O}_2$ . Black dots in Figure 11a indicate the location of Argo  $\text{O}_2$  profiles in October. In Figure 11c, locations south of  $55^\circ\text{S}$  are masked out if less than five Argo T/S observations are available in a  $1^\circ$  bin centered at the grid point of interest.

Experiments with different sampling scenarios show reduction of bias and variance in RF (T/S-based) estimates with growing sample size, unless a relatively large amount of data is included from small regions that are not representative of the large-scale O<sub>2</sub> distribution. The estimate does not necessarily improve when the additional profiles are randomly distributed in the SO (Table 3). Although this result is counterintuitive, it should be considered in the context of the RF variable importance analysis (Figure 7). The main predictors for O<sub>2</sub> from RF regression are T/S, hence sampling in more locations in space and time may provide redundant information. Latitude, longitude, month, and year provide information regarding O<sub>2</sub>, however this relationship is not causal. The importance associated with them in the RF originates from atmospheric and oceanic forcing, which is also manifest in T/S structure. However these coordinate predictors may contribute to improving representation of phenomena (e.g., biological activity) that are not directly accounted for in T/S, and this will likely be resolvable as the number of profiles increase. Larger importance of biological activity (compared to ventilation, hence T/S) will also be more apparent at shallower depths. Similarly, we find that including sea surface height and surface chlorophyll from satellite data in the set of predictors does not lead to a significant improvement of the 150 m O<sub>2</sub> estimates, because information about those variables is already partially represented by other predictors in the RF (i.e., T/S, longitude, latitude, month, and year). The six-predictor RFs have the advantage of being computationally less expensive. The importance of Chl may increase as the O<sub>2</sub> data coverage (for the training dataset) extends to more regions, e.g., west of South America.

While highly accurate (yet sparse in space and time) ship-based measurements are crucial for calibration of profiling floats (for O<sub>2</sub> as well as T/S), profiling floats have great potential to measure O<sub>2</sub> and other biogeochemical variables in the 3-D ocean, with no seasonal bias. An extensive spatial and temporal coverage of O<sub>2</sub> profiling floats remains the long-term goal for the scientific community in order to facilitate basin scale studies from high quality O<sub>2</sub> observations. The framework presented here has the potential to improve our understanding of O<sub>2</sub> annual mean fields and variability with the limited measurements available at present. As the data set grows, the RF method will remain a valuable tool for multivariate mapping. In fact, even with a large size of the O<sub>2</sub> data set, using cross-covariances of O<sub>2</sub> and T/S should provide a better estimate of O<sub>2</sub> fields, compared to mapping O<sub>2</sub> observations alone. The framework here can also be used to simulate sampling scenarios and guide enhancements to the current array of O<sub>2</sub> profiling floats (e.g., SOCCOM floats; Johnson et al., 2017). Finally, this method may prove useful for other biogeochemical variables (e.g., nutrients and carbon).

#### Acknowledgments

This work was sponsored by NSF's Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project under the NSF award PLR-1425989, with additional support from NOAA and NASA. The Argo data used here were collected and made freely available by the International Argo Program and by the national programs that contribute to it (<https://doi.org/10.17882/42182>). The Roemmich and Gilson (2009) climatology can be found at [http://sio-argo.ucsd.edu/RG\\_Climatology.html](http://sio-argo.ucsd.edu/RG_Climatology.html); oxygen profiles with bias correction by Drucker and Riser (2016) at <http://runt.ocean.washington.edu/o2/>; the Southern Ocean State Estimate at [http://sose.ucsd.edu/BSOSE\\_iter105\\_solution.html](http://sose.ucsd.edu/BSOSE_iter105_solution.html); the World Ocean Atlas Climatology at <https://www.nodc.noaa.gov/OC5/woa13/>; the Mixed Layer Argo product by Holte et al. (2017) at <http://mixedlayer.ucsd.edu/>. Thanks to Lynne Talley and members of the working group on Statistical Oceanography within the SAMSI Program on Mathematical and Statistical Methods for Climate and the Earth System (CLIM) for helpful discussions. Thanks to the two anonymous reviewers for helpful suggestions.

#### References

- Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, 147, 77–89.
- Altabet, M. A. (2007). Constraints on oceanic N balance/imbalance from sedimentary <sup>15</sup>N records. *Biogeosciences*, 4, 75–86.
- Berk, R. A. (2016). *Statistical learning from a regression perspective* (2nd ed.). Basel, Switzerland: Springer International Publishing.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L., Friedman, J. H., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. New York, NY: CRC Press.
- Bushinsky, S. M., & Emerson, S. R. (2015). Marine biological production from in situ oxygen measurements on a profiling float in the subarctic Pacific Ocean. *Global Biogeochemical Cycles*, 29, 2050–2060. <https://doi.org/10.1002/2015GB005251>
- Carter, B. R., Williams, N. L., Gray, A. R., & Feely, R. A. (2016). Locally interpolated alkalinity regression for global alkalinity estimation. *Limnology and Oceanography: Methods*, 14, 268–277.
- Chapman, C. (2017). New perspectives on frontal variability in the Southern Ocean. *Journal of Physical Oceanography*, 47, 1151–1168.
- Chapman, C., & Charantonis, A. A. (2017). Reconstruction of subsurface velocities from satellite observations using iterative self-organizing maps. *IEEE Geoscience and Remote Sensing Letters*, 14, 617–620.
- Diaz, R. J., & Rosenberg, R. (2008). Spreading dead zones and consequences for marine ecosystem. *Science*, 321, 926–929.
- Drucker, R., & Riser, S. C. (2016). In situ phase-domain calibration of oxygen Optodes on profiling floats. *Methods in Oceanography*, 17, 296–318.
- Emerson, S. R., & Bushinsky, S. (2014). Oxygen concentrations and biological fluxes in the open ocean. *Oceanography*, 27, 168–171.
- Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, 56, 229–238.
- Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., et al. (2014). World Ocean Atlas 2013, volume 3: Dissolved oxygen, apparent oxygen utilization, and oxygen saturation. In S. Levitus (Ed.), A. Mishonov (Tech. Ed.), NOAA atlas NESDIS (Vol. 75, 27 pp.). Silver Spring, MD: National Oceanic and Atmospheric Administration.
- Giglio, D., & Johnson, G. C. (2016). Subantarctic and polar fronts of the Antarctic Circumpolar Current and Southern Ocean heat and freshwater content variability: A view from Argo. *Journal of Physical Oceanography*, 46, 749–768.
- Giglio, D., & Johnson, G. C. (2017). Middepth decadal warming and freshening in the South Atlantic. *Journal of Geophysical Research: Oceans*, 122, 973–979. <https://doi.org/10.1002/2016JC012246>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

- Holte, J., Talley, L. D., Gilson, J., & Roemmich, D. (2017). An Argo mixed layer climatology and database. *Geophysical Research Letters*, 44, 5618–5626. <https://doi.org/10.1002/2017GL073426>
- Johnson, K. S., Plant, J. N., Coletti, L. J., Jannasch, H. W., Sakamoto, C. M., Riser, S. C., et al. (2017). Biogeochemical sensor performance in the SOCCOM profiling float array. *Journal of Geophysical Research: Oceans*, 122, 6416–6436. <https://doi.org/10.1002/2017JC012838>
- Kamikawaiji, Y., Matsuyama, H., Fukui, K., Hosoda, S., & Ono, S. (2016). *Decision tree-based feature function design in conditional random field applied to error detection of ocean observation data*. Paper presented at the Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI).
- Kutkina, O., & Feuerriegel, S. (2016). *Deep learning in R*. Albert-Ludwigs-Universität Freiburg R Blog. March 7, 2016.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.
- Mazloff, M., Heimbach, P., & Wunsch, C. (2010). On the meridional extent and fronts of the Antarctic Circumpolar Current. *Journal of Physical Oceanography*, 40, 880–899.
- Orsi, A. H., Whitworth, T., & Nowlin, W. D. (1995). An eddy-permitting Southern Ocean State Estimate. *Journal of Physical Oceanography*, 42, 641–673.
- Roemmich, D., & Gilson, J. (2009). The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program. *Progress in Oceanography*, 82, 81–100.
- Sauzède, R., Bittig, H. C., Claustre, H., Pasqueron de Fommervault, O., Gattuso, J. P., Legendre, L., et al. (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, 128.
- Sauzède, R., Claustre, H., Jamet, C., Uitz, J., Ras, J., Mignot, A., et al. (2015). Retrieving the vertical distribution of chlorophyll *a* concentration and phytoplankton community composition from in situ fluorescence profiles: A method based on a neural network with potential for global-scale applications. *Journal of Geophysical Research: Oceans*, 120, 451–470. <https://doi.org/10.1002/2014JC010355>
- Sauzède, R., Claustre, H., Uitz, J., Jamet, C., Dall'Olmo, G., D'Ortenzio, F., et al. (2016). A neural network-based method for merging ocean color and Argo data to extend surface bio-optical properties to depth: Retrieval of the particulate backscattering coefficient. *Journal of Geophysical Research: Oceans*, 121, 2552–2571. <https://doi.org/10.1002/2015JC011408>
- Takeshita, Y., Martz, T. R., Johnson, K. S., Plant, J., Gilbert, D., Riser, S., et al. (2013). A climatology-based quality control procedure for profiling float oxygen data. *Journal of Geophysical Research: Oceans*, 118, 5640–5650. <https://doi.org/10.1002/jgrc.20399>
- Verdy, A., & Mazloff, M. (2017). A data assimilating model for estimating Southern Ocean biogeochemistry. *Journal of Geophysical Research: Oceans*, 122, 6968–6988. <https://doi.org/10.1002/2016JC012650>
- Williams, N. L., Juranek, L. W., Johnson, K. S., Feely, R. A., Riser, S. C., Talley, L. D., et al. (2016). Empirical algorithms to estimate water column pH in the Southern Ocean. *Geophysical Research Letters*, 43, 3415–3422. <https://doi.org/10.1002/2016GL068539>
- Williams, N. L., Juranek, L. W., Feely, R. A., Johnson, K. S., Sarmiento, J. L., Talley, L. D., et al. (2017). Calculating surface ocean pCO<sub>2</sub> from biogeochemical Argo floats equipped with pH: An uncertainty analysis. *Global Biogeochemical Cycles*, 31, 591–604. <https://doi.org/10.1002/2016GB005541>