

**Random forest regression on multi-platform in-situ ocean observations:
Investigating high-frequency nutrient dynamics in the Southern Ocean**

Sangmin Song^a, Paige D. Lavin^{b,c}, Alison R. Gray^a, Zachary Nachod^d, Dhruv Balwada^e, Lilian
A. Dove^f, Andrew F. Thompson^g

^a *School of Oceanography, University of Washington, Seattle, WA.*

^b *Cooperative Institute for Satellite Earth System Studies/Earth System Science Interdisciplinary
Center (CISESS/ESSIC), University of Maryland, College Park, MD.*

^c *NOAA/NESDIS Center for Satellite Applications and Research, College Park, MD.*

^d *University of Hawai‘i at Mānoa, Honolulu, HI.*

^e *Lamont-Doherty Earth Observatory, Columbia University, New York City, NY.*

^f *Brown University, Providence, RI.*

^g *Environmental Science and Engineering, California Institute of Technology, Pasadena, CA.*

Corresponding author: Sangmin Song, sangsong@uw.edu

14 ABSTRACT: Nutrient cycling in the ocean is mediated by physical mixing processes that span
15 diverse spatial and temporal scales. New biogeochemical profiling floats (BGC-Argo) have begun to
16 observe nutrient distributions globally, but their 10-day cycling period limits the types of processes
17 they can capture. Small-scale dynamics, occurring on $O(1)$ day and $O(1)$ km, remain particularly
18 difficult to observe in-situ. Here, we show that random forest regression (RFR) can recover high-
19 frequency information by leveraging the sampling strategies of multiple ocean profilers. Our
20 RFR is trained, validated, and tested on BGC-Argo and shipboard data to within $\sim 3\%$ accuracy,
21 then applied to observations from two rapid-sampling Seagliders deployed during the Southern
22 Ocean Glider Observations of the Submesoscale (SOGOS) experiment in 2019. This approach
23 generates novel nitrate distributions at 50 times the horizontal resolution of the original float data.
24 Using the high-resolution RFR outputs, we identify signatures of nutrient injection into the mixed
25 layer that coincide with enhanced stirring in a turbulent region downstream of the Southwest
26 Indian Ridge. Relating these intermittent transport events to biological time series suggests that
27 small-scale stirring mediates additional nutrient drawdown and primary production in this region.
28 In our exploration of high-frequency nitrate variability in the Southern Ocean, RFR extends the
29 capabilities of the BCG-Argo array and allows for deeper understanding of biogeochemical cycling
30 at a more comprehensive set of scales. As a flexible approach that can be generalized to suit
31 other multi-platform observing systems, RFR presents new opportunities to maximize value from
32 existing datasets.

33 SIGNIFICANCE STATEMENT: We use a regional random forest regression (RFR) to leverage
34 data from multiple ocean observing instruments that offer different advantages. In our study of
35 the Southern Ocean, we use RFR to produce new nutrient maps at 50 times higher resolution
36 than previously possible. By estimating small-scale information, RFR reveals interactions between
37 physical and biological processes during rapid mixing events that are normally difficult to observe.
38 These short-lived interactions appear to be important in determining local nutrient content and
39 therefore biological activity in this important ocean basin. Similar machine learning approaches
40 that can be applied regionally will extend possibilities for how we use data from increasingly
41 advanced ocean platforms.

42 1. Introduction

43 The ocean is mixed by physical processes spanning a wide range of temporal and spatial scales,
44 which mediate biogeochemical interactions in distinct ways. Nutrient distributions differ at basin-
45 wide scales and are broadly determined by the global circulation. At the mesoscale (horizontal
46 scales on $O(20\text{--}200\text{ km})$, evolving over weeks to months), lateral and vertical nutrient transport
47 processes are associated with coherent eddies that modify the density structure of the upper ocean
48 (Su et al. 2021; Patel et al. 2020; Levy and Martin 2013; Mahadevan and Archer 2000). These
49 widespread features subject local nutrient distributions to competing processes that are both depth-
50 dependent (biological utilization and remineralization) and density-dependent (transport along
51 isopycnals; Omand and Mahadevan (2013)). When eddy-associated motions transport nutrients
52 across the base of the mixed layer, increased availability in the biologically active surface layer can
53 enable more primary production if that nutrient was previously limited.

54 Over smaller distances and shorter periods, nutrient distributions are also impacted by high-
55 frequency motions at the submesoscale (scales of $O(1)$ day and $O(1)$ km; Lévy et al. (2024, 2018);
56 Mahadevan (2016); Lévy et al. (2012); Thomas et al. (2008); Mahadevan and Tandon (2006)).
57 Submesoscale dynamics are associated with enhanced vertical velocities, typically localized to
58 filaments and fronts, that cause vigorous exchange between the interior and surface ocean (Taylor
59 and Thompson 2023; Freilich and Mahadevan 2019; Thomas et al. 2008; Brannigan 2016). These
60 processes can exert a strong influence on local primary production and carbon export since light
61 and nutrient availability change rapidly with depth (Lévy et al. 2018; Klein and Lapeyre 2009;

62 Erickson and Thompson 2018). Nutrient transport at submesoscales is particularly important for
63 mediating biological processes since motions at these scales occur on the same timescales as
64 phytoplankton growth (Mahadevan 2016). Vertical nutrient mixing along strongly tilted density
65 surfaces, including at submesoscale fronts, may be a significant pathway for nutrient fluxes in
66 highly turbulent regions depending on the structure of the nutricline (Freilich and Mahadevan
67 2019).

68 The Southern Ocean is one such energetic region, enriched in mesoscale eddies and submesoscale
69 stirring, that has outsized importance in the global climate and carbon system due to its distinctive
70 physical circulation and biogeochemical distribution (Gray 2024; Henley et al. 2020; Rintoul and
71 Naveira Garabato 2013; Talley 2013). Model simulations of Southern Ocean dynamics have
72 suggested that submesoscale stirring is a significant mechanism by which deep waters, enriched in
73 iron and other remineralized nutrients like nitrate and phosphate, exchange tracers into the surface
74 mixed layer (Uchida et al. 2019; Balwada et al. 2018). One experiment found that increasing the
75 resolution of a Southern Ocean biogeochemical model from 100 km to 2 km resulted in a nearly
76 two-fold increase in production due to nutrient injection into the mixed layer (Uchida et al. 2020).
77 Other numerical simulations suggest that where there are strong flow-topography interactions,
78 submesoscale upwelling may be the predominant mechanism by which dissolved iron reaches the
79 upper Southern Ocean (Rosso et al. 2016). These turbulent contributions to the nutrient supply
80 have global effects since the Southern Ocean is a central site of water mass formation (Talley
81 2013). Unfortunately, the challenges of sampling at the resolution required to observe these
82 small-scale dynamics have restricted our ability to constrain high-frequency nutrient fluxes using
83 in-situ observations. The impact of short-lived, filamentary mixing processes on upper nutrient
84 distributions remains poorly constrained using observations.

85 New observing technologies have emerged in the last twenty years to address this knowledge
86 gap, including autonomous profiling instruments called Argo floats. The deployment of thousands
87 of Argo floats has enabled new observation-based studies in previously undersampled regions (e.g.
88 Wong et al. (2020); Swart et al. (2023)). Some newer Argo floats, known as Biogeochemical-
89 Argo (BGC-Argo) floats, have been further equipped with biologically relevant sensors for oxygen,
90 nitrate, optical backscatter, and chlorophyll fluorescence (Claustre et al. 2020; Sarmiento et al.
91 2023; Roemmich et al. 2019). Although the Argo floats provide invaluable coverage of the global

oceans, their characteristic profiling period of 10 days makes them more suitable for observing processes at the mesoscale and larger. Ocean gliders are a different type of autonomous instruments that sample much more rapidly than floats, returning profiles of the ocean interior every 4–6 hours, and have provided novel observation-based insights into submesoscale physical mixing processes (e.g. Rudnick (2016); Thompson et al. (2016); Erickson et al. (2016); Erickson and Thompson (2018); Balwada et al. (2024)). Recent gliders have been equipped with sensors for oxygen, nitrate, phosphate, pH, and optical properties (e.g. Possenti et al. (2021); Vincent et al. (2018); Birchill et al. (2021)), but most Seagliders to date have not been deployed with a full suite of biogeochemical sensors due to challenges with power demand, temporal responsiveness, and measurement drift (Chai et al. 2020).

Here, we leverage these different sampling strategies using random forest regression (RFR) in order to characterize small-scale, biogeochemical-physical interactions from in-situ ocean observations. Focusing specifically on nitrate variability in a turbulent region of the Southern Ocean, we design a RFR model to maximize the information gained by two Seagliders deployed during the Southern Ocean Glider Observations of the Submesoscale (SOGOS) mission in 2019. The regional RFR is trained using nearby ship-based and BGC-Argo observations, including one BGC-Argo float deployed in tandem with the SOGOS Seagliders. After validating model performance, we apply the RFR to predict high-resolution, depth-resolved nitrate fields along the glider tracks and identify new signs of enhanced nitrate variability at high frequencies.

The observational data are described in Section 2, followed by an explanation of the RFR methodology and performance evaluation in Section 3a–b. Methods of characterizing the RFR nitrate using time series and wavelet analysis are given in Section 3c and Section 3d, respectively. In Section 4a, we demonstrate the strong performance of our Southern Ocean RFR before applying it to Seaglider observations to generate novel, high-resolution nitrate maps. Section 4b highlights relationships between different physical and biological time series in the upper ocean, while Section 4c explores what timescales of variability are important for nutrient transport events. We conclude with broader implications of this work in Section 5 by commenting on the utility of deploying heterogenous in-situ platforms together, as well as the applicability of RFR for bridging data gaps in other observing systems.

2. Data

During the 2019 SOGOS experiment, two Seagliders (SG659 and SG660) were deployed alongside a BGC-Argo float (WMO 5906030; hereafter the SOGOS float) on the I06 Global Ship-based Hydrographic Investigations Program (GO-SHIP) cruise. The two Seagliders were outfitted with temperature (T), salinity (S), and pressure (p) sensors on an unpumped CTD (Conductivity-Temperature-Depth instrument; CT-Sail) and an oxygen (O_2) optode (Aanderaa 4330 standard foil). Chlorophyll fluorescence and optical backscatter data at 470 nm and 700 nm were also collected (WETLabs ECO puck). GO-SHIP stations provided T and S measurements as well as discrete nitrate and oxygen concentrations from bottle data. The SOGOS BGC-Argo float, as well as the other six BGC-Argo floats in this region (see Appendix) that are used for RFR training, were equipped with sensors for T, S, p, O_2 as well as nitrate, fluorescence, and backscatter. To match the vertical range of the Seagliders, only BGC-Argo and GO-SHIP observations down to depths of 1000 m are used. For all platforms, time is reported as days elapsed since January 1, 2019 (yearday). Conservative temperature (CT), absolute salinity (SA), potential density(σ_0), spice (τ), Brunt-Väisälä buoyancy frequency (N^2), and oxygen saturation (O_{2sat}) were calculated using the Thermodynamic Equation of Seawater 2010 (TEOS-10) Python toolbox (IOC et al. 2010). Processing of all platform data is further described in the Appendix.

The Seagliders sampled for 86 days, from May 1, 2019 to July 25, 2019, covering a region spanning approximately 30–40°E and 50–54°S (Figure 1a). SG659 and SG660 completed 456 and 502 V-shaped dives, respectively, to ~1000 m depth and sampled during both the descent and ascent. The gliders generally surfaced every 4–6 hours (profiles every 2–3 hours) while the BGC-Argo float sampled with a 5-day profiling frequency (16 profiles within the duration of the Seaglider missions). After being deployed along the 30°E I06 cruise track around –51.5°S, the three SOGOS profilers were advected eastward by the Antarctic Circumpolar Current (ACC). The Seagliders pass through a standing meander region by the Southwest Indian Ridge (SWIR) on yeardays 120–150, which is a known hotspot of enhanced eddy kinetic energy (EKE) (Balwada et al. 2024; Yung et al. 2022). Further details on the deployment and trajectories of the Seagliders are described in Dove et al. (2021).

Surface EKE over the SOGOS deployment is characterized using the delayed-time multi-satellite gridded product (1/4° resolution) for sea level heights and derived variables pro-

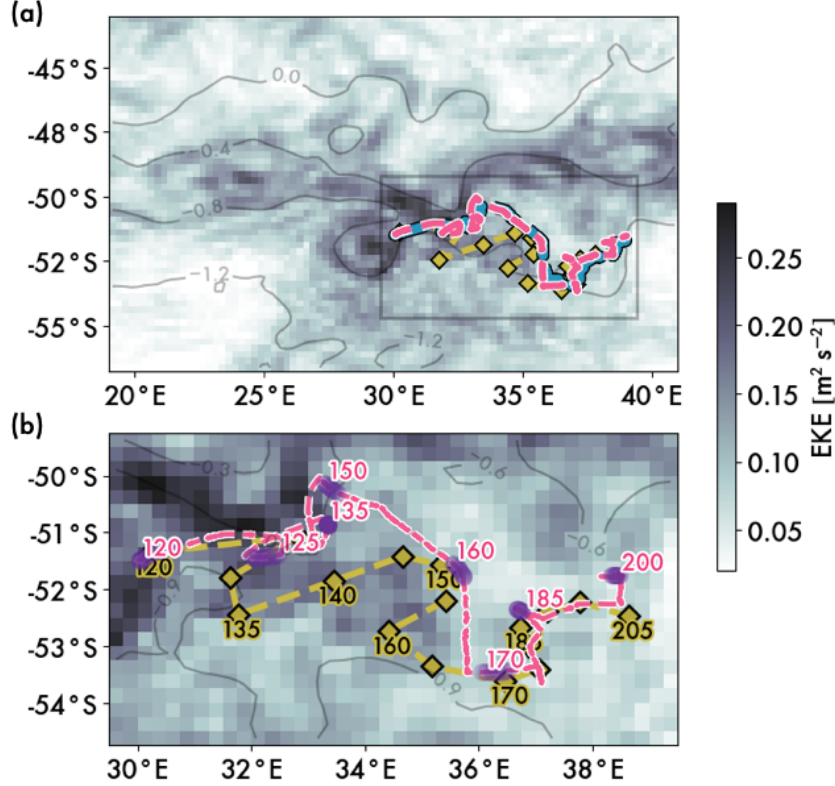


FIG. 1: (a) Trajectories for Seagliders SG659 (blue) and SG660 (magenta) and SOGOS float profile locations (yellow diamonds). Grey box denotes the area in panel (b), showing an expanded view of the float (yellow) and glider SG660 (magenta) as they traverse from the high EKE region (yeardays 120–150) to the low EKE region (yeardays 170–200). Magenta text indicates yearday of the subset of observations highlighted in purple. Background colored by eddy kinetic energy (EKE, $\text{m}^2 \text{s}^{-2}$) averaged over the Seaglider deployment period.

duced and distributed by Copernicus Marine Environment Monitoring Service (CMEMS; <https://doi.org/10.48670/moi-00148>). The altimetry product provides zonal and meridional geostrophic velocity anomalies (u'_g and v'_g , respectively) from which EKE is calculated as $\frac{1}{2}\sqrt{u_g'^2 + v_g'^2}$. Following the framework in Dove et al. (2021), we delineate a high EKE region (yeardays 120–150; $\text{EKE} > 0.124 \text{ m}^2 \text{s}^{-2}$) and a low EKE region (yeardays 170–200; $\text{EKE} < 0.109 \text{ m}^2 \text{s}^{-2}$) within the SOGOS deployment.

In our analysis of mixed layer properties along the glider trajectories, we utilize satellite-based estimates of photosynthetically active radiation (PAR) from Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data at 4 km resolution in 8-day composites (NASA/GSFC OBPG; Dataset ID: erdMH1par08day). To determine where gliders may be sampling across subme-

161 soscale fronts, we also use the finite-size Lyapunov exponent (FSLE) value-added L4 product from
162 delayed-time merged Global Ocean Gridded Absolute Geostrophic Velocities (1/25° resolution;
163 backward-in-time) provided by AVISO (DOI 10.24400/527896/a01-2022.002). FSLE values have
164 been shown to correspond to localized regions of high strain rate; filaments of large negative FSLEs
165 (computed backward-in-time) indicate lines of strong stretching that constrain fluid motion and
166 transport around coherent structures, including along submesoscale fronts and around mesoscale
167 eddies (d’Ovidio et al. 2004; Siegelman et al. 2020).

168 **3. Methods**

169 *a. Random Forest Regression*

170 Random forest regression (RFR) is a supervised machine learning method that uses an ensemble
171 (forest) of n “weak” learners (decision trees) to make strong predictions (Breiman 2001). The RFR
172 is given a set of observed predictor variables (“features”) as inputs, and a split condition based on
173 a feature is determined at each branch of each decision tree. During training, each split criterion
174 aims to best separate the target observations (here, of nitrate) into branches, or regions R_j , with
175 the least variance. Each resulting region after splitting is assigned the average target value \hat{y}_{R_j} .
176 After training is complete, new input observations are sorted into their respective R_j ’s using the
177 determined conditions, and then assigned \hat{y}_{R_j} as the predicted target value. The decision at each
178 node split in RFR is limited to a random subset of all features provided. As a result, the trees
179 of RFR are less correlated than those in a family of “bagged” trees which consider all possible
180 features at all nodes. RFR also uses “bootstrapping”, or subsampling data with replacement, so
181 that each decision tree is trained on a slight variation of training data. By introducing randomness
182 both in the node features considered and in the training data used for each decision tree, RFR is
183 less prone to overfitting on data that are not randomly distributed in space and time (Stock 2022;
184 Sharp et al. 2022b).

185 RFR returns a measure for each feature called “feature importance” that reflects its predictive
186 value. For each split where a given feature is used to determine the split criterion, the difference in
187 the variance of the pre-split node compared to the two post-split nodes is calculated (and weighted
188 by the number of samples in the pre-split node). These values are summed across all of the relevant
189 splits in a given decision tree; the average of these sums across all trees represents the feature

190 importance of the given variable. The feature importances are computed for all variables used in
 191 the model, and are scaled relative to each other so that the sum of all of the final feature importances
 192 is 1. The feature importances are often screened during RFR development to help select a final set
 193 of features.

194 Since nitrate variability is dependent on a wide range of physical, chemical, and biological
 195 factors, we explored the use of many possible “feature lists” (the set of predictor features provided
 196 for RFR) for our RFR. Predictive features such as CT, SA, σ_0 , τ , N^2 , O_2 , O_2sat , as well as
 197 coordinate features such as latitude (lat), longitude (lon), time, and season (sz_1 and sz_2) were
 198 considered in different versions of the model. τ is a measure of temperature and salinity variations,
 199 and acts as a tracer of water masses in the interior ocean. We applied a logarithmic transformation
 200 to the N^2 data since its distribution is initially non-Gaussian. Seasonal information was encoded
 201 as two training variables, sz_1 and sz_2 , following Sharp et al. (2022b):

$$\begin{aligned} sz_1 &= \cos(2\pi * yd/365), \\ sz_2 &= \sin(2\pi * yd/365), \end{aligned} \tag{1}$$

202 where yd refers to the yearday (days elapsed) referenced to January 1, 2019. The two variables
 203 together add a cyclical signal representing the time of year, or season. Of the feature lists considered,
 204 results from a selection of seven models with increasing complexity are presented in Section 4a.1.
 205 Each RFR was constructed using the scikit-learn Python RandomForestRegressor package with
 206 $n = 1000$ trees; a random one-third subset of the features were considered at each node (Breiman
 207 2001), with a minimum node sample size of 5 for a split to be considered.

208 During RFR development, observations are separated into different datasets for training, vali-
 209 dation, and testing steps (Figure 2). We note that the observations within a single glider or float
 210 profile are highly correlated; since the individual profiles are the “independent” units here rather
 211 than the pointwise samples, we keep observations from vertical profiles together during data split-
 212 ting. Our RFR is trained and validated using GO-SHIP I06 bottle data and observations from six
 213 BGC-Argo floats (WMO: 5904469, 5904659, 5905368, 5905996, 5906031, 5906207; Figure 3a),
 214 which provided coverage of the Antarctic Southern Zone (ASZ) bounded by the Polar Front and
 215 Sea Ice Edge (Sauvé et al. 2023). The BGC-Argo float training observations cover a time period
 216 between May 8, 2017 and July 7, 2021, within approximately two years before and after the SO-

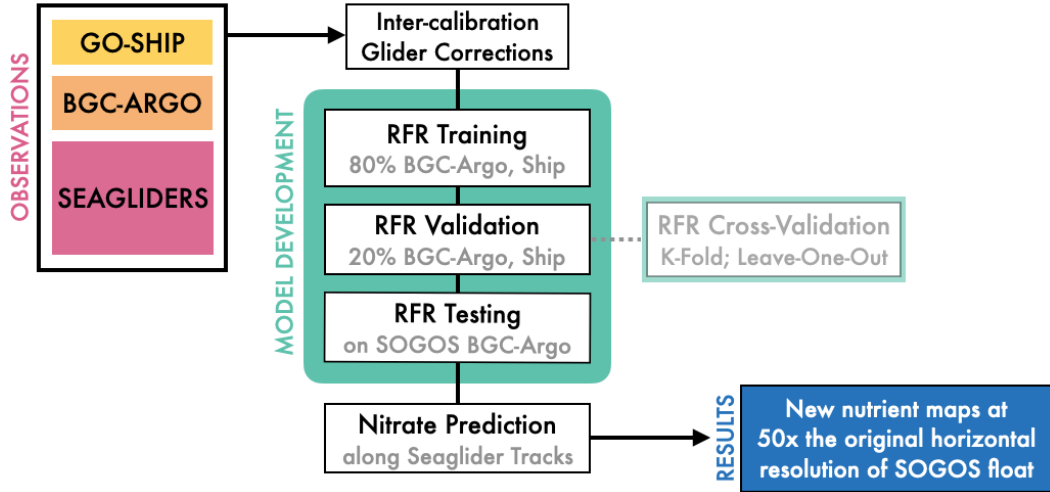


FIG. 2: Available ship, float, and glider observations are intercalibrated and quality controlled before model development. The regional GO-SHIP and BGC-Argo data (excluding the SOGOS float) are split into training and validation subsets, while the SOGOS float is withheld for testing. The RFR outputs are nitrate profiles with high horizontal resolution reaching depths of ~ 1000 m along the Seaglider tracks.

GOS glider deployment (Figure 3b). The combined GO-SHIP and BGC-Argo dataset is split into the training (80%) and validation (20%) datasets as is common for machine learning workflows. After the RFR model is developed on the training data, we compute validation errors on this 20% withheld validation data (hereafter called the “simple holdout” validation dataset; Figure 4a) to give a first-order estimate of model performance.

Using training observations that are spatiotemporally correlated can lead to validation error biases that obscure the true performance of the model when applied to data outside of the training dataset (Millard and Richardson 2015; Stock and Subramaniam 2022). We therefore rely on two other cross-validation techniques; the first is k-fold, in which RFR is validated iteratively on a random $1/k$ th sample, or “fold” of the data (Figure 4b). Combining results across all folds produces a more representative validation error distribution for each version of the model (Kohavi 1995). Additionally, we apply a variation of spatial leave-one-out (SLOO) cross-validation, which examines model performance by withholding all observations from one profiling float at a time (Figure 4c; e.g. Le Rest et al. (2014); Stock and Subramaniam (2022)).

Typical machine learning workflows construct test datasets by removing a small (1–10%) random subset of the data available for training. However, our goal for testing in this particular use case

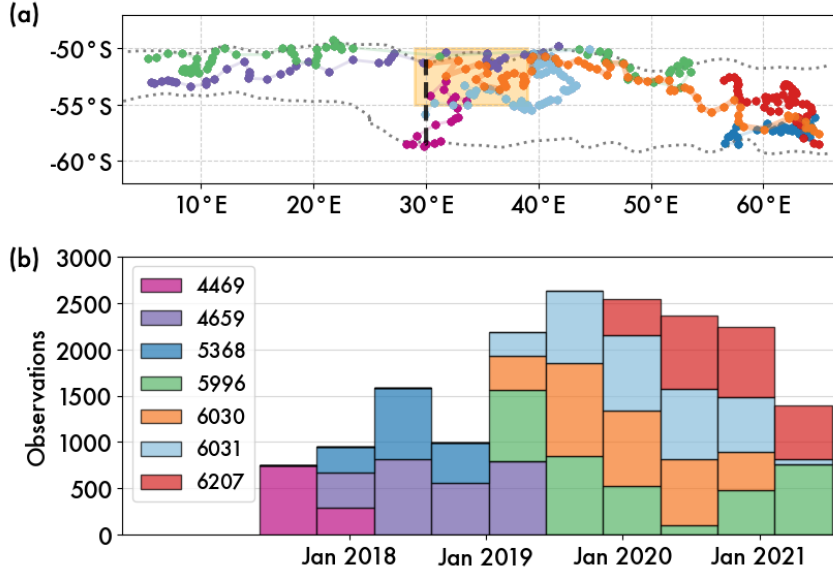


FIG. 3: (a) Location of BGC-Argo float profiles used for training and validation (WMO: 5904469, 5904659, 5905368, 5905996, 5906031, 5906207) and those used for testing (WMO: 5906030). GO-SHIP station locations in black diamonds. Polar Front and Sea Ice Edge 2019 mean front positions in dotted lines (Sauvé et al. 2023). (b) Time coverage by the BGC-Argo training floats; “590” truncated from all float labels in legend.

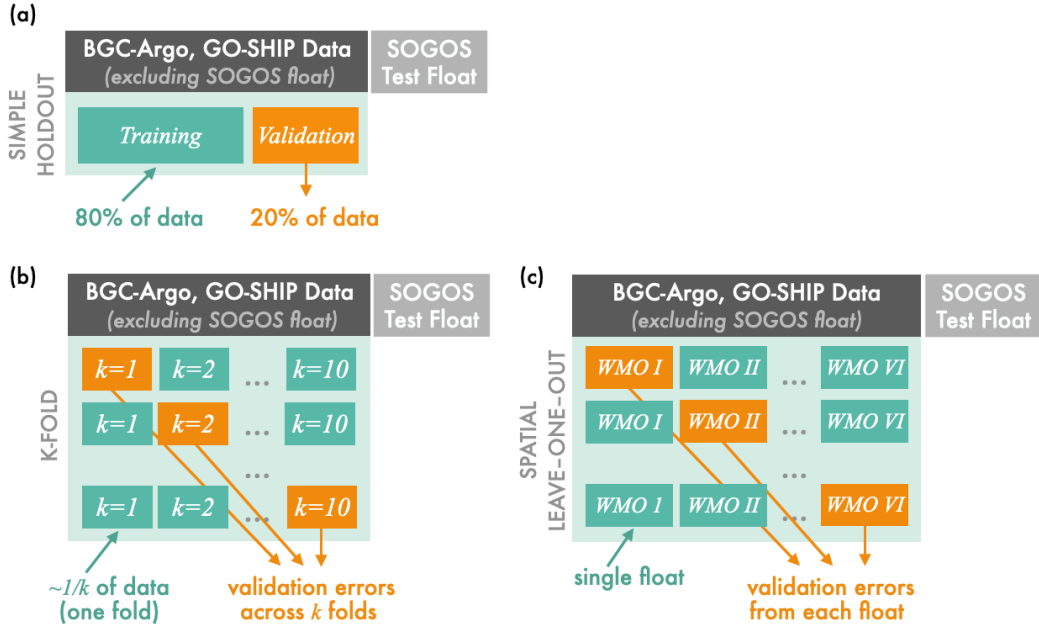


FIG. 4: Schematics of (a) simple holdout validation with an 80% and 20% split, (b) k-fold cross-validation, and (c) spatial leave-one-out (SLOO) cross-validation.

is to get a final independent estimate of model error when applied onto the Seaglider tracks. Since the Seagliders sample close to the SOGOS float by experimental design, the entirety of the SOGOS float data (WMO 5906030) is reserved for testing. The training dataset consists of 11645 observations, and the validation dataset has 2933 observations; the test dataset from the withheld SOGOS float has 3308 observations from 16 profiles. We emphasize that RFR applications in the geosciences should consider the specific goal of RFR when determining what data constitutes the training, validation, and test datasets.

After selecting the optimal feature list during validation and testing, we train a final RFR model using the chosen feature list. At this stage, we combine the training, validation, and test datasets (representing all available GO-SHIP and BGC-Argo data) to serve as an expanded training dataset for the the best model performance. This RFR is applied to the Seaglider observations to estimate high-resolution nitrate fields.

b. Alternative Machine Learning Models

We compare RFR performance to that of two currently existing machine learning algorithms that have been applied for nutrient estimation. The CARbonate system and Nutrients concentration from hYdrological properties and Oxygen using a Neural-network version B (CANYON-B) uses Bayesian neural networks to predict nutrient variables (nitrate, phosphate, silicate) from a set of predictive features including T, S, p, O₂, latitude, longitude, and/or day of the year (Bittig et al. 2018). The Empirical Seawater Property Estimation Routines (ESPER's) are another set of predictive biogeochemical algorithms (Carter et al. 2021), and the ESPER-Mixed routine combines estimates from neural networks and locally interpolated regressions. CANYON-B and ESPER-Mixed are trained primarily on cruise data from different releases of the Global Data Analysis Project (Olsen et al. 2019). To produce nitrate estimates from ESPER-Mixed, we use input features of T, S, p, O₂, latitude, and longitude; for CANYON-B, the same features plus time were considered.

c. Mixed Layer Characterization at High Frequencies

We analyze mixed layer properties of both physical and biogeochemical parameters over the course of the SOGOS deployment, which we divide into regions of high and low EKE (Section 2). Along-track glider EKE is calculated by interpolating the satellite surface EKE to the average

location of each glider profile. Mixed layer depth (MLD) was calculated following Dove et al. (2021) as the depth at which density is first 0.05 kg m^{-3} greater than the density observed at 10 m. The absolute horizontal buoyancy gradient in the mixed layer ($|\nabla_h b|$) is estimated for each i th profile following $|\nabla_h b|_i = (|b_{i+1} - b_{i-1}|)(\Delta d)^{-1}$, where b_i is the mean buoyancy in the mixed layer for the i th profile and Δd is the horizontal distance between profiles at indices $i - 1$ and $i + 1$. Overall, $\nabla_h b$ is an underestimate of the strength of the fronts because the gliders do not always sample these features perpendicularly (Thompson et al. 2016). However, we expect a few outlying, large $\nabla_h b$ values, which can occur when the depth-averaged current causes the glider separation Δd to be small due to the surfacing position being close to the dive position. We choose not to account for the impact of the depth-averaged current in this analysis.

The mixed layer mean nitrate (\bar{N}_{ML} , in $\mu\text{mol kg}^{-1}$) represents the integrated nitrate content in the mixed layer, normalized by the thickness of the mixed layer (h_{ML} , in m):

$$\bar{N}_{ML} = \frac{1}{h_{ML}} \int_{-h_{ML}}^0 [NO_3^-] \rho \, dz, \quad (2)$$

where $[NO_3^-](z)$ is the RFR-nitrate prediction in $\mu\text{mol kg}^{-1}$, and $\rho(z)$ is in-situ density in kg m^{-3} . In practice, the integral is estimated using trapezoidal sums on the irregularly vertically spaced observations. We also analyze the difference in nitrate concentration across the base of the mixed layer (ΔN_{ML}), which is defined as \bar{N}_{ML} minus the mean nitrate concentration 20 m below the MLD (calculated over the depth range 10–30 m below the MLD). Horizontal variance of the mixed layer nitrate (s_{H,NO_3}^2 , in $\mu\text{mol}^2 \text{kg}^{-2}$) is calculated as the variance in \bar{N}_{ML} when binned into 10-profile (~ 2 day) windows that overlap by 2 profiles (~ 10 hours). To relate the nitrate time series to evidence of biological production, we also track the mixed layer mean optical backscatter at 470 nm ($\overline{bbp_{470}}$), which is calculated following the same method in Equation 2). Backscatter measures the scattering of light by particles present in the water, and serves as a proxy for particulate organic carbon.

d. Wavelet Analysis for Timescales of Variability

We apply wavelet analysis to the RFR glider estimates to detect important timescales of nitrate variability in the mixed layer. A wavelet represents a local function in the frequency-time space, centered around 0, with a particular frequency distribution. Using wavelet analysis to return time-

localized information (over spectral analysis) is particularly useful when analyzing climate data because of non-stationarity and persistence in the time series (Grinsted et al. 2004; Torrence and Compo 1998). Here, we apply a version of a continuous wavelet transform (CWT) called the weighted-wavelet Z transform (WWZ), which decomposes the time series into dilated and shifted transformations of a complex Morlet wavelet (a sinusoidal wave with Gaussian decay). In essence, WWZ returns information on how closely a signal resembles the wavelet defined at a given time and frequency (Foster 1996). WWZ is suitable for irregularly spaced data since it does not rely on interpolation, which can misrepresent the true spectral content of the time series (Torrence and Compo 1998). All signal processing is performed using the Python package Pyleoclim designed for use on climate time series (Khider et al. 2022). Significance testing is done by comparison to the CWTs of 1000 iterations of the theoretical red noise, first order autoregressive process AR(1) benchmark (Torrence and Compo 1998).

4. Results

a. Random Forest Regression

1) RFR TRAINING

During RFR training, we explore the performance of different feature lists by evaluating the feature importance and validation errors for each model version; results from seven example models are presented here. Model A represents the simplest RFR feature list, using only two variables, τ and σ_0 , while models D–G incorporate additional spatiotemporal information (Figure 5a). RFR feature importance is one measure of the predictive capability of a given predictor variable (Methods Section 3a), but the limitations of using feature importances alone for final feature selection has been noted previously (e.g. Strobl et al. (2007)). Whereas SA, p, and O₂ consistently appear to have strong predictive capability in our RFR, other spatial and temporal features (like latitude, longitude, and time) return low feature importance (Figure 5b). However, when we evaluate the models on validation data, these spatiotemporal features improve the performance of the RFR as indicated by reductions in the error bias, median absolute error (MAE), and interquartile range of the absolute errors (IQR-AE) (Table 1; Figure 6). To give an example of how feature importance can misrepresent predictive power, we consider a feature (e.g. latitude) that may be primarily useful toward the top of a decision tree for an initial splitting of the data, but less useful than other

features further down in the tree. Given how feature importances are calculated, latitude would then return low feature importance despite providing a critical “pre-sorting” of the data that allows the high feature importance predictors to be effective in the rest of the tree. We emphasize that without also evaluating validation error distributions of the different models, using only feature importance could lead to sub-optimal feature selection.

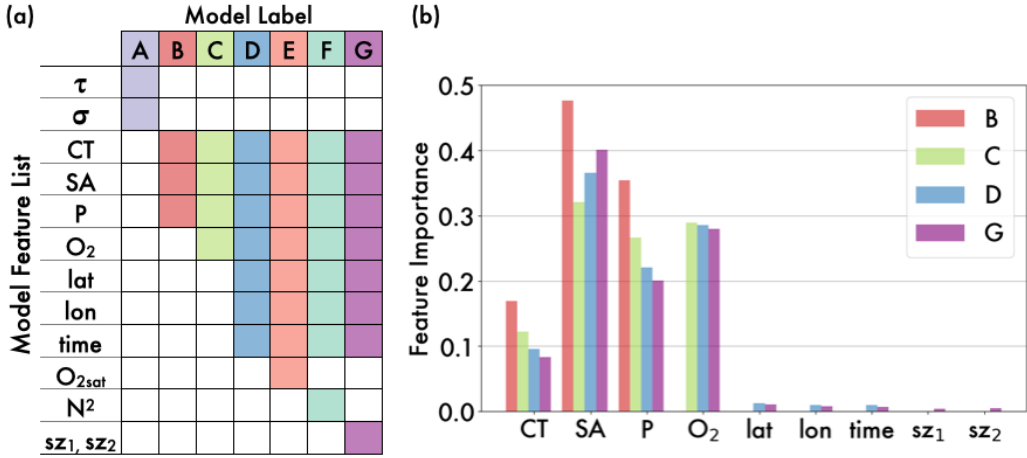


FIG. 5: (a) Feature lists for models A through F, showing which variables are considered during RFR training for each version of the model. (b) Feature importances [unitless], as described in Section 3a, for models B, C, D and G. The sum of all feature importances for a given model is one.

2) RFR VALIDATION

We next use simple holdout validation, which computes prediction errors on 20% of withheld GO-SHIP and BGC-Argo data, to assess suitability of the feature lists (Figure 4a). This holdout validation is a common first-order approach to assessing performance from different feature lists. Replacing τ and σ_0 in model A with CT and SA, and adding biogeochemical information (O_2) with spatiotemporal coordinates (latitude, longitude, yearday, and season) in model D all contribute to decreasing the simple holdout validation MAE (\pm IQR-AE) from $0.309 \mu\text{mol kg}^{-1}$ (± 0.461) to $0.142 \mu\text{mol kg}^{-1}$ (± 0.205) (Table 1). If a particular feature can be calculated from the other variables in the feature list, its inclusion does not seem to improve performance. For example, adding O_{2sat} to a feature list that already has O_2 as a predictor will yield nearly the same results. Likewise, the addition of N^2 does not improve performance noticeably in model F over model D. This characteristic is likely because CT, SA, and p already encode much of the same information,

and/or because the variable is too noisy to be an effective predictor. From the simple holdout validation alone, the distinctions in performance between model D and G are unclear since the models have very similar MAE (\pm IQR-AE) values ($0.146 (\pm 0.211) \mu\text{mol kg}^{-1}$ for model D and $0.142 (\pm 0.205) \mu\text{mol kg}^{-1}$ for model G; Table 1).

Model	Validation MAE	Validation IQR-AE	Validation Median Bias
A	0.3089	0.4605	-0.0554
B	0.2669	0.3686	-0.0241
C	0.2431	0.3533	-0.0141
D	0.1455	0.2106	-0.0154
E	0.1459	0.2140	-0.0172
F	0.1525	0.2245	-0.0149
G	0.1422	0.2047	-0.0171

TABLE 1: Simple holdout (20%) validation errors for models A through G, whose feature lists are given in Figure 5. Median absolute error (MAE), interquartile range of the absolute errors (IQR-AE), and median bias are reported in $\mu\text{mol kg}^{-1}$.

3) RFR CROSS-VALIDATION

To better distinguish the feature list with the best predictive performance and lowest overfitting tendency, we use both k-fold cross-validation and spatial leave-one-out (SLOO) cross-validation. During k-fold cross-validation, we generate a larger set of validation errors to differentiate performance between models. For $k=10$ in our case, ten RFR models are trained, each time training on observations from nine folds and withholding one fold for validation (Figure 4b). The distribution of 10 k-fold MAEs for each model reflects similar results to those from simple holdout validation above, where models D–G significantly improve performance over models A–C (Table 2; Figure 6a). For models D–G, the validation errors across folds have a small spread despite being trained on shuffled data; the consistent performance suggests that the model is robust and generalizes well to new data.

We next combine the k-fold cross-validation errors across folds to estimate a probability density (Gaussian kernel density estimate; KDE) for each of the models (Figure 7a). All models have slightly negative bias (center of curves in Figure 7a) but the bias is significantly improved for models D–G ($\sim 0.02 \mu\text{mol kg}^{-1}$), seemingly due to the inclusion of latitude, longitude, and yearday

Model	K-Fold MAE	K-Fold IQR-AE	K-Fold Median Bias
A	0.2857	0.4419	-0.0531
B	0.2683	0.3837	-0.0351
C	0.2463	0.3474	-0.0389
D	0.1419	0.2118	-0.0190
E	0.1456	0.2144	-0.0173
F	0.1477	0.2207	-0.0201
G	0.1380	0.2023	-0.0152

TABLE 2: K-fold cross-validation errors for models A through G (combined across all folds), whose feature lists are given in Figure 5. Median absolute error (MAE), interquartile range of the absolute errors (IQR-AE), and median bias are reported in $\mu\text{mol kg}^{-1}$.

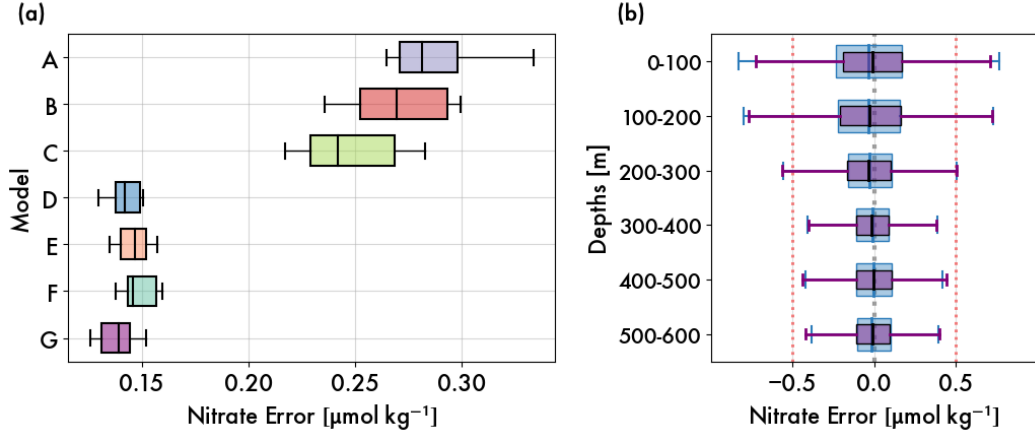


FIG. 6: (a) Spread of aggregated k-fold cross-validation MAEs ($\mu\text{mol kg}^{-1}$) across models A–G. (b) Cross-validation errors in 100-m depth bins for models D (blue) and G (purple). Float nitrate measurement uncertainty of $\pm 0.5 \mu\text{mol kg}^{-1}$ in dashed red lines (Maurer et al. 2021).

features (Table 2). The peak of the KDEs for models D–G suggest that the inclusion of seasonal variables s_{z1} and s_{z2} as predictors in model G leads to the best performance out of the four models. From aggregating validation errors into 100 m bins to examine the depth-dependence of performance, we find that the improvements in model G over model D are mostly in the upper 200 m near the surface (Figure 6b).

We conclude validation by using a spatial leave-one-out (SLOO) cross-validation technique on models D and G, which is a useful technique in geoscience contexts for assessing the impact of spatiotemporal correlations in the observations used for training (Stock 2022; Stock and Subramaniam

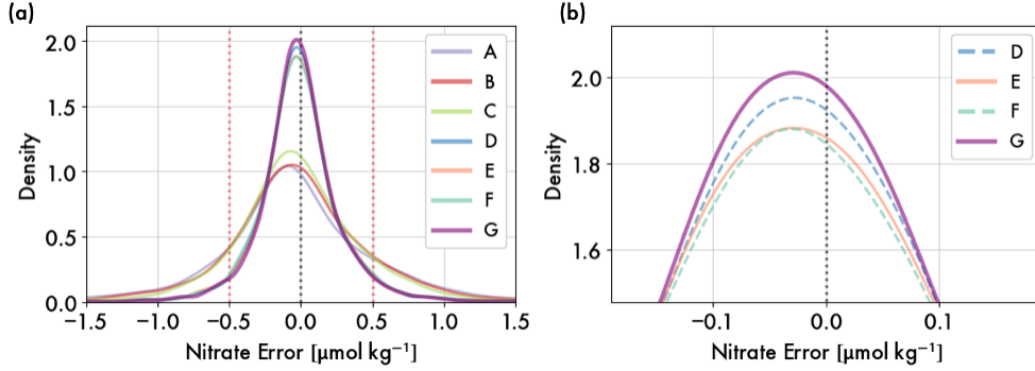


FIG. 7: (a) Gaussian kernel density estimate (KDE) for k-fold cross-validation errors from models A–G. The probability density is most closely centered around 0 for model G. Float nitrate measurement uncertainty of $\pm 0.5 \mu\text{mol kg}^{-1}$ in dashed red lines (Maurer et al. 2021). (b) Zoomed in view highlighting the peaks of the KDE’s for models D–G.

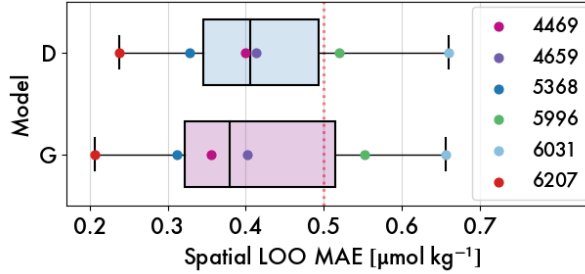


FIG. 8: Spatial leave-one-out validation errors on six floats. Colored dots denote the validation MAE’s for each withheld float WMO; “590” truncated from all float labels in legend. Float nitrate measurement uncertainty of $\pm 0.5 \mu\text{mol kg}^{-1}$ in dashed red line (Maurer et al. 2021)

2022). BGC-Argo training floats were selected due to their temporal and spatial proximity to the SOGOS float and gliders, but these training observations are taken at different points in time and in slightly different regions (Figure 3). Here, we iteratively leave out one float to serve as a validation dataset, training on the remaining five floats (Figure 4c). The six SLOO models return an average validation MAE (\pm IQR-AE) of $0.406 \mu\text{mol kg}^{-1}$ (± 0.148) for model D, and $0.379 \mu\text{mol kg}^{-1}$ (± 0.192) for model G. The SLOO validation errors on a single float tend to be higher than those from simple holdout validation, but in general model G returns lower validation errors than model D for five of the six floats. Only the validation errors on float WMO 59055996 increase slightly from model D to model G. We use training observations from this particular float that are taken farther west than the SOGOS deployment (10°E versus 30°E) since the overall tracer distributions appear to match those of the SOGOS float. However, float WMO 59055996 is near the Polar Front

371 and spatially separated from many of the other training floats in the east, such that the training
372 data may be less representative of the float WMO 5905996 observations withheld from validation.
373 During early model development, floats that returned especially poor SLOO cross-validation errors
374 were removed from the training data before the final set of BGC-Argo floats was chosen. From
375 both k-fold and SLOO cross-validation results, we select feature list G for the next steps of RFR
376 development.

377 4) RFR TESTING

378 We test our RFR (model G) using the withheld SOGOS float data (Figure 9a), and find a test
379 validation MAE (\pm IQR-AE) of $0.203 (\pm 0.290) \mu\text{mol kg}^{-1}$, corresponding to a 0.22% relative
380 error and mean bias of $+0.082 \mu\text{mol kg}^{-1}$. When the float test MAE is computed only over the
381 period of Seaglider deployment (yeardays 120–200), the MAE is $0.345 \mu\text{mol kg}^{-1} (\pm 0.428)$,
382 which is slightly higher than that calculated over the larger SOGOS float dataset extending into
383 2020 (Figure 9c). This difference may be due to the fact that the SOGOS deployment begins in one
384 of the most energetic regions of the global oceans. The pattern of overestimation reaches yearday
385 ~ 250 (end of August), which is around the end of austral winter. Even so, 84% of all test MAEs
386 are $\leq 0.5 \mu\text{mol kg}^{-1}$ (float nitrate measurement uncertainty for BGC-Argo; Maurer et al. (2021))
387 and 95% of test MAEs are $\leq 0.732 \mu\text{mol kg}^{-1}$.

388 The RFR test errors exhibit significant variability both in time (horizontal axis) and along depth
389 (vertical axis in Figure 9c). Under the mixed layer, the RFR test errors show small but persistent
390 overestimation. These positive errors in the interior may result from the fact that the larger
391 BGC-Argo training dataset covers a slightly wider range of nitrate values than the SOGOS float
392 observations alone, particularly at lower concentrations. The narrower range of nitrate observed
393 by the SOGOS float is not unexpected, given that the SOGOS float samples during austral fall and
394 winter, while the broader BGC-Argo training dataset covers the annual cycle. All of the BGC-Argo
395 and GO-SHIP data used for training were from the same region in the ASZ that the SOGOS
396 float samples, so we expect that the same water masses were observed in the training/validation
397 and test data (Figure 3). Above the MLD, there are different patterns of alternating over- and
398 underestimation that may be related to nitrate anomalies associated with mesoscale structures.

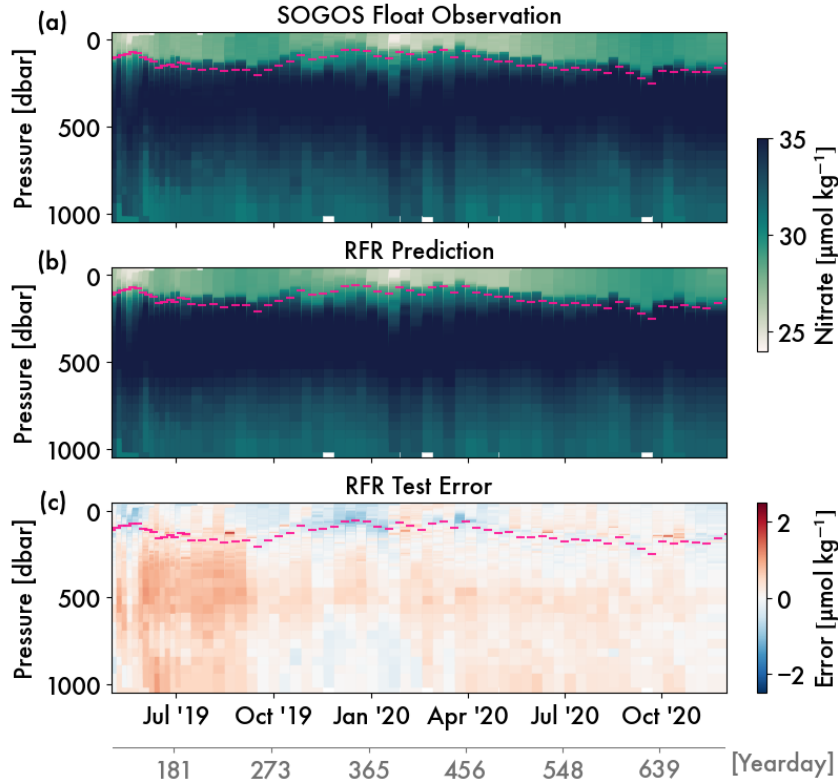


FIG. 9: Nitrate ($\mu\text{mol kg}^{-1}$) (a) observed by the SOGOS float and (b) predicted by RFR model G for the upper 1000 dbar. (c) Nitrate prediction error (predicted – observed; $\mu\text{mol kg}^{-1}$) over the same pressure range. MLD in magenta lines. Horizontal time axis in month 'year format; corresponding yearday (relative to Jan 2019) in grey.

For a final evaluation of our RFR model performance, we compare the output to predictions made from two other robust machine learning algorithms, CANYON-B and ESPER-Mixed (Methods Section 3b; Bittig et al. (2018); Carter et al. (2021)). The two algorithms perform similarly with excellent performance under 200 m (Figure 10b-c). Strong performance of both models in the ocean interior underscores the utility of machine learning methods for tracer estimation. However, within the mixed layer, both CANYON-B and ESPER-Mixed outputs persistently and substantially underestimate nitrate (Figure 10b-d). The underestimation may be due to temporal biases where the cruise observations used to train CANYON-B and ESPER-Mixed are heavily biased towards austral summer (Bittig et al. 2018). Mixed layer nitrate is expected to be lower during these months when biological utilization is most active. In contrast, our observations are later in austral fall when the mixed layer deepens and productivity decreases to low background levels (Su et al.

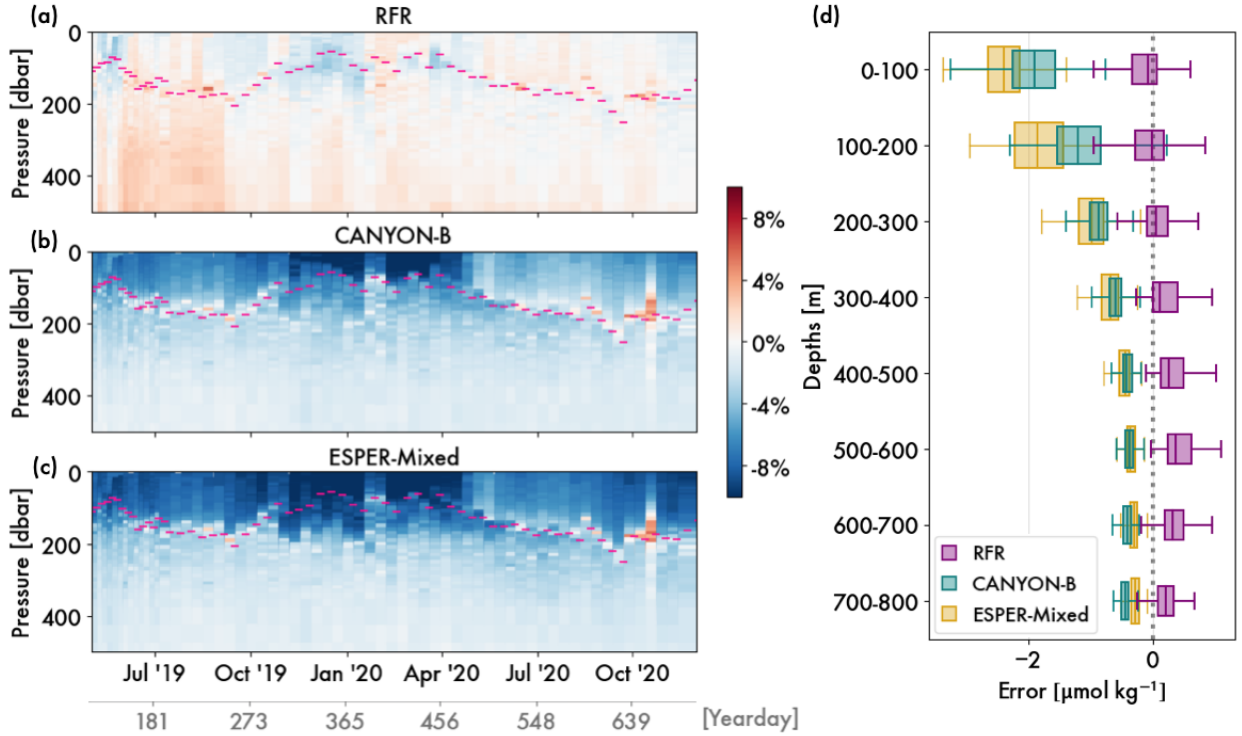


FIG. 10: Relative prediction errors $[(\text{predicted} - \text{observed}) / \text{observed}]$ for SOGOS test float data in the upper 500 dbar using (a) RFR, (b) CANYON-B, and (c) ESPER-Mixed. Results not shown between 500 dbar and 1000 dbar are similar to those under ~ 300 dbar. MLD in magenta lines. Horizontal time axis in month 'year format; corresponding yearday (relative to Jan 2019) in grey. (d) Test errors ($\mu\text{mol kg}^{-1}$) in 100-m depth bins for RFR (purple), CANYON-B (teal), ESPER-Mixed (tan).

(2022)). CANYON-B, which considers day of year, returns better predictions in the upper 200 m than ESPER-B, which does not use time as a predictor. This increased performance in the upper ocean of CANYON-B relative to ESPER-B mirrors how the addition of seasonal variables to our RFR model G improved performance over model D (Figure 6). Descriptions of the algorithms in Carter et al. (2021) and Bittig et al. (2018) mention that the seasonality and exact values of their models' output should be taken with caution in the upper ocean. The relative success of our RFR method in estimating mixed layer distributions suggest that targeted regional models may be able to recover useful information about the upper ocean that is lost in globally trained models.

Using feature list G, we train a final RFR model using all the available GO-SHIP and BGC-Argo data, including those from the SOGOS float. This approach utilizes all the observations possible to yield the best predictive power, but only after cross-validation is complete. Once the final RFR has been trained, we supply high-frequency Seaglider observations as inputs to the model and

422 generate novel nitrate distributions at high horizontal resolution along the glider tracks (Figures
 423 11b, 12). The gliders have average horizontal distance of ~ 1.5 km between profiles (downcast
 424 and upcast are separate profiles), whereas the BGC-Argo float profiles are separated by ~ 70 – 80
 425 km. Application of RFR therefore results in a ~ 50 -fold increase in horizontal resolution for the
 426 nitrate distributions in this region (Figure 11b). Though the float profiles are sparse relative to
 427 those from the gliders, the SOGOS float's sampling frequency of 5 days is already faster than the
 428 typical profiling period of 10 days for most Argo floats. Scientific questions that can be addressed
 429 using the global BGC-Argo array may be limited by resolution, but an RFR approach can extend
 430 rich regional datasets from floats by leveraging the spatiotemporal resolution offered by different
 431 platforms.

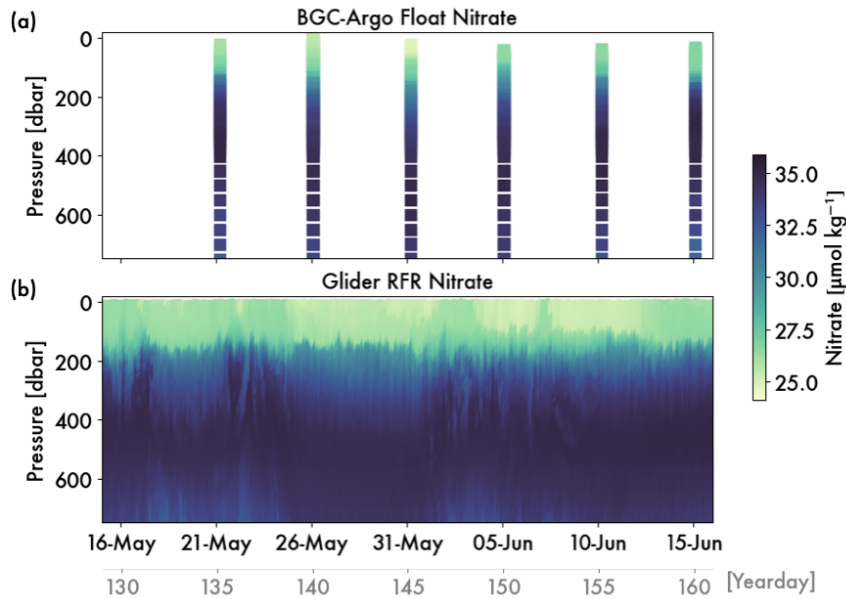


FIG. 11: Nitrate ($\mu\text{mol kg}^{-1}$) (a) observed by the SOGOS float over a period of ~ 30 days (6 profiles)
 and (b) predicted by RFR model G for glider SG660 over the same time period (428 profiles).
 Horizontal time axis in day-month format for 2019; corresponding yearday (relative to Jan 2019)
 in grey.

432 *b. Mixed Layer Variability from High-Frequency RFR Estimates*

433 The RFR-derived nitrate distributions enable analysis of mixed layer variability at much higher
 434 temporal resolution than previously possible in this region. High-frequency variability in the
 435 nitrate distributions is most evident below the base of the mixed layer (Figure 12). We note that CT

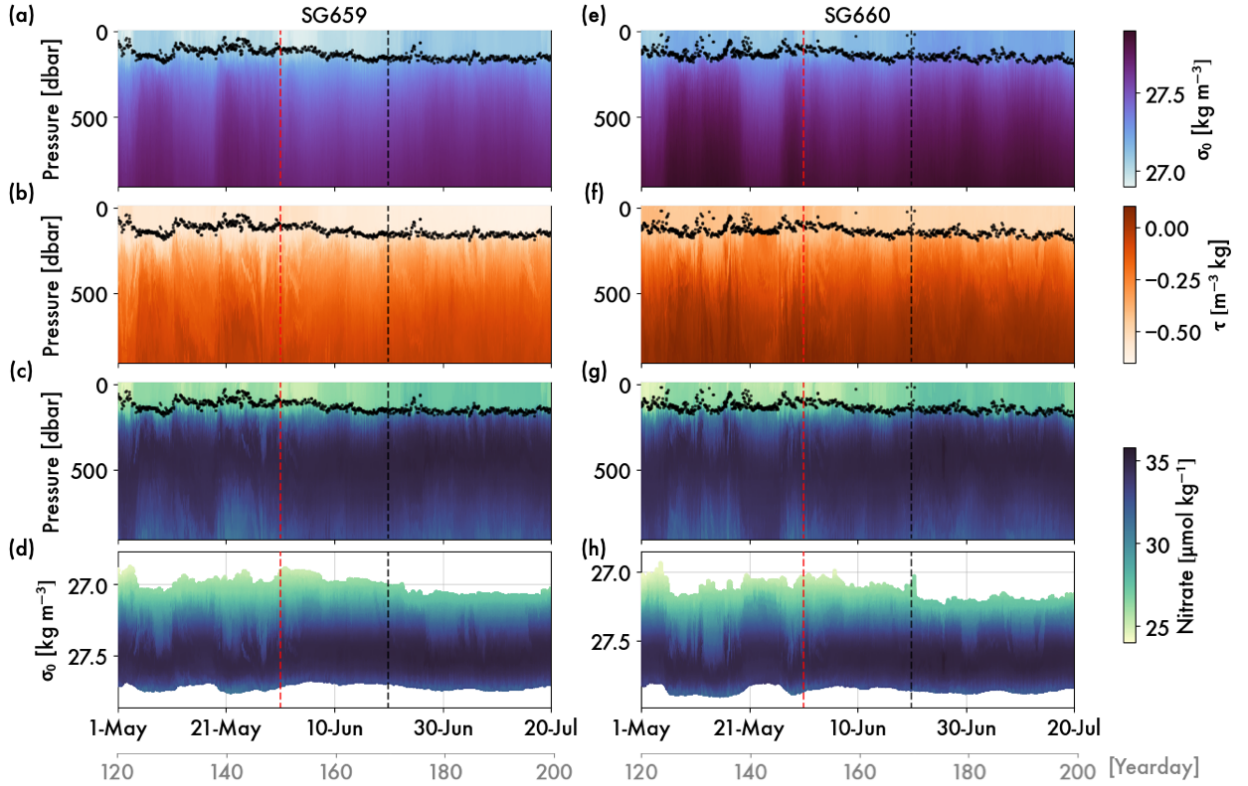


FIG. 12: (a) Potential density referenced to surface (σ_0), (b) spice (τ), and (c) nitrate ($\mu\text{mol kg}^{-1}$) sections from SG659 plotted in time-pressure space; (d) nitrate plotted in time-density space. (d-f) Same variables as left, but for SG660. Dashed red lines bound the high EKE region (yeardays 120–150) and dashed black lines the low EKE region (yeardays 170–200). Horizontal time axis in month-day format for 2019; corresponding yearday (relative to Jan 01 2019) in grey. Sections showing the full nitrate prediction to 1000 m are given in Supplementary Information.

and SA are impacted by atmospheric surface forcing (including heating, cooling, evaporation, and precipitation) that typically occurs at larger scales, such that small-scale variability generated by stirring tends to be reduced in the mixed layer relative to the interior. Since RFR uses both CT and SA for nitrate prediction, these parameters may contribute to similar erasure of rapid variability in the mixed layer nitrate. Although the dampening of high-frequency variability in the mixed layer could be partially due to artifacts of the RFR model, biological drawdown of nitrate near the surface would also tend to decrease nitrate variability at short timescales when rapid injection is met with utilization.

We focus our next analysis on processes affecting the mixed layer, using various time series to quantify the strength and timing of nutrient injection into the upper ocean. Nutrient supply into the mixed layer can be mediated both by uplifting of isopycnals and changing of MLD, or direct tracer

transport along isopycnals across the base of the mixed layer (Freilich and Mahadevan 2019). To interpret relationships between the nutrient and physical dynamics of this region, we divide the SOGOS deployment into two regions of different EKE (Section 2). The three SOGOS platforms observe a high EKE region by the Southwest Indian Ridge (yeardays 120-150) before passing into a low EKE region downstream (yeardays 170–200; Figures 1a, 13a). Previous physical characterization of this region in Dove et al. (2021) suggests that the high EKE region is rich in submesoscale instabilities that affect vertical stratification at the base of the mixed layer and promote greater biogeochemical exchange between the mixed layer and ocean interior.

We use changes in the RFR-derived mean mixed layer nitrate \bar{N}_{ML} as a proxy for deep, nutrient-rich waters reaching the upper ocean. Any motions that stir nitrate-rich filaments into the mixed layer would also deliver other remineralized nutrients. Since large areas of the Southern Ocean are limited by iron, local decreases in \bar{N}_{ML} may be attributed to additional biological utilization spurred by iron availability. Values for \bar{N}_{ML} tend to be lower in the high EKE region (yeardays 120–150) than in the low EKE region (yeardays 170–200; Figure 13d), which we associate with higher levels of productivity using satellite light availability (PAR) and optical backscatter (bbp_{470} ; Figure 13g,h). The decrease in \bar{N}_{ML} coincides with an increase in optical backscatter, indicating more particulate organic carbon in the upper ocean. At this point in austral fall, summer blooms have already utilized available nutrients in the mixed layer and light availability is decreasing, so background biological activity is relatively low (Su et al. 2022). Still, additional nutrient input into the mixed layer can spur productivity, although full bloom initiation could take weeks or occur only weakly. There are occasionally significant differences in the \bar{N}_{ML} observed by the different platforms (e.g. yeardays 150–165; Figure 13d), which can be partially attributed to periods of increased spatial separation between the gliders and SOGOS float (Figure 1b).

For other characteristics like MLD and the difference in nitrate across the base of the mixed layer (ΔN_{ML}), the time series from the three SOGOS platforms reflect that the SOGOS float misses the rapid variability captured by gliders (Figure 13b,e). The ΔN_{ML} calculated from the SOGOS float observations are consistently small within the high EKE region (yeardays 120–150). In contrast, the ΔN_{ML} estimated from RFR glider nitrate changes rapidly and reaches values twice as large as those from the float (Figure 13e). Though long-term trends from both platforms may be similar, the long profiling period of the BGC-Argo obscures significant patterns evident in the glider nutrient

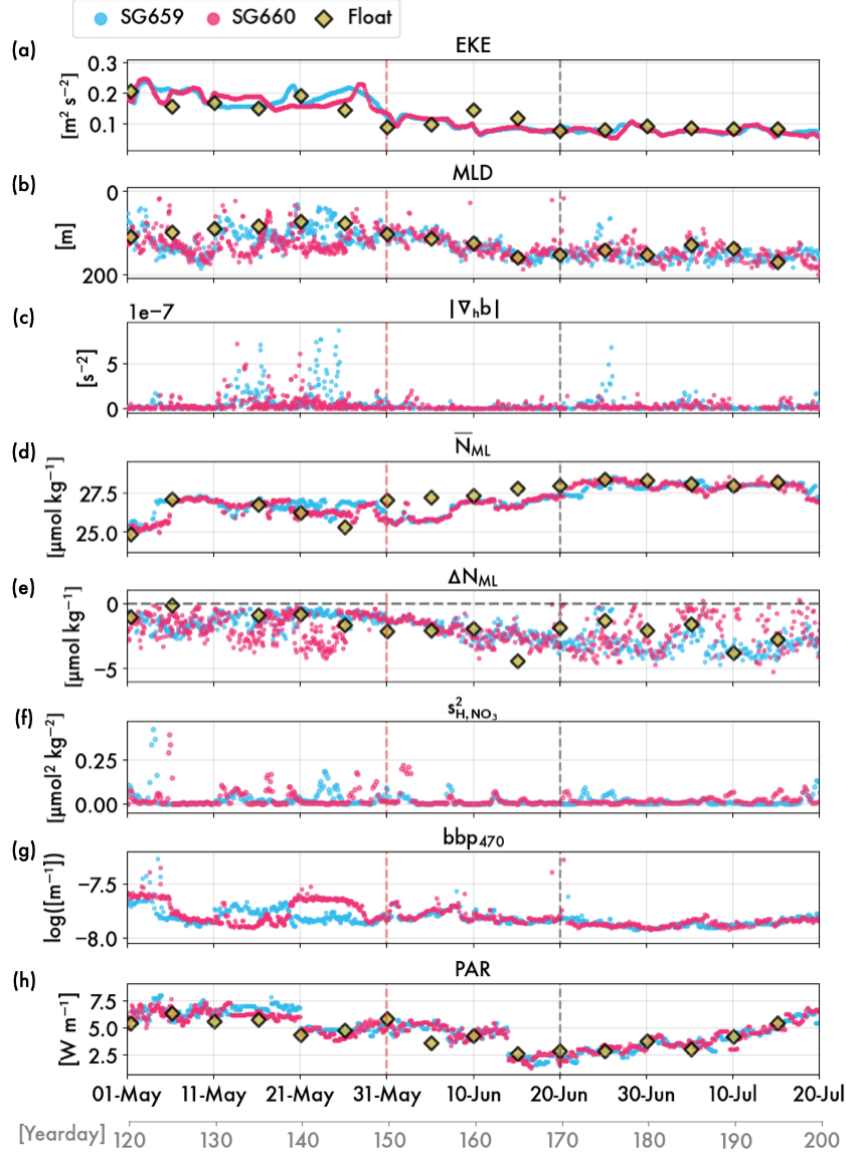


FIG. 13: For SG659 (teal), SG660 (magenta) and SOGOS float (yellow diamonds, when available), (a) along-track eddy kinetic energy (EKE); (b) mixed layer depth (MLD); (c) horizontal buoyancy gradient ($\nabla_h b$) in the mixed layer; (d) mixed layer mean nitrate concentration (\bar{N}_{ML}); (e) difference in nitrate concentration across the base of the mixed layer (ΔN_{ML}); (f) horizontal variance in mean mixed layer nitrate (s^2_{H,NO_3}) for 10-profile window (~ 2 days); (g) logarithm of mixed layer mean backscatter at 470 nm ($bb p_{470}$), (h) photosynthetically active radiation (PAR). Horizontal time axis in day-month format for 2019; corresponding yearday (relative to Jan 01 2019) in grey. Dashed red lines bound the high EKE region (yeardays 120–150) and dashed black lines bound the low EKE region (yeardays 170–200). Panels d–f use the RFR-predicted nitrate fields while the rest are observed quantities.

477 signals. We comment further on the platforms' distinct sampling strategies and the respective
478 benefits in Section 5.

479 Strong fluctuations in the mixed layer nitrate signals coincide with signs of increased physical
480 stirring at small scales. When the gliders sample the high EKE region (yeardays 120–150), they
481 observe intermittent shoaling of the mixed layer (Figure 13b) as well as enhanced horizontal
482 buoyancy gradients that indicate submesoscale structures (Figure 13c; previously explored in Dove
483 et al. (2021)). Analogously, we use the RFR glider nitrate estimates to demonstrate that the
484 high EKE region frequently exhibits higher horizontal variance in mixed layer nitrate (s_{H,NO_3}^2)
485 as compared to the low EKE region (Figure 13e). Higher values of s_{H,NO_3}^2 in the high EKE
486 region suggest that the RFR glider estimates are resolving filaments of water masses with distinct
487 nutrient characteristics, sourced from different regions, as they are stirred at mesoscales and
488 submesoscales. The timing of elevated s_{H,NO_3}^2 coincides with the sharp gradients in MLD and
489 intensified lateral buoyancy gradients observed by the gliders (e.g. yeardays 125, 131, 137, 148
490 for SG660). Altogether, characteristics of the RFR-derived nitrate over time appear consistent
491 with enhanced submesoscale upwelling of nitrate through the steepening of density gradients and
492 weakening of stratification at the base of the mixed layer.

493 *c. Timescales of Nutrient Variability*

494 We next characterize the important temporal scales at which upper ocean nutrient content varies
495 by applying wavelet analysis; the high-frequency RFR nitrate estimates allow us to assess variability
496 at a more comprehensive range of periods (from ~5 hours to ~50 days). Along-isopycnal nitrate
497 and spice are tracked on a range of density surfaces that are generally within the nutricline, with
498 average depths of the corresponding isopycnals (\bar{d}) ranging from ~170–400 m for SG660 (Table
499 3; Figure 12h). Applying continuous wavelet transform (CWT) on the along-isopycnal nitrate and
500 spice signals shows which frequencies of variability are dominant in each signal, and at what point
501 in the deployment; higher CWT amplitudes indicate enhanced variability at that given frequency
502 and time.

503 The along-isopycnal nitrate CWTs reflect significant frequencies of nitrate variability that are both
504 in the mesoscale and submeoscale range. We caution that significance should only be interpreted
505 outside of the shaded “cone of influence” (COI). This excluded region is a result of the finite nature

σ_0	\bar{d}
27.30	174.63
27.40	223.68
27.50	296.80
27.60	396.92

TABLE 3: Average depth (\bar{d}) of analyzing isopycnal (σ_0) for SG660

of the time series; wavelets defined at periods of X days can only be considered significant X days after the start or from the end of deployment. At mesoscale periods ≥ 20 days, there appear to be significant bands of high CWT amplitudes in both nitrate and spice during yeardays 140–150 when the glider samples the eddy-rich, high EKE region (Figure 14a,b). These patterns of enhanced high-frequency variability extend down to the base of the nutricline at ~ 400 m (Figure 14).

At submesoscale periods (~ 0.2 days to ~ 2 days), the CWT plots show several short-lived events (e.g. yeardays 130, 137, 150) during which both nitrate and spice variability are strongly enhanced in the interior between 200 m and 400 m depths. Increased CWT amplitudes at submesoscales continue to occur sporadically, although weakly, in the low EKE region after yearday 170 along the shallow isopycnals $\bar{d} \leq 223$ m (Figure 14). High CWT amplitudes at submesoscales tend to occur where there are sharp gradients in MLD and shoaling of the analyzing isopycnal (yeardays 130, 137), presumably due to eddy activity (Figure 15a–c). During these short events, the glider appears to sample across filaments of enhanced FSLE (Figure 15d–f), which are typically found around submesoscale fronts or between mesoscale eddies (Siegelman et al. 2020). Not all times at which the glider appears to sample strong MLD gradients are associated with enhanced submesoscale variability in the nitrate CWTs (e.g. yeardays 124, 147, Figure 15a–c).

Our wavelet analysis suggests that nitrate variability is often dominated by mesoscale modulation of water masses at periods between ~ 20 to 50 days. However, especially in the high EKE region, rapid changes in along-isopycnal nitrate appear to occur in intermittent, short-lived events at timescales associated with submesoscale stirring (~ 0.5 to 2 days). Although not within the scope of this paper, a different method using wavelet transform coherence could be used to measure the correlation between CWTs, i.e. the shared variability of two signals (Foster 1996). This type of analysis would explore how physical and biological time series co-vary, and what mixing

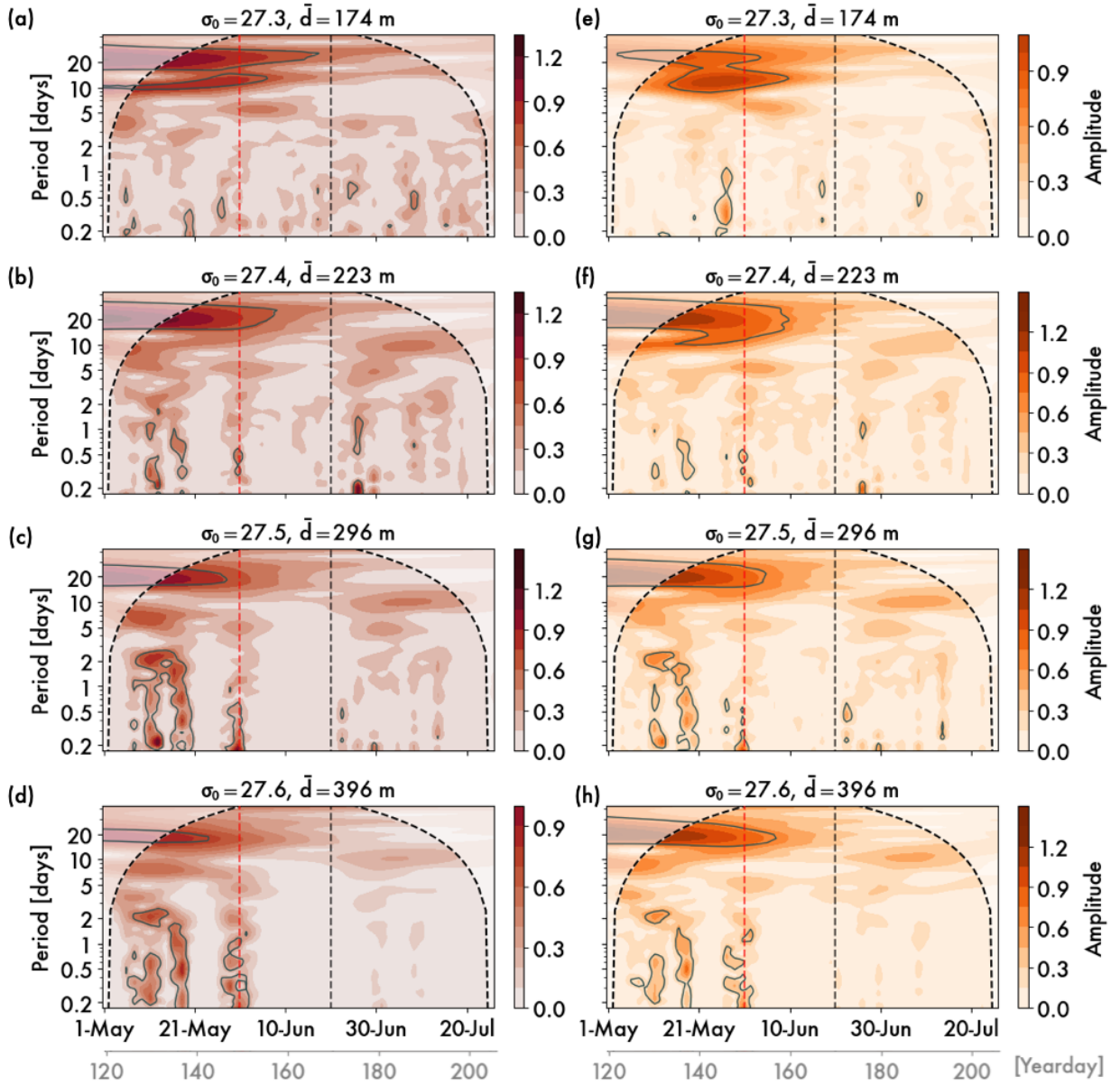


FIG. 14: (a-d) CWTs of RFR nitrate estimates from SG660 along four analyzing isopycnals (3). (e-h) CWTs of observed spice from SG660 along the same isopycnals. Dark grey lines are 95% significance contours. Dashed red lines bound the high EKE region (yeardays 120–150) and dashed black lines the low EKE region (yeardays 170–200). Inertial period at this latitude is ~ 0.643 days, or 15.4 hours.

mechanisms would lead to strong or weak covariance at different frequencies and times. Continued progress in nutrient mapping at higher resolutions, here achieved by RFR, will invite new methods of quantifying tracer variability at a comprehensive range of scales.

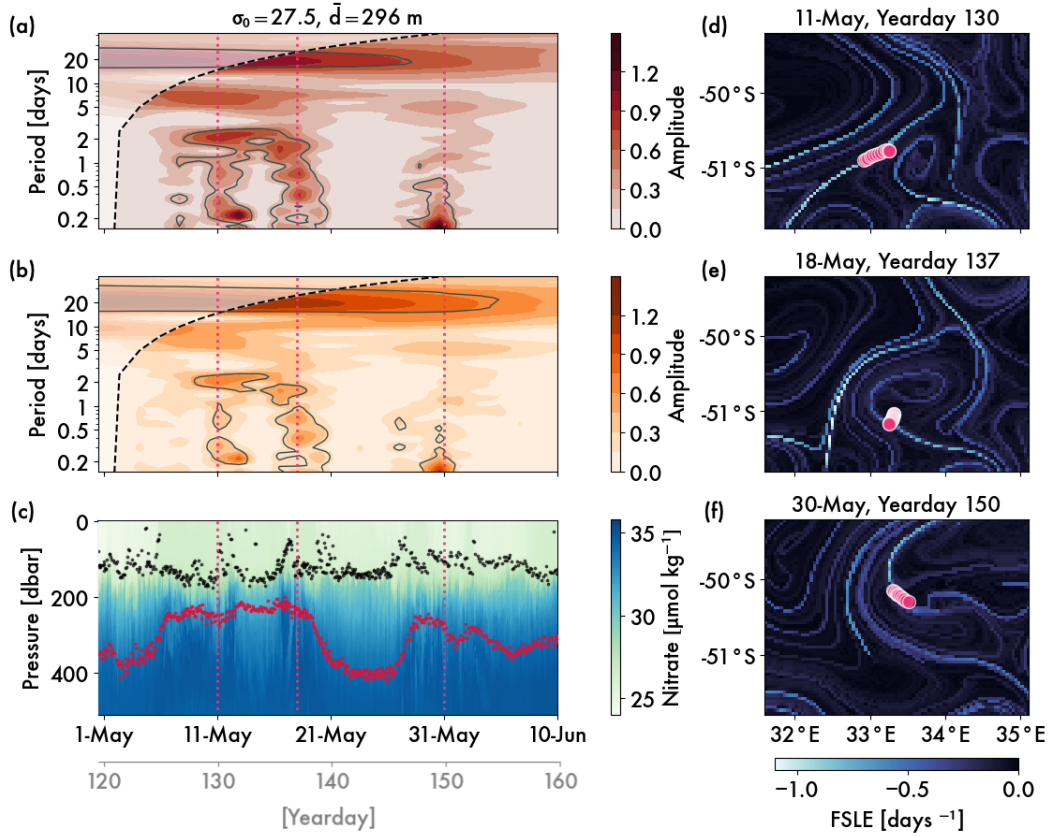


FIG. 15: (a, b) Nitrate and spice CWTs from SG660 for analyzing isopycnal $\sigma_0 = 27.5$, which is on average at $\bar{d}=296$ m. Dark grey lines are 95% significance contours. Cone of influence (COI) plotted in dashed black; results insignificant within white shaded region. (c) Glider SG660 nitrate section with analyzing isopycnal $\sigma_0=27.4$ in red dots; MLD in black dots. Dotted magenta lines in panels a–c indicate the year days represented in panels d–f. (d–f) Surface FSLE at daily resolution, with SG660 profile locations in magenta dots for a given year day.

5. Discussion and Conclusions

Our motivation for developing a regional random forest regression (RFR) for nutrient prediction was to bridge observational gaps at short timescales and extend insights into Southern Ocean nutrient variability. We train the RFR model on nutrient observations from regional BGC-Argo floats, then apply the RFR on inputs from rapid-sampling Seagliders to generate upscaled nitrate distributions. Using the observation-based RFR estimates, we find enhanced high-frequency variability in mixed layer nitrate in a turbulent region with enhanced submesoscale stirring. Quantifying the dominant timescales of variability over time with wavelet analysis suggests that nutrients are sporadically injected into the upper ocean in small-scale filamentary structures during short-lived

541 events; such rapid variability is only evident in the RFR glider estimates and not in the original
 542 BGC-Argo float observations.

543 The multi-platform SOGOS experimental design is well-suited for RFR because each of the
 544 observing platforms has distinct advantages. The SOGOS float measures an additional variable
 545 (nitrate), while the Seagliders measure at higher resolution (2 profiles every 4–6 hours, instead of 1
 546 profile every 5 days). By deriving high-frequency nitrate estimates along the Seaglider tracks with
 547 RFR, we can explore what nutrient information is missed by the SOGOS float alone. Even within
 548 the low EKE region (yeardays 170–200) when the gliders and float tend to sample close together,
 549 the float observes a much narrower range of nitrate values than the gliders (Figure 11, Figure 16a).
 550 The mean value of the float distribution is higher than that of the glider distribution, even when
 551 accounting for RFR prediction errors. The floats’ inability to capture small-scale dynamics may
 552 therefore bias long-term float averages used in other applications.

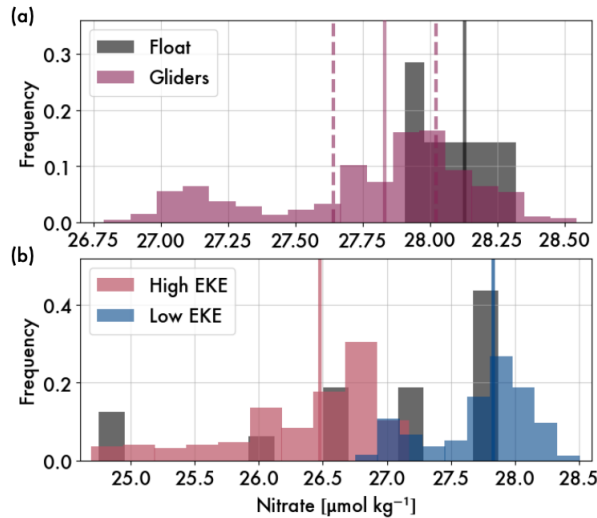


FIG. 16: (a) Distributions of mean layer mixed nitrate (\bar{N}_{ML} ;) from the RFR glider estimates (purple) and and the SOGOS float observations (grey) when the platforms sample close to another in the low EKE region (yeardays 170–200). Means of the distributions are shown by solid lines; dashed purple lines indicate the bounds of uncertainty using the RFR test MAE calculated for the low EKE region. (b) Distributions of \bar{N}_{ML} from the RFR glider estimates in the the high (red) versus low EKE regions (blue); means of the distributions in solid lines. Float observations (grey) from both the high and low EKE regions are grouped.

553 Likewise, differentiating the nutrient distributions between the high and low EKE regions would
 554 be impossible based on the sparse SOGOS float observations alone. On the other hand, the RFR
 555 glider nitrate estimates better resolve the two distinct distributions (Figure 16b), and return a

556 statistically significant difference between upper ocean nitrate in the high EKE versus low EKE
557 regions (Welch t-test statistic: -45.2, p-value ≈ 0 , dof=1357). The SOGOS float actually sampled
558 at double the resolution of the typical Argo float (~ 5 days rather than ~ 10 days), so the difference
559 in high-frequency coverage would be even greater in comparison to standard Argo floats. Given
560 in-situ sampling limitations, approaches like RFR can transfer the benefits of rapid sampling to
561 variables that are not represented in the gliders' sensor array. We encourage future observational
562 deployments to consider utilizing heterogeneous arrays of instruments, especially where machine
563 learning can be applied to fill in missing information (e.g. Salam and Hsieh (2023); Salcedo-Sanz
564 et al. (2020); Renosh et al. (2023); Lermusiaux et al. (2017); Chai et al. (2020)).

565 Among machine learning approaches, RFR is a relatively simple algorithm that can be trained on
566 regional datasets too small for deep learning. RFR has been successfully applied to a range of cases
567 in oceanography (e.g. Sharp et al. (2022a); Callens et al. (2020); Tong et al. (2019)), including for
568 oxygen prediction in the Southern Ocean using BGC-Argo float data (Giglio et al. (2018)). Further-
569 more, many oceanographic applications highlight RFR as a useful tool for geospatial observations
570 because of its reduced overfitting tendency and ability to handle non-linear relationships between
571 variables (Zhou et al. 2023; Sharp et al. 2022b). Deep learning methods are not necessarily better
572 than simpler algorithms for data that are non-uniformly distributed; where multiple algorithms
573 produce similar regional predictions, simple learners can offer greater stability and interpretability
574 (Domingos 2012).

575 RFR, like all other machine learning models, is still sensitive to the representativeness of training
576 data and is subject to certain performance limitations (Millard and Richardson 2015). The nature of
577 in-situ sampling with ocean profilers like floats and gliders poses a challenge for model development
578 because the training observations are not randomly distributed throughout the region and time
579 period of interest. Although we select training observations from BGC-Argo floats that appear
580 to sample tracer characteristics of the same ASZ region in which the SOGOS experiment takes
581 place (Figure 3), information from these six floats is not sufficient to represent the full range of
582 tracer relationships in this region, nor how they change over time. Another caveat to RFR is that
583 we choose a feature list for model training based on the assumption that nitrate content correlates
584 strongly with given predictive variables like CT, SA, or O_2 , but this correlation may be weaker
585 in different parts of the Southern Ocean (Ishizu and Richards 2013). The same submesoscale

586 processes we attempt to diagnose using the high-resolution RFR nitrate may already be responsible
587 for greater decoupling between nitrate and other variables in-situ (Mahadevan 2016; Omand and
588 Mahadevan 2013). Despite these limitations, our regional RFR produces remarkably low test
589 prediction errors in one of the most turbulent areas of the global oceans. The success of our RFR
590 approach highlights the potential for machine learning to improve mapping of ocean fields.

591 Machine learning and artificial intelligence have been increasingly applied for oceanographic
592 applications (Sonnewald et al. 2021; Sun et al. 2022), including improving satellite altimetry
593 products (Martin et al. 2023; Cohen 2019; Fan et al. 2021), estimating biogeochemical distributions
594 (Bittig et al. 2018; Carter et al. 2021), and identifying eddy activity (Ashkezari et al. 2016; Zhang
595 et al. 2023), among many others. Future application of RFR or other machine learning approaches
596 on other multi-platform datasets can be used to address a wide range of questions depending on
597 what different types of measurements can be synthesized (Salam and Hsieh 2023; Salcedo-Sanz
598 et al. 2020). Here, our observation-based RFR approach to nitrate prediction in the Southern Ocean
599 extends previous simulation and theory on high-frequency nutrient dynamics. Given increasing
600 observational coverage of the global oceans by Argo floats and other drifting profilers, RFR
601 presents opportunities to derive additional value from these sometimes incomplete biogeochemical
602 datasets. Such efforts to bridge observational gaps using new ocean technologies and machine
603 learning techniques will expand our knowledge of global biogeochemical cycles at previously
604 inaccessible scales.

Acknowledgments. This work is supported by NSF awards OCE-1756956 and OCE-1756882. SS and ARG are also supported by NASA award NNX80NSSC19K1252, the U.S. Argo Program through NOAA award NA20OAR4320271, NSF award OCE-2148434, and NSF's Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project through award OPP-1936222. PDL was supported by the NOAA grant NA19NES4320002 (Cooperative Institute for Satellite Earth System Studies, CISESS) at the University of Maryland/ESSIC. We thank Geoff Shilling and Craig Lee at APL for their efforts in reprocessing of the glider data. We also extend sincere thanks to Yuichiro Takeshita for offering his insights and code for processing the glider oxygen optode logs. We use colormaps obtained from the cmocean package (Thyng et al. 2016).

Data availability statement. The SOGOS data for Seagliders SG659 and SG660 can be accessed through Balwada (2023); DOI 10.5281/zenodo.8361656. Argo float data were collected and made freely available by the International Argo Program and the national programs that contribute to it (<https://argo.ucsd.edu>, <https://www.ocean-ops.org>). The Argo Program is part of the Global Ocean Observing System; Argo float data and metadata from Global Data Assembly Centre (Argo (2021); DOI 10.17882/42182). Shipboard data were collected and made publicly available by the International Global Ship-based Hydrographic Investigations Program (GO-SHIP; <http://www.go-ship.org/>) and the national programs that contribute to it. The satellite altimetry data are freely available through the E.U. Copernicus Marine Environment Monitoring Service (CMEMS; DOI 10.48670/moi-00148), and the value-added FSLE product is provided by AVISO (<https://www.aviso.altimetry.fr/en/data/products/value-added-products/fsle-finite-size-lyapunov-exponents.html>; DOI 10.24400/527896/a01-2022.002). MODIS-Aqua satellite data are hosted by NOAA and provided through the NASA Ocean Biology Processing Group (<https://coastwatch.pfeg.noaa.gov/erddap/griddap/erdMH1par08day.html>).

APPENDIX

Data Quality Control and Processing

a. Seaglider Processing

The glider data was reprocessed into an L2 xarray Dataset (courtesy of Geoff Shilling and Craig Lee, Applied Physics Lab) which separates the data into glider profiles and averages the

raw observations in 1 m depth intervals from 0 to 1000 m. We use the despiked L3 product, which interpolates observations vertically (gaps ≤ 50 m in depth) removes outliers more than 2 standard deviations from the running mean. With quality-controlled BGC-Argo and GO-SHIP data for reference, additive corrections for CT and SA were determined by finding the closest glider matches to the quality-controlled Argo measurements, which are assumed to be true reference values. A threshold of ~ 5 m depth difference and distance of ~ 10 km were used as an upper threshold to filter matches, consistent with float and bottle match thresholds for standard BGC-Argo quality control (Maurer et al. 2021). The difference dT and dS for each observation pair was calculated to represent the glider offsets (glider minus float), and had statistically insignificant slopes along depth. A single additive correction was made for all profiles from one glider. For temperature, 0.0629°C was added to all profiles from SG659, and 0.030°C added to those from SG660. For salinity, SG659 had negligible corrections while 0.18 psu was subtracted from SG660 measurements. These corrections are comparable to those performed on the same dataset from Dove et al. (2021).

We also correct oxygen since oxygen optodes on the rapidly sampling gliders are prone to a time response lag. As the gliders ascend and descend, a small boundary layer develops around the head of the optode. The oxygen measurement therefore lags behind the true oxygen concentration, creating a tendency for gliders to slightly overestimate oxygen on a downward cast, and to underestimate on an upward cast. Methods for optode lag corrections on Argo floats (code courtesy of Yuichiro Takeshita, Stoer et al. (2023)) were adapted to the glider optodes. For the standard foil Aanderaa optodes on the gliders, a boundary layer thickness of ~ 40 μm was chosen, while the time response is calculated internally following Bittig et al. (2018). The oxygen sensor is also corrected for an offset using a gain correction (Johnson et al. 2015). For SG659, we calculated the corrected oxygen as $\text{O}_{2\text{corr}} = 1.126(\text{O}_2) - 3.256$; for SG660, $\text{O}_{2\text{corr}} = 1.0866(\text{O}_2) - 0.146$.

b. GO-SHIP Processing

Bottle data from the GO-SHIP line I06 in 2019 were accessed through the CLIVAR and Carbon Hydrographic Data Office (CCHDO; <https://cchdo.ucsd.edu/cruise/325020190403>). Quality control of the bottle data is performed using the provided flags (2: no problems noted).

c. BGC-Argo Processing

Delayed-mode BGC-Argo data from seven floats (WMO: 5904469, 5904659, 5905368, 5905996, 5906030, 5906031, 5906207) are downloaded from an Argo Global Data Assembly Center (GDAC) using the Python BGC-Argo Toolbox (<https://github.com/go-bgc/workshop-python>). Quality control is performed using standard BGC-Argo QC flags (1: good data, 2: probably good data, 8: interpolated value).

References

Argo, 2021: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE, <https://doi.org/10.17882/42182>.

Ashkezari, M. D., C. N. Hill, C. N. Follett, G. Forget, and M. J. Follows, 2016: Oceanic eddy detection and lifetime forecast using machine learning methods. *Geophys. Res. Lett.*, **43** (23), 12,234–12,241, <https://doi.org/10.1002/2016GL071269>.

Balwada, D., 2023: Tracer stirring and variability in the antarctic circumpolar current near the Southwest Indian Ridge Analysis Code and Data. Zenodo, <https://doi.org/10.5281/zenodo.8361656>.

Balwada, D., A. R. Gray, L. A. Dove, and A. F. Thompson, 2024: Tracer Stirring and Variability in the Antarctic Circumpolar Current Near the Southwest Indian Ridge. *J. Geophys. Res. Oceans*, **129** (1), e2023JC019811, <https://doi.org/10.1029/2023JC019811>.

Balwada, D., K. S. Smith, and R. Abernathey, 2018: Submesoscale Vertical Velocities Enhance Tracer Subduction in an Idealized Antarctic Circumpolar Current. *Geophys. Res. Lett.*, **45** (18), 9790–9802, <https://doi.org/10.1029/2018GL079244>.

Birchill, A. J., and Coauthors, 2021: Exploring Ocean Biogeochemistry Using a Lab-on-Chip Phosphate Analyser on an Underwater Glider. *Front. Mar. Sci.*, **8**, <https://doi.org/10.3389/fmars.2021.698102>.

Bittig, H. C., T. Steinhoff, H. Claustre, B. Fiedler, N. L. Williams, R. Sauzède, A. Körtzinger, and J.-P. Gattuso, 2018: An Alternative to Static Climatologies: Robust Estimation of Open Ocean

CO₂ Variables and Nutrient Concentrations From T, S, and O₂ Data Using Bayesian Neural Networks. *Front. Mar. Sci.*, **5**, <https://doi.org/10.3389/fmars.2018.00328>.

Brannigan, L., 2016: Intense submesoscale upwelling in anticyclonic eddies. *Geophys. Res. Lett.*, **43** (7), 3360–3369, <https://doi.org/10.1002/2016GL067926>.

Breiman, L., 2001: Random Forests. *Mach. Learn.*, **45** (1), 5–32, <https://doi.org/10.1023/A:1010933404324>.

Callens, A., D. Morichon, S. Abadie, M. Delpy, and B. Lique, 2020: Using Random forest and Gradient boosting trees to improve wave forecast at a specific location. *Appl. Ocean Res.*, **104**, 102 339, <https://doi.org/10.1016/j.apor.2020.102339>.

Carter, B. R., and Coauthors, 2021: New and updated global empirical seawater property estimation routines. *Limnol. Oceanogr.: Methods*, **19** (12), 785–809, <https://doi.org/10.1002/lom3.10461>.

Chai, F., and Coauthors, 2020: Monitoring ocean biogeochemistry with autonomous platforms. *Nat. Rev. Earth. Environ.*, **1** (6), 315–326, <https://doi.org/10.1038/s43017-020-0053-y>.

Claustre, H., K. S. Johnson, and Y. Takeshita, 2020: Observing the Global Ocean with Biogeochemical-Argo. *Ann. Rev. Mar. Sci.*, **12** (Volume 12, 2020), 23–48, <https://doi.org/10.1146/annurev-marine-010419-010956>.

Cohen, M. X., 2019: A better way to define and describe Morlet wavelets for time-frequency analysis. *NeuroImage*, **199**, 81–86, <https://doi.org/10.1016/j.neuroimage.2019.05.048>.

Domingos, P., 2012: A few useful things to know about machine learning. *Commun. ACM*, **55** (10), 78–87, <https://doi.org/10.1145/2347736.2347755>.

Dove, L. A., A. F. Thompson, D. Balwada, and A. R. Gray, 2021: Observational Evidence of Ventilation Hotspots in the Southern Ocean. *J. Geophys. Res. Oceans*, **126** (7), <https://doi.org/10.1029/2021JC017178>.

d’Ovidio, F., V. Fernández, E. Hernández-García, and C. López, 2004: Mixing structures in the Mediterranean Sea from finite-size Lyapunov exponents. *Geophys. Res. Lett.*, **31** (17), <https://doi.org/10.1029/2004GL020328>.

713 Erickson, Z. K., and A. F. Thompson, 2018: The Seasonality of Physically Driven Export at
 714 Submesoscales in the Northeast Atlantic Ocean. *Global Biogeochem. Cycles*, **32** (8), 1144–
 715 1162, <https://doi.org/10.1029/2018GB005927>.

716 Erickson, Z. K., A. F. Thompson, N. Cassar, J. Sprintall, and M. R. Mazloff, 2016: An advective
 717 mechanism for deep chlorophyll maxima formation in southern Drake Passage. *Geophys. Res.*
 718 *Lett.*, **43** (20), <https://doi.org/10.1002/2016GL070565>.

719 Fan, Y., and Coauthors, 2021: OC-SMART: A machine learning based data analysis platform for
 720 satellite ocean color sensors. *Remote Sens. Environ.*, **253**, 112 236, [https://doi.org/10.1016/j.rse.](https://doi.org/10.1016/j.rse.2020.112236)
 721 2020.112236.

722 Foster, G., 1996: Wavelets for period analysis of unevenly sampled time series. *Astron. J.*, **112**,
 723 1709, <https://doi.org/10.1086/118137>.

724 Freilich, M. A., and A. Mahadevan, 2019: Decomposition of Vertical Velocity for Nutrient
 725 Transport in the Upper Ocean. *J. Phys. Oceanogr.*, **49** (6), 1561–1575, [https://doi.org/10.1175/](https://doi.org/10.1175/JPO-D-19-0002.1)
 726 JPO-D-19-0002.1.

727 Giglio, D., V. Lyubchich, and M. R. Mazloff, 2018: Estimating Oxygen in the Southern Ocean Using
 728 Argo Temperature and Salinity. *J. Geophys. Res. Oceans*, **123** (6), 4280–4297, [https://doi.org/](https://doi.org/10.1029/2017JC013404)
 729 10.1029/2017JC013404.

730 Gray, A. R., 2024: The Four-Dimensional Carbon Cycle of the Southern Ocean. *Ann. Rev. Mar.*
 731 *Sci.*, **16** (Volume 16, 2024), 163–190, <https://doi.org/10.1146/annurev-marine-041923-104057>.

732 Grinsted, A., J. C. Moore, and S. Jevrejeva, 2004: Application of the cross wavelet transform and
 733 wavelet coherence to geophysical time series. *Nonlinear Process. Geophys.*, **11** (5/6), 561–566,
 734 <https://doi.org/10.5194/npg-11-561-2004>.

735 Henley, S. F., and Coauthors, 2020: Changing Biogeochemistry of the Southern Ocean and Its
 736 Ecosystem Implications. *Front. Mar. Sci.*, **7**, <https://doi.org/10.3389/fmars.2020.00581>.

737 IOC, SCOR, and IAPSO, 2010: *The International thermodynamic equation of seawater, 2010: cal-*
 738 *culation and use of thermodynamic properties*. No. 56, Intergovernmental Oceanographic Com-
 739 mission, Manuals and Guides, UNESCO, URL [https://policycommons.net/artifacts/8333631/](https://policycommons.net/artifacts/8333631/the-international-thermodynamic-equation-of-seawater-2010/9264182/)
 740 the-international-thermodynamic-equation-of-seawater-2010/9264182/.

- Ishizu, M., and K. J. Richards, 2013: Relationship between oxygen, nitrate, and phosphate in the world ocean based on potential temperature. *J. Geophys. Res. Oceans*, **118** (7), 3586–3594, <https://doi.org/10.1002/jgrc.20249>.
- Johnson, K. S., J. N. Plant, S. C. Riser, and D. Gilbert, 2015: Air Oxygen Calibration of Oxygen Optodes on a Profiling Float Array. *J. Atmos. Oceanic Tech.*, **32** (11), 2160–2172, <https://doi.org/10.1175/JTECH-D-15-0101.1>.
- Khider, D., J. Emile-Geay, F. Zhu, A. James, J. Landers, V. Ratnakar, and Y. Gil, 2022: Pyleoclim: Paleoclimate Timeseries Analysis and Visualization With Python. *Paleoceanogr. Paleoclimatology*, **37** (10), e2022PA004 509, <https://doi.org/10.1029/2022PA004509>.
- Klein, P., and G. Lapeyre, 2009: The Oceanic Vertical Pump Induced by Mesoscale and Submesoscale Turbulence. *Annu. Rev. Mar. Sci.*, **1** (1), 351–375, <https://doi.org/10.1146/annurev.marine.010908.163704>.
- Kohavi, R., 1995: A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1137–1143, IJCAI'95.
- Le Rest, K., D. Pinaud, P. Monestiez, J. Chadoeuf, and V. Bretagnolle, 2014: Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Glob. Ecol. Biogeogr.*, **23** (7), 811–820, <https://doi.org/10.1111/geb.12161>.
- Lermusiaux, P. F. J., and Coauthors, 2017: A future for intelligent autonomous ocean observing systems. *J. Mar. Res.*, **75** (6), 765–813.
- Levy, M., and A. P. Martin, 2013: The influence of mesoscale and submesoscale heterogeneity on ocean biogeochemical reactions. *Global Biogeochem. Cycles*, **27** (4), 1139–1150, <https://doi.org/10.1002/2012GB004518>.
- Lévy, M., D. Couespel, C. Haëck, M. G. Keerthi, I. Mangolte, and C. J. Prend, 2024: The Impact of Fine-Scale Currents on Biogeochemical Cycles in a Changing Ocean. *Ann. Rev. Mar. Sci.*, **16** (Volume 16, 2024), 191–215, <https://doi.org/10.1146/annurev-marine-020723-020531>.
- Lévy, M., P. J. S. Franks, and K. S. Smith, 2018: The role of submesoscale currents in structuring marine ecosystems. *Nat. Commun.*, **9** (1), 4758, <https://doi.org/10.1038/s41467-018-07059-3>.

- 769 Lévy, M., D. Iovino, L. Resplandy, P. Klein, G. Madec, A. M. Tréguier, S. Masson, and K. Taka-
770 hashi, 2012: Large-scale impacts of submesoscale dynamics on phytoplankton: Local and
771 remote effects. *Ocean Model.*, **43-44**, 77–93, <https://doi.org/10.1016/j.ocemod.2011.12.003>.
- 772 Mahadevan, A., 2016: The Impact of Submesoscale Physics on Primary Productivity of Plankton.
773 *Ann. Rev. Mar. Sci.*, **8** (1), 161–184, <https://doi.org/10.1146/annurev-marine-010814-015912>.
- 774 Mahadevan, A., and D. Archer, 2000: Modeling the impact of fronts and mesoscale circulation on
775 the nutrient supply and biogeochemistry of the upper ocean. *J. Geophys. Res. Oceans*, **105** (C1),
776 1209–1225, <https://doi.org/10.1029/1999JC900216>.
- 777 Mahadevan, A., and A. Tandon, 2006: An analysis of mechanisms for submesoscale vertical motion
778 at ocean fronts. *Ocean Model.*, **14** (3), 241–256, <https://doi.org/10.1016/j.ocemod.2006.05.006>.
- 779 Martin, S. A., G. E. Manucharyan, and P. Klein, 2023: Synthesizing Sea Surface Temperature and
780 Satellite Altimetry Observations Using Deep Learning Improves the Accuracy and Resolution
781 of Gridded Sea Surface Height Anomalies. *J. Adv. Model Earth Sy.*, **15** (5), e2022MS003 589,
782 <https://doi.org/10.1029/2022MS003589>.
- 783 Maurer, T. L., J. N. Plant, and K. S. Johnson, 2021: Delayed-Mode Quality Control of Oxy-
784 gen, Nitrate, and pH Data on SOCCOM Biogeochemical Profiling Floats. *Front. Mar. Sci.*, **8**,
785 <https://doi.org/10.3389/fmars.2021.683207>.
- 786 Millard, K., and M. Richardson, 2015: On the Importance of Training Data Sample Selection in
787 Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote*
788 *Sens.*, **7** (7), 8489–8515, <https://doi.org/10.3390/rs70708489>.
- 789 Olsen, A., and Coauthors, 2019: GLODAPv2.2019 – an update of GLODAPv2. *Earth Syst. Sci.*
790 *Data*, **11** (3), 1437–1461, <https://doi.org/10.5194/essd-11-1437-2019>.
- 791 Omand, M. M., and A. Mahadevan, 2013: Large-scale alignment of oceanic nitrate and density. *J.*
792 *Geophys. Res. Oceans*, **118** (10), 5322–5332, <https://doi.org/10.1002/jgrc.20379>.
- 793 Patel, R. S., J. Llorc, P. G. Strutton, H. E. Phillips, S. Moreau, P. Conde Pardo, and A. Lenton,
794 2020: The Biogeochemical Structure of Southern Ocean Mesoscale Eddies. *J. Geophys. Res.*
795 *Oceans*, **125** (8), e2020JC016 115, <https://doi.org/10.1029/2020JC016115>.

- 796 Possenti, L., and Coauthors, 2021: Air-Sea Gas Fluxes and Remineralization From a Novel
797 Combination of pH and O₂ Sensors on a Glider. *Front. Mar. Sci.*, **8**, <https://doi.org/10.3389/fmars.2021.696772>.
798
- 799 Renosh, P. R., J. Zhang, R. Sauzède, and H. Claustre, 2023: Vertically Resolved Global Ocean
800 Light Models Using Machine Learning. *Remote Sens.*, **15** (24), 5663, <https://doi.org/10.3390/rs15245663>.
801
- 802 Rintoul, S. R., and A. C. Naveira Garabato, 2013: Dynamics of the South-
803 ern Ocean Circulation. *International Geophysics*, Vol. 103, Elsevier, 471–492,
804 <https://doi.org/10.1016/B978-0-12-391851-2.00018-0>, URL <https://linkinghub.elsevier.com/retrieve/pii/B9780123918512000180>.
805
- 806 Roemmich, D., and Coauthors, 2019: On the Future of Argo: A Global, Full-Depth, Multi-
807 Disciplinary Array. *Front. Mar. Sci.*, **6**, <https://doi.org/10.3389/fmars.2019.00439>.
- 808 Rosso, I., A. M. Hogg, R. Matear, and P. G. Strutton, 2016: Quantifying the influence of sub-
809 mesoscale dynamics on the supply of iron to Southern Ocean phytoplankton blooms. *Deep-Sea Res. I*, **115**, 199–209, <https://doi.org/10.1016/j.dsr.2016.06.009>.
810
- 811 Rudnick, D. L., 2016: Ocean Research Enabled by Underwater Gliders. *Ann. Rev. Mar. Sci.*,
812 **8** (Volume 8, 2016), 519–541, <https://doi.org/10.1146/annurev-marine-122414-033913>.
- 813 Salam, T., and M. A. Hsieh, 2023: Heterogeneous robot teams for modeling and prediction of
814 multiscale environmental processes. *Auton. Robots*, **47** (4), 353–376, <https://doi.org/10.1007/s10514-023-10089-6>.
815
- 816 Salcedo-Sanz, S., and Coauthors, 2020: Machine learning information fusion in Earth observation:
817 A comprehensive review of methods, applications and data sources. *Inform. Fusion*, **63**, 256–272,
818 <https://doi.org/10.1016/j.inffus.2020.07.004>.
- 819 Sarmiento, J. L., and Coauthors, 2023: The Southern Ocean carbon and climate observations
820 and modeling (SOCCOM) project: A review. *Prog. Oceanogr.*, **219**, 103 130, <https://doi.org/10.1016/j.pocean.2023.103130>.
821
- 822 Sauv  , J., A. R. Gray, C. J. Prend, S. M. Bushinsky, and S. C. Riser, 2023: Carbon Outgassing
823 in the Antarctic Circumpolar Current Is Supported by Ekman Transport From the Sea Ice Zone

in an Observation-Based Seasonal Mixed-Layer Budget. *J. Geophys. Res. Oceans*, **128** (11), e2023JC019815, <https://doi.org/10.1029/2023JC019815>.

Sharp, J. D., A. J. Fassbender, B. R. Carter, G. C. Johnson, C. Schultz, and J. P. Dunne, 2022a: GOBAI-O2: A Global Gridded Monthly Dataset of Ocean Interior Dissolved Oxygen Concentrations Based on Shipboard and Autonomous Observations (NCEI Accession 0259304). NOAA National Centers for Environmental Information, URL <https://www.ncei.noaa.gov/archive/accession/0259304>, <https://doi.org/10.25921/Z72M-YZ67>.

Sharp, J. D., A. J. Fassbender, B. R. Carter, P. D. Lavin, and A. J. Sutton, 2022b: A monthly surface $p\text{CO}_2$ product for the California Current Large Marine Ecosystem. *Earth Syst. Sci. Data*, **14** (4), 2081–2108, <https://doi.org/10.5194/essd-14-2081-2022>.

Siegelman, L., P. Klein, A. F. Thompson, H. S. Torres, and D. Menemenlis, 2020: Altimetry-Based Diagnosis of Deep-Reaching Sub-Mesoscale Ocean Fronts. *Fluids*, **5** (3), 145, <https://doi.org/10.3390/fluids5030145>.

Sonnenwald, M., R. Lguensat, D. C. Jones, P. D. Dueben, J. Brajard, and V. Balaji, 2021: Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environ. Res. Lett.*, **16** (7), 073008, <https://doi.org/10.1088/1748-9326/ac0eb0>.

Stock, A., 2022: Spatiotemporal distribution of labeled data can bias the validation and selection of supervised learning algorithms: A marine remote sensing example. *ISPRS J. Photogramm. Remote Sens.*, **187**, 46–60, <https://doi.org/10.1016/j.isprsjprs.2022.02.023>.

Stock, A., and A. Subramaniam, 2022: Iterative spatial leave-one-out cross-validation and gap-filling based data augmentation for supervised learning applications in marine remote sensing. *GISci. Remote Sens.*, **59** (1), 1281–1300, <https://doi.org/10.1080/15481603.2022.2107113>.

Stoer, A. C., and Coauthors, 2023: A census of quality-controlled Biogeochemical-Argo float measurements. *Front. Mar. Sci.*, **10**, 1233289, <https://doi.org/10.3389/fmars.2023.1233289>.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn, 2007: Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, **8** (1), 25, <https://doi.org/10.1186/1471-2105-8-25>.

- 851 Su, J., C. Schallenberg, T. Rohr, P. G. Strutton, and H. E. Phillips, 2022: New Estimates of Southern
852 Ocean Annual Net Community Production Revealed by BGC-Argo Floats. *Geophys. Res. Lett.*,
853 **49 (15)**, e2021GL097372, <https://doi.org/10.1029/2021GL097372>.
- 854 Su, J., P. G. Strutton, and C. Schallenberg, 2021: The subsurface biological structure of Southern
855 Ocean eddies revealed by BGC-Argo floats. *J. Mar. Syst.*, **220**, 103569, [https://doi.org/10.1016/](https://doi.org/10.1016/j.jmarsys.2021.103569)
856 [j.jmarsys.2021.103569](https://doi.org/10.1016/j.jmarsys.2021.103569).
- 857 Sun, Z., and Coauthors, 2022: A review of Earth Artificial Intelligence. *Comput. Geosci.*, **159**,
858 105034, <https://doi.org/10.1016/j.cageo.2022.105034>.
- 859 Swart, S., and Coauthors, 2023: The Southern Ocean mixed layer and its boundary fluxes: fine-
860 scale observational progress and future research priorities. *Philos. Transact. A Math. Phys. Eng.*
861 *Sci.*, **381 (2249)**, 20220058, <https://doi.org/10.1098/rsta.2022.0058>.
- 862 Talley, L. D., 2013: Closure of the Global Overturning Circulation Through the Indian, Pacific,
863 and Southern Oceans: Schematics and Transports. *Oceanog.*, **26 (1)**, 80–97, [https://doi.org/](https://doi.org/10.5670/oceanog.2013.07)
864 [10.5670/oceanog.2013.07](https://doi.org/10.5670/oceanog.2013.07).
- 865 Taylor, J. R., and A. F. Thompson, 2023: Submesoscale Dynamics in the Upper
866 Ocean. *Ann. Rev. Fluid Mech.*, **55 (Volume 55, 2023)**, 103–127, [https://doi.org/10.1146/](https://doi.org/10.1146/annurev-fluid-031422-095147)
867 [annurev-fluid-031422-095147](https://doi.org/10.1146/annurev-fluid-031422-095147).
- 868 Thomas, L. N., A. Tandon, and A. Mahadevan, 2008: Submesoscale processes and dynam-
869 ics. *Geophysical Monograph Series*, M. W. Hecht, and H. Hasumi, Eds., Vol. 177, Ameri-
870 can Geophysical Union, Washington, D. C., 17–38, <https://doi.org/10.1029/177GM04>, URL
871 <https://onlinelibrary.wiley.com/doi/10.1029/177GM04>.
- 872 Thompson, A. F., A. Lazar, C. Buckingham, A. C. Naveira Garabato, G. M. Damerell, and
873 K. J. Heywood, 2016: Open-Ocean Submesoscale Motions: A Full Seasonal Cycle of Mixed
874 Layer Instabilities from Gliders. *J. Phys. Oceanogr.*, **46 (4)**, 1285–1307, [https://doi.org/10.1175/](https://doi.org/10.1175/JPO-D-15-0170.1)
875 [JPO-D-15-0170.1](https://doi.org/10.1175/JPO-D-15-0170.1).
- 876 Thyng, K., C. Greene, R. Hetland, H. Zimmerle, and S. DiMarco, 2016: True Colors of Oceanog-
877 raphy: Guidelines for Effective and Accurate Colormap Selection. *Oceanog.*, **29 (3)**, 9–13,
878 <https://doi.org/10.5670/oceanog.2016.66>.

- 879 Tong, S., X. Liu, Q. Chen, Z. Zhang, and G. Xie, 2019: Multi-Feature Based Ocean Oil Spill
880 Detection for Polarimetric SAR Data Using Random Forest and the Self-Similarity Parameter.
881 *J. Remote Sens.*, **11** (4), 451, <https://doi.org/10.3390/rs11040451>.
- 882 Torrence, C., and G. P. Compo, 1998: A Practical Guide to Wavelet Analysis.
883 *Bull.Amer.Meteor.Soc.*, **79** (1), 61–78, [https://doi.org/10.1175/1520-0477\(1998\)079<0061:
884 APGTWA>2.0.CO;2](https://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2).
- 885 Uchida, T., D. Balwada, R. Abernathey, G. McKinley, S. Smith, and M. Lévy, 2019: The Contri-
886 bution of Submesoscale over Mesoscale Eddy Iron Transport in the Open Southern Ocean. *J.*
887 *Adv. Model Earth Sy.*, **11** (12), 3934–3958, <https://doi.org/10.1029/2019MS001805>.
- 888 Uchida, T., D. Balwada, R. P. Abernathey, G. A. McKinley, S. K. Smith, and M. Lévy, 2020: Vertical
889 eddy iron fluxes support primary production in the open Southern Ocean. *Nat. Commun.*, **11** (1),
890 1125, <https://doi.org/10.1038/s41467-020-14955-0>.
- 891 Vincent, A. G., R. W. Pascal, A. D. Beaton, J. Walk, J. E. Hopkins, E. M. S. Woodward, M. Mowlem,
892 and M. C. Lohan, 2018: Nitrate drawdown during a shelf sea spring bloom revealed using a
893 novel microfluidic *in situ* chemical sensor deployed within an autonomous underwater glider.
894 *Mar. Chem.*, **205**, 29–36, <https://doi.org/10.1016/j.marchem.2018.07.005>.
- 895 Wong, A. P. S., and Coauthors, 2020: Argo Data 1999–2019: Two Million Temperature-Salinity
896 Profiles and Subsurface Velocity Observations From a Global Array of Profiling Floats. *Front.*
897 *Mar. Sci.*, **7**, <https://doi.org/10.3389/fmars.2020.00700>.
- 898 Yung, C. K., A. K. Morrison, and A. M. Hogg, 2022: Topographic Hotspots of Southern Ocean
899 Eddy Upwelling. *Front. Mar. Sci.*, **9**, <https://doi.org/10.3389/fmars.2022.855785>.
- 900 Zhang, G., R. Chen, X. Li, L. Li, H. Wei, and W. Guan, 2023: Temporal Variability of Global
901 Surface Eddy Diffusivities: Estimates and Machine Learning Prediction. *J. Phys. Oceanogr.*,
902 **53** (7), 1711–1730, <https://doi.org/10.1175/JPO-D-22-0251.1>.
- 903 Zhou, Z., C. Qiu, and Y. Zhang, 2023: A comparative analysis of linear regression, neural networks
904 and random forest regression for predicting air ozone employing soft sensor models. *Sci. Rep.*,
905 **13**, 22 420, <https://doi.org/10.1038/s41598-023-49899-0>.