A Multivariate Empirical Orthogonal Function Method to Construct Nitrate Maps in the Southern Ocean

YU-CHIAO LIANG

Department of Earth System Science, University of California, Irvine, Irvine, California

MATTHEW R. MAZLOFF AND ISABELLA ROSSO

Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

SHIH-WEI FANG AND JIN-YI YU

Department of Earth System Science, University of California, Irvine, Irvine, California

(Manuscript received 7 February 2018, in final form 30 April 2018)

ABSTRACT

The ability to construct nitrate maps in the Southern Ocean (SO) from sparse observations is important for marine biogeochemistry research, as it offers a geographical estimate of biological productivity. The goal of this study is to infer the skill of constructed SO nitrate maps using varying data sampling strategies. The mapping method uses multivariate empirical orthogonal functions (MEOFs) constructed from nitrate, salinity, and potential temperature (N-S-T) fields from a biogeochemical general circulation model simulation Synthetic N-S-T datasets are created by sampling modeled N-S-T fields in specific regions, determined either by random selection or by selecting regions over a certain threshold of nitrate temporal variances. The first 500 MEOF modes, determined by their capability to reconstruct the original N-S-T fields, are projected onto these synthetic N-S-T data to construct time-varying nitrate maps. Normalized root-mean-square errors (NRMSEs) are calculated between the constructed nitrate maps and the original modeled fields for different sampling strategies. The sampling strategy according to nitrate variances is shown to yield maps with lower NRMSEs than mapping adopting random sampling. A *k*-means cluster method that considers the N-S-T combined variances to identify key regions to insert data is most effective in reducing the mapping errors. These findings are further quantified by a series of mapping error analyses that also address the significance of data sampling density. The results provide a sampling framework to prioritize the deployment of biogeochemical Argo floats for constructing nitrate maps.

1. Introduction

Nitrate, mostly in its dissolved form NO_3^-, is an essential element for supplying and sustaining marine biological productivity in the global oceans (Moore et al. 2013). The amount of nitrate serves as an important limiting nutrient, altering the structure and function of phytoplankton communities (Dugdale and Goering 1967; Church et al. 2000; Moore et al. 2013); and studies have been suggested to regulate the strength of the biological pump (Elderfield 2006, chapter 6; Ducklow et al. 2001; Ardyna et al. 2017), which is a pivotal part of the global biogeochemical cycles (Deppeler and Davidson 2017 Nitrate drawdown is also a good indicator of the net community production (Arrigo 2005; Munro et al. 2015; Plant et al. 2016; Johnson et al. 2017). Johnson et al. (2017) estimated that the annually averaged net community production in the Southern Ocean (SO) is 1.3 PgC yr^{-1}, which accounts for about 13% of the global annual net community production. Therefore, constructing comprehensive, accurate nitrate maps offers geographical estimates of bioproductivity to help understand the marine biogeochemical state and subsequent impacts on global climate.

In the SO, the large-scale nitrate distribution is largely determined by lateral and vertical transport processes (Williams and Follows 2003). Letscher et al. (2016) estimated that 17%–20% of the total nitrate in the low-latitude SO regions is transported by nutrient-rich water masses from high latitudes. Verdy and Mazloff (2017)

Corresponding author: Yu-Chiao Liang, yuchiaol@uci.edu

DOI: 10.1175/JTECH-D-18-0018.1

recently made a consistent transport estimate using an SO biogeochemical state estimate. Other studies, based on in situ observations, found that the Antarctic Circumpolar Current can carry a great amount of nitrate into downstream areas to cause abrupt phytoplankton blooms (Hoppe et al. 2015). As for vertical nitrate transport, the biological pump mechanism is capable of exchanging nitrate between the surface and interior ocean (Williams and Follows 2003; Elderfield 2006, chapter 6). Because advective transport processes also redistribute other properties (e.g., salinity and potential temperature), there exist certain large-scale correspondences between these properties and nitrate fields, such as the nitrate–potential temperature relationship identified by Ishizu and Richards (2013). These spatial correspondences shed light on the possibility to inform SO nitrate information with the assistance of other tracer fields.

The major challenge of constructing SO nitrate maps is the scarcity of in situ measurements. However, the situation is improving. The Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) program recently reported that 31 profiling floats carrying nitrate sensors have successfully transmitted 40 complete nitrate annual cycles (Johnson et al. 2017), and additional float deployments are being carried out. Multiple methods and techniques have been developed to reconstruct fields in regions with sparse observational data, such as optimal interpolation (Reynolds and Smith 1994; Schneider 2001), model-based gap-filling techniques (e.g., data assimilation; Stammer et al. 2002; Wunsch and Heimbach 2007; Mazloff et al. 2010; Verdy and Mazloff 2017), and empirical orthogonal function (EOF)-based methods (e.g., Smith et al. 1996; Kaplan et al. 1997; Beckers and Rixen 2003; Alvera-Azcárate et al. 2005; Kondrashov and Ghil 2006; Alvera-Azcárate et al. 2007; Alvera-Azcárate et al. 2011; Nikolaidis et al. 2014; Alvera-Azcárate et al. 2016). The EOF-based methods, in particular, show advantages over the other methods in terms of ease of implementation and accuracy relative to computational costs (Alvera-Azcárate et al. 2005).

The multivariate EOF (MEOF) method, a variant of the archetypal EOF method, has been widely used for investigating large-scale atmospheric and oceanic coupled variability structures because of its salient ability to incorporate different variables with their combined variances (Xue et al. 2000; Sparnocchia et al. 2003; Wheeler and Hendon 2004, Alvera-Azcárate et al. 2007). The MEOF method has also been applied to reconstruct maps with the assistance of several related fields. For example, the MEOF method has been used to synthesize temperature–salinity information (De Mey and Robinson 1987; Fukumori and Wunsch 1991), and reconstruct sea surface temperatures with the assistance of chlorophyll-*a* and wind fields from satellite observations (Alvera-Azcárate et al. 2007). However, the applicability of the MEOF method for constructing SO nitrate maps has not yet been investigated.

Furthermore, sampling strategies determining observational requirements to construct maps target spatial correlation structure, but they often neglect the fact that temporal variance is extremely heterogeneous (Dormann et al. 2007; Wang et al. 2012). The implementation of the MEOF method considering both the spatial and temporal information may allow improved mapping of nitrate. The goal of this study is to apply the MEOF method to construct SO nitrate maps and to assess optimal data sampling strategies addressing both signal structure and amplitude.

In section 2 we describe the biogeochemical general circulation model used in this study that provides the reference nitrate–salinity–potential temperature (N-S-T) fields for the mapping task and further analyses. The basics of the MEOF calculations and *k*-means cluster method are also introduced. Section 3 presents the procedure and explains how we sample N-S-T data and construct SO nitrate maps, accompanied with a series of mapping error and sampling density analyses. In section 4 we discuss the results and caveats in the context of prioritizing deployment of in situ measurements, such as biogeochemical Argo floats, to best inform mapping of SO nitrate.

2. Model and methodology

a. Biogeochemical general circulation model

The biogeochemical general circulation model (GCM) used in this study to provide the reference nitrate, salinity, and potential temperature (temperature hereinafter) fields is the MITgcm (Marshall et al. 1997) coupled to the modified Biogeochemistry with Light, Iron, Nutrients, and Gas (BLING) model (Galbraith et al. 2010). A sea ice component is also included (Losch et al. 2010). This biogeochemical GCM setup has been applied to estimate SO dynamical and biogeochemical states (Verdy and Mazloff 2017). The model domain is 78°–30°S at 1/3° resolution with a Mercator projection, and then the resolution telescopes to a coarser resolution from 30°S to the equator. The vertical z-coordinate grid has 52 layers with varied thickness from about 4 m at the surface to 400 m at depth. The bathymetry is derived from ETOP01 (Amante and Eakins 2009). For this work we consider an analysis domain spanning 64.8°–30.4°S,

and subsample the model on a Mercator grid with 2° resolution in longitude and approximately 1.1° in latitude. The sample spacing ranges from 96 km at 64.8°S to 190 km at 30.4°S.

The biogeochemical component, adapted from the original BLING model (Galbraith et al. 2010), includes nitrogen cycling and phytoplankton dynamics (Verdy and Mazloff 2017). Evolutions and interactions of eight prognostic tracers [i.e., inorganic/organic forms of nitrogen and phosphorus, dissolved inorganic carbon (DIC), alkalinity, oxygen, and iron] are calculated in the model, representing important biogeochemical processes, such as the conversion between DIC and organic matters, phytoplankton evolution, and net community production (Verdy and Mazloff 2017; Rosso et al. 2017).

A number of datasets are utilized to initiate and to force the biogeochemical GCM. The atmospheric state is obtained from ERA-Interim products (Dee et al. 2011). The initial biogeochemical tracer fields are derived from the Global Ocean Data Analysis Project, version 2 (GLODAPv2), climatology (Lauvset et al. 2016; Key et al. 2015); the *World Ocean Atlas 2013* climatologies (Garcia et al. 2013a,b); and a coupled model simulation with BLING, version 2 (E. Galbraith 2013, personal communication). The river and Antarctic freshwater discharge are derived from continental freshwater products of Dai and Trenberth (2002) and Hammond and Jones (2016). The model is run for 130 years with a time step of 1 h by looping the 2005–14 forcing conditions. The N-S-T fields in the latter 60-yr period (i.e., model years 71–130) are used in our analyses. Monthly averaged fields are output for diagnostics. We use only the N-S-T fields at 100-m depth, which is approximately the average depth of nutrocline in the SO (not shown). The simulated N-S-T fields are used as the reference N-S-T fields in the following analyses. The N-S-T anomaly fields are calculated by subtracting the monthly mean fields over the 60-yr analysis period, thus representing the departure from the seasonal cycle.

b. MEOF method

In this study we adopt the MEOF approach to construct SO nitrate maps and consider sets of synthetic N-S-T data in order to evaluate the method and sampling strategy. Here we briefly summarize the procedures of the MEOF calculation. First, the N-S-T anomaly fields are divided by their total (spatial and temporal) standard deviations [$\sigma_N = 0.001\,07$ (mol N^2 m^{-2}), $\sigma_S = 0.0984$ (psu), $\sigma_T = 0.405$ (°C)] and transformed into a data matrix $\mathbf{X}$ of the form

$$\begin{bmatrix} N_{1,1} & N_{1,2} & \cdots & \cdots & N_{1,n} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ N_{m,1} & N_{m,2} & \cdots & \cdots & N_{m,n} \\ S_{1,1} & S_{1,2} & \cdots & \cdots & S_{1,n} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ S_{m,1} & S_{m,2} & \cdots & \cdots & S_{m,n} \\ T_{1,1} & T_{1,2} & \cdots & \cdots & T_{1,n} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ T_{m,1} & T_{m,2} & \cdots & \cdots & T_{m,n} \end{bmatrix},$$

where m represents the grid points of the reference N-S-T fields with land points cropped; n is the total time steps; and the superscript T denotes transpose of the matrix. Then we perform the singular value decomposition (SVD) method on $\mathbf{X}$ to isolate the spatial and temporal MEOF information in $\mathbf{U}$ and $\mathbf{V}$

$$\mathbf{UDV}^{\mathrm{T}} = \mathbf{X}, \tag{1}$$

where $\mathbf{D}$ is a diagonal matrix, in which the diagonal elements (i.e., D_{ii}) represent the eigenvalues according to the rank in amplitude (from largest to smallest). We determine the MEOF mode (spatial pattern) and its principal component (PC; time information) from the vector components in $\mathbf{U}$ and $\mathbf{V}$, respectively, according to the amplitude of their corresponding eigenvalues The percentage of variance accounted for by the ith MEOF mode is calculated as

$$D_{ii}^2/\mathrm{trace}(\mathbf{DD}^{\mathrm{T}}) \times 100\%, \tag{2}$$

where trace($\mathbf{DD}^{\mathrm{T}}$) is the sum of all diagonal elements in $\mathbf{DD}^{\mathrm{T}}$. As we have 720 time records, we obtain 720 MEOF modes. We examine the spatial characteristics of the leading MEOF modes in section 3. A schematic chart to clarify the details of the MEOF calculation is shown in Fig. 1 (step 1).

c. k-means cluster method

The k-means cluster method is designed to partition one or multiple datasets into k clusters in which each datum is assigned to a certain cluster according to the nearest mean (Hartigan and Wong 1979). Formally, its algorithm aims to minimize the sum of squared distances between the data and the mean within each cluster (Forgy 1965; MacQueen 1967; Hartigan and Wong 1979), which can be formulated as

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{S}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2, \tag{3}$$

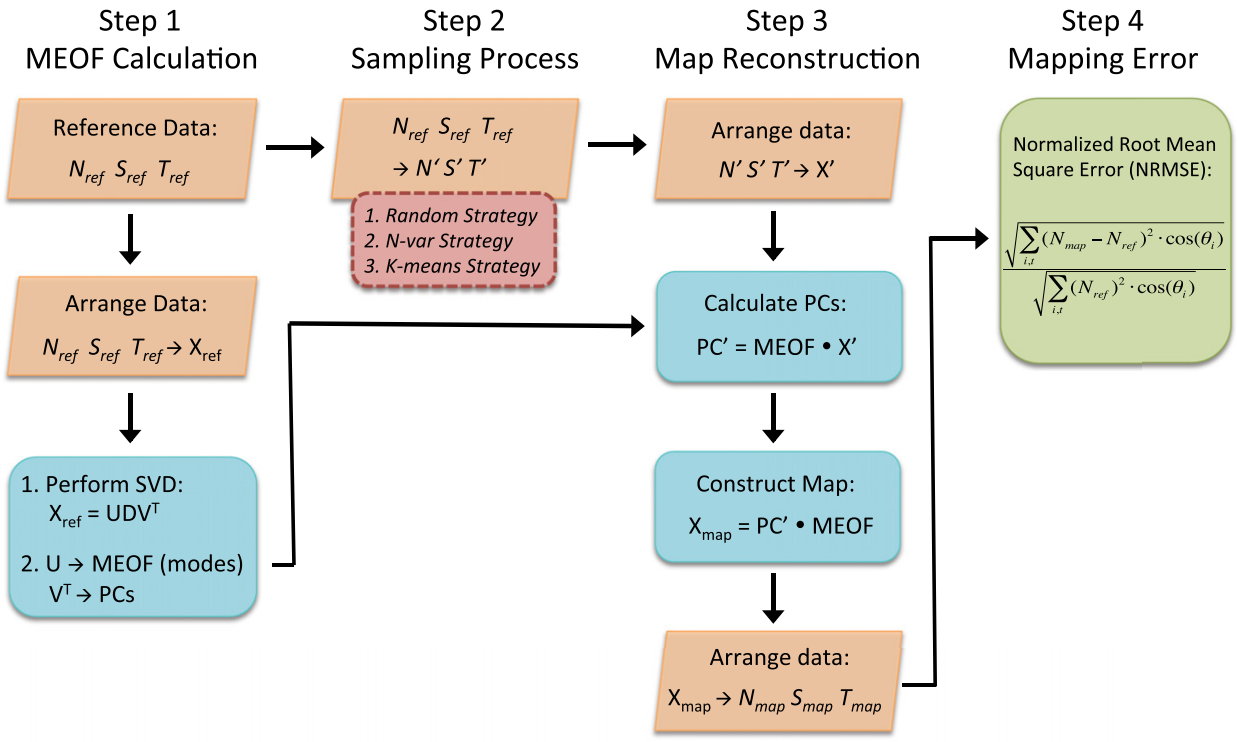

FIG. 1. Schematic of the MEOF mode calculation, sampling processes, and map construction used in this study.

where $\mathbf{x}$ is the data and $\boldsymbol{\mu}_i$ is the mean of data in cluster $\boldsymbol{S}_i$. Because of its ease of implementation and relatively smaller computation and storage costs compared to other clustering methods (Hartigan and Wong 1979; Firdaus and Uddin 2015), the *k*-means cluster method has served as a prototype of unsupervised learning algorithms and has been successfully applied to many problems associated with categorization or regression (Shirkhorshidi et al. 2014). In marine biogeochemical studies, it has been used to explore common features or relationships between different fields in regional and global oceans (e.g., D'Ortenzio and Ribera d'Alcalà 2009; D'Ortenzio et al. 2012; Lacour et al. 2015; Mayot et al. 2016; Ardyna et al. 2017). For example, Ardyna et al. (2017) used the *k*-means cluster method on satellite-derived chlorophyll-*a* concentration data to define bioregions in the SO, each of which contains unique biogeochemical phenology.

In this study we use the *k*-means cluster from a Python machine learning package, called scikit-learn, v0.19.0 (Pedregosa et al. 2011, also see details on the scikit-learn official website: http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html). We take log with base 10 on the N-S-T variances (see Figs. 2d–f) and organize them into an array [i.e., $\mathbf{x}$ in Eq. (3)] with size $N \times M$, where N is the grid size of the SO domain (4857) and M represents the number of variables (three: nitrate, salinity, and temperature). The array is then clustered with an agglomerative hierarchical clustering model (Hartigan and Wong 1979). We set the cluster number as five, meaning five clusters or groups are determined based on the N-S-T variances. The model is iterated until it reaches a convergence criterion of relative tolerance less than 0.0001 [defined with regard to the magnitude calculated in Eq. (3)], or 300 maximum iteration steps. The *k*-means algorithm is performed 10 times, and the best result in terms of smallest relative tolerance is selected as the final clusters. The resulting five clusters divide the SO domain into five subregions, which are used to determine the *k*-means sampling strategy in later analyses and are discussed in section 3.

After five SO subregions are determined by the *k*-means cluster method, we perform the Kruskal–Wallis *H* test (Kruskal and Wallis 1952), with a null hypothesis that assumes the medians of each group are the same, to inform whether these regions are significantly different in their N-S-T mean variance fields. Significant results are found, as all p values are far less than 0.0001, indicating at least one region differs from all others in terms of N-S-T mean variances. The mean variances for each region are summarized in Table 1.

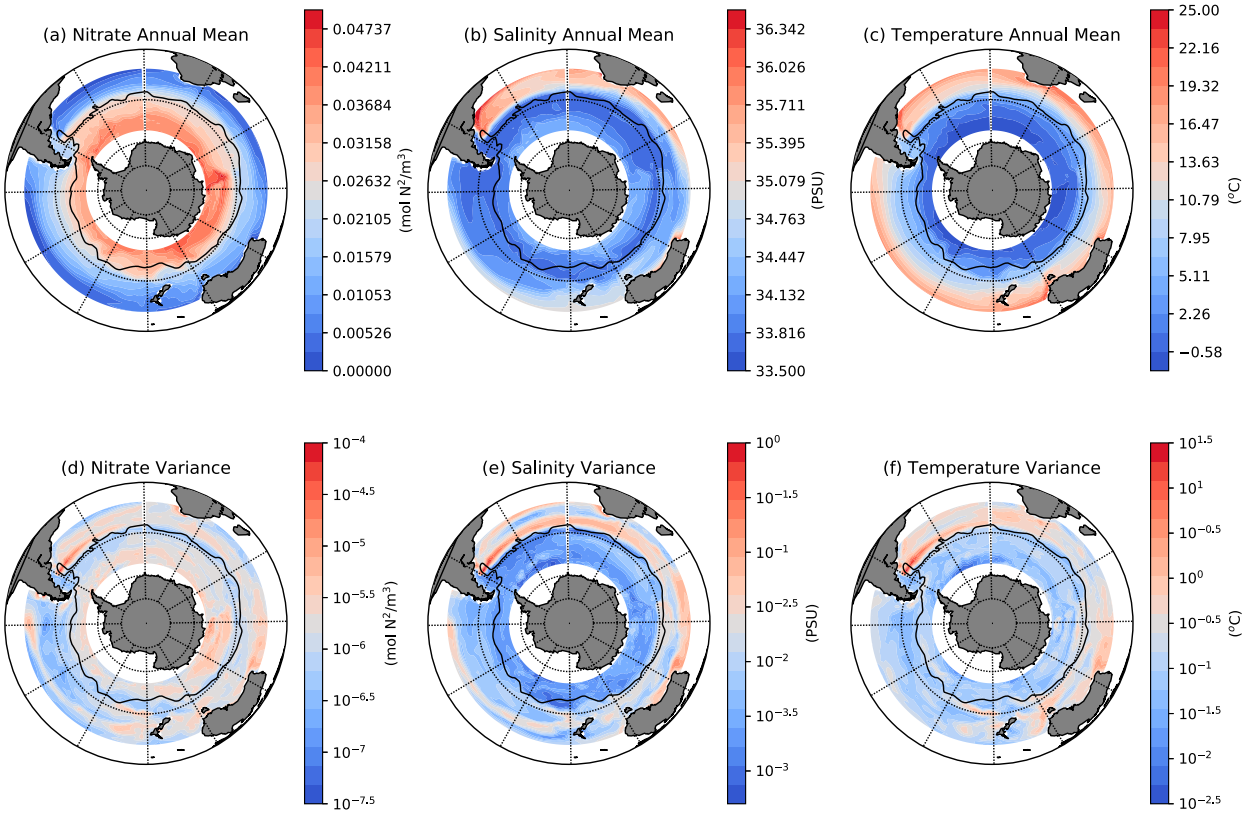

FIG. 2. The maps of the mean (a) nitrate, (b) salinity, and (c) potential temperature fields at 100 m from the biogeochemical GCM. (d)–(f) As in (a)–(c), but for variance in $\log_{10}$ scale. In each panel the black curve circling low-latitude oceans represents the Subantarctic Front according to Orsi et al. 1995.

3. Construction of SO nitrate maps

a. MEOF analysis

We perform the MEOF analysis on the reference N-S-T fields simulated by the biogeochemical GCM over the 60-yr period (i.e., model years 71–130; see section 2a for details) to construct SO nitrate maps. We first investigate the mean and variance characteristics of the reference N-S-T fields. Figures 2a–2c show the N-S-T mean values in the SO. The mean nitrate field is characterized by an evident north–south structure with high values located south of the Subantarctic Front (the black curve in Fig. 2) and low values to the north. Similar geographical features can be seen in the annual mean salinity and temperature fields but with low and high values reversed with latitudes (Figs. 2b,c). The spatial correspondences between mean N-S-T fields in the SO reflect the fact that they are largely determined by similar large-scale physical processes. It is also noted that their meridional gradients (i.e., north–south

TABLE 1. Geographic information, N-S-T variances, and NRMSE reduction rate of each K-region.

	K1	K2	K3	K4	K5	Random selection
Grid No.[a]	430	987	1337	1135	968	—
Area (km^2)	12 264 539	27 441 742	18 734 790	15 978 503	23 490 489	—
N mean variance ($mol\,N^2\,m^{-3}$)	3.89×10^{-6}	1.53×10^{-6}	2.61×10^{-6}	1.32×10^{-6}	6.51×10^{-7}	—
S mean variance (psu)	7.21×10^{-2}	2.02×10^{-2}	3.28×10^{-3}	1.83×10^{-3}	7.02×10^{-3}	—
T mean variance (°C)	1.11	3.08×10^{-1}	1.29×10^{-1}	6.12×10^{-1}	1.05×10^{-1}	—
N NRMSE reduction rate	5.28×10^{-4}	2.19×10^{-4}	1.71×10^{-4}	8.89×10^{-5}	8.89×10^{-5}	1.82×10^{-4}
S NRMSE reduction rate	8.21×10^{-4}	2.42×10^{-4}	9.99×10^{-5}	5.69×10^{-5}	8.55×10^{-5}	1.98×10^{-4}
T NRMSE reduction rate	7.89×10^{-4}	2.23×10^{-4}	1.09×10^{-4}	6.00×10^{-5}	7.98×10^{-5}	1.99×10^{-4}

[a] One grid cell covers approximately 2° (longitude) × 1.1° (latitude) ≈ 27 192 km^2.

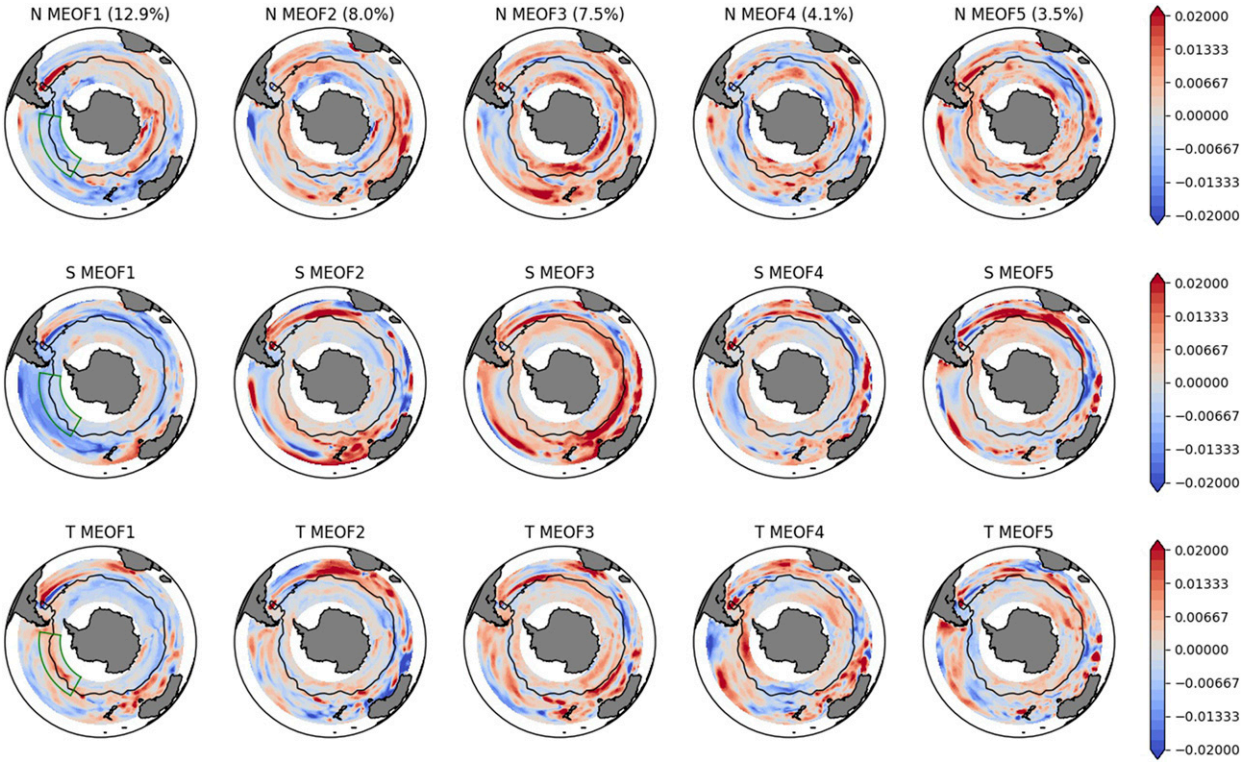

FIG. 3. The first five leading MEOF modes. (top to bottom) The maps are the spatial patterns of the MEOF modes associated with (top) nitrate, (middle) salinity, and (bottom) temperature fields. The green boxes (65°–50°S, 150°–80°W) in the first column denote the Bellingshausen–Amundsen Sea regions. The black curve is as in Fig. 2.

changes) are not uniform throughout the latitudes, but they appear sharper in the 50°–60°S latitude band that approximately coincides with the Subantarctic Front (the black curve in Figs. 2a–c). Comparisons with these modeled mean features and the N-S-T mean fields shown in recent studies (e.g., Verdy and Mazloff 2017; Rosso et al. 2017) indicate that the biogeochemical GCM used in this study reasonably captures the distinctive large-scale N-S-T features in the SO.

The large-scale similarities can also be found in the N-S-T variance maps (in $\log_{10}$ scale; Figs. 2d,e). Particularly high nitrate variances collocate with high salinity and temperature variances at the confluence of the Brazil and Malvinas Currents off the Argentine coast and its downstream areas (Figs. 2d–f). These collocations imply parts of the high N-S-T variances are sourced from the fluctuations and instabilities of the Subantarctic Front (black curve in Figs. 2d–f), which was also reported by a recent study (Ferrari et al. 2017). However, mismatches of the N-S-T variance structure appear in high latitudes, poleward of approximately 60°S, where patterns of nitrate variance do not closely resemble the patterns of salinity and temperature variances. The results imply that the ocean dynamics explains a large portion of the N-S-T variability in the latitudes near the Subantarctic Front, but differences either in background gradients or as a result of biogeochemical processes are significant in the subpolar regions.

We next perform the MEOF analysis over the reference N-S-T fields and obtain 720 MEOF modes (see section 2b and Fig. 1, step 1, for details) for the nitrate map construction. Figure 3 demonstrates the spatial patterns of the leading five MEOF modes, which combine to explain 36% of the total N-S-T combined variance. The N-S-T patterns of the first MEOF mode capture important features and show resemblances with each other at low latitudes (see the first column panels in Fig. 3), particularly at the confluence of the Brazil and Malvinas Currents, and in the downstream regions where the positive nitrate anomalies and out-of-phase salinity and temperature anomalies are collocated. In contrast, the N-S-T patterns do not resemble each other in higher latitudes. In the Bellingshausen–Amundsen Sea regions (green boxes in the first column panels in Fig. 3) the anomaly pattern of strong positive temperature signals differs from those of moderate negative nitrate and salinity patterns. These latitude-dependent

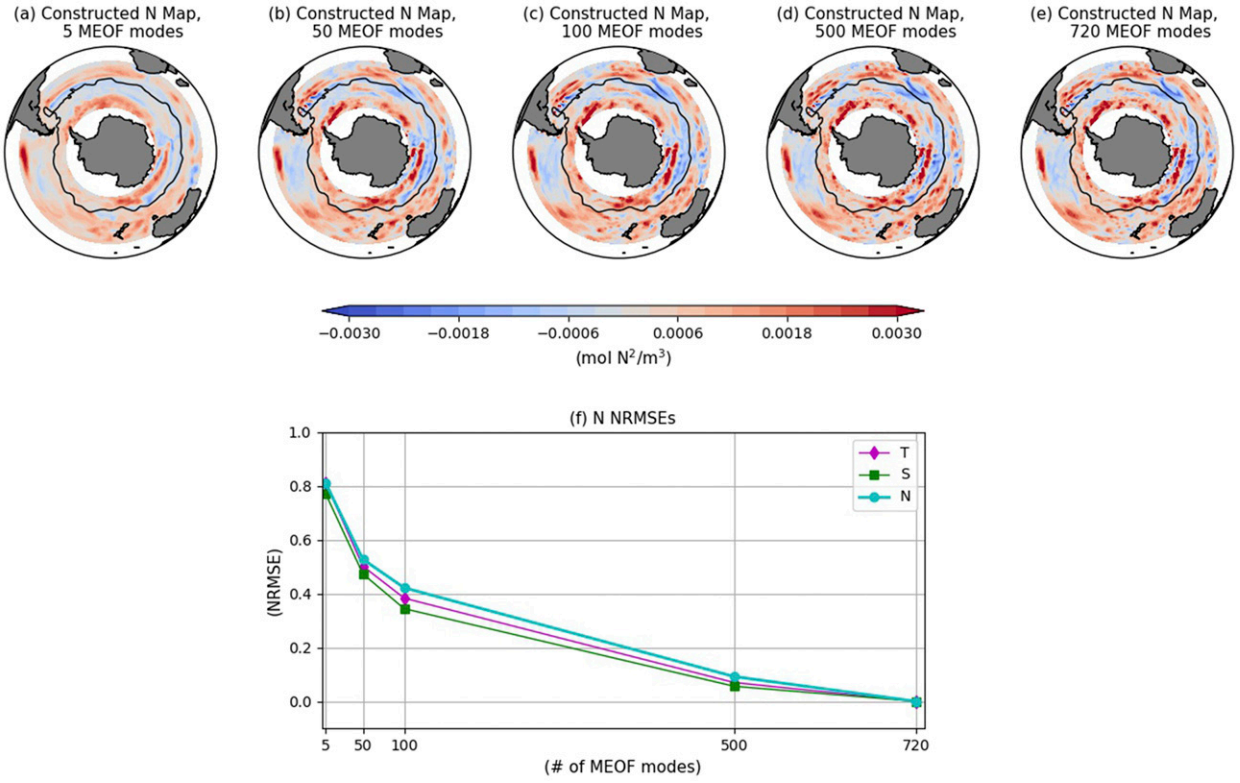

FIG. 4. The constructed nitrate anomaly maps using the first (a) 5, (b) 50, (c) 100, (d) 500, and (e) 720 MEOF modes in August of a random model year. The black curve is as in Fig. 2. (f) The NRMSEs between each constructed nitrate map and the reference nitrate anomaly fields (cyan line). Also shown in (f) are the NRMSEs for the salinity (green line) and temperature fields (magenta line).

similarities and differences seem to be the general features of the other four MEOF modes (shown in the second to fifth column panels in Fig. 3) and other lower-order MEOF modes (not shown).

To evaluate the relationships between the number of MEOF modes and the capability to reproduce the reference nitrate anomaly field, we show the snapshots of the mapped nitrate anomalies for one random August in the biogeochemical GCM simulation from using the first five MEOF modes to using total 720 modes (Figs. 4a–e). The more MEOF modes are used, the more detailed features of nitrate anomalies are manifested. Taking the anomalies at the Brazil–Malvinas confluence and its downstream regions as an example, when we use only five MEOF modes, the reproduced nitrate anomalies show two parallel, out-of-phase anomaly bands extending from the coastal region into the South Atlantic (Fig. 4a), whereas the meandering structures of nitrate anomalies become evident when using more MEOF modes (Figs. 4b–e).

The capability of capturing the details can be quantified by examining the normalized root-mean-square errors (NRMSEs) between the mapped nitrate anomaly field (N_{map}) and the reference nitrate field (N_{ref}). The NRMSE for nitrate maps is defined as

$$\text{NRMSE} = \frac{\sqrt{\sum_{t,i}(N_{\text{map}} - N_{\text{ref}})^2 \times \cos(\theta_i)}}{\sqrt{\sum_{t,i}(N_{\text{ref}})^2 \times \cos(\theta_i)}}, \qquad (4)$$

where t, i indicate the time step and the spatial grid point, respectively; and θ_i is the latitude at grid i (rad). Figure 4f shows the nitrate NRMSEs with an increasing number of MEOF modes used in recovering the reference nitrate anomaly field (cyan line). The first five MEOF modes result in about 0.81 NRMSE (36% variance explained), which is greatly reduced to about 0.092 (99% variance explained) using 500 modes. A similar reduction of NRMSEs can also be found for recovering salinity and temperature fields (green and magenta lines, respectively, in Fig. 4f). We chose to use 500 MEOF modes for the following nitrate map construction, as this number is sufficient to reduce the N-S-T NRMSEs to less than 0.1 and to explain more than 99% of the total N-S-T combined variance.

b. Sampling strategies and nitrate map construction

The nitrate variance is highly heterogeneous in the SO domain as shown in Fig. 2d. Likewise, the spatial structures of variability are not isotropic as can be seen from the MEOF modes in Figs. 3 and 4. It follows that observations in different locations have varying levels of information content. To investigate this hypothesis, we create synthetic N-S-T datasets according to three sampling strategies: 1) random selection, 2) certain thresholds of nitrate variance (N-var strategy), and 3) the *k*-means cluster method (*k*-means strategy). The sampling strategies are listed in Fig. 1 (step 2) for clarification. We apply the MEOF analysis to these synthetic data to reconstruct SO nitrate maps and to compare the maps to the reference fields to assess these sampling strategies (see Fig. 1, steps 3 and 4).

Figure 5 shows five SO subregions with color markings determined by the *k*-means cluster method (see section 2c for cluster details). The K1 region in red, containing 430 grid points ($\sim$12 264 539 km^2), largely coincides with high N-S-T variance regions that cover the northern areas of the Subantarctic Front extending from the coast of Argentina toward the coast of South Africa and regions surrounding the Australian continent. The K2 region in dark yellow, containing 987 grid points ($\sim$27 441 742 km^2), covers low-latitude regions outside the K1 area. The K3 and K4 regions in green and light blue, respectively, containing 1337 and 1135 grid points, respectively ($\sim$18 734 790 km^2 and $\sim$15 978 503 km^2), together encompass almost all low N-S-T variance regions poleward of the Subantarctic Front (inside the black curve in Fig. 5). The K5 region in blue, containing 968 grid points ($\sim$23 490 489 km^2), covers wide areas in the South Pacific where relatively low N-S-T variance regions are located. The five SO subregions from the *k*-means cluster method determines the *k*-means strategy to assess the importance of sampling locations in map construction. The grid number, area coverage, and mean N-S-T variances of each K region are summarized in Table 1.

We determine the random selection strategy and N-var strategy with certain threshold values of the nitrate variance that give rise to the same number of grid points as each of the K1–5 regions. For example, when comparing strategies to the K1 region (Fig. 6c), we randomly select 430 grid points over the SO domain (Fig. 6a). We also determined that there are 430 grid points with nitrate variance larger than 3.2436×10^{-6} (mol N^2 m^{-2}) and used this as our threshold to determine the N-var strategy (Fig. 6b). Thus, our three sampling strategies all have an equal amount of synthetic N-S-T data. It is noted that we assume the synthesized N-S-T datasets are perfectly sampled in each grid over a 60-yr period; in other words, no gaps in time are considered.

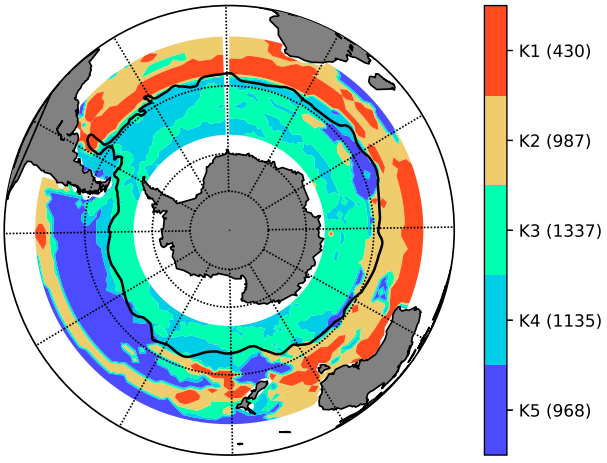

FIG. 5. Color shading denotes the five subregions of the SO domain determined by the *k*-means cluster method. The number of mapping grid points included in each subregion is denoted in the parentheses next to the color bar. The black curve is as in Fig. 2.

The random selection strategy, as expected, inserts data with no specific spatial preferences (Fig. 6a). The inserted data are spread throughout the region without a specific structure. On the contrary, the 430 N-S-T data are distributed in a structured manner using the N-var and *k*-means strategies (Figs. 6b,c). The N-var sampling strategy exhibits data grouping in the Davis Sea–Dumont d'Urville Sea regions (red box in Fig. 6b), at the Brazil–Malvinas confluence and its downstream region (magenta box in Fig. 6b), and in the southeastern Indian Ocean (black box in Fig. 6b). Similar data placements are adopted by the *k*-means strategy, but the inserted data concentrates in low-latitude regions (magenta and black boxes in Fig. 6c) rather than in high-latitude Davis Sea–Dumont d'Urville Sea regions (red box in Fig. 6c). The N-S-T data are inserted in one band extending from the Brazil–Malvinas confluence to South Africa, coinciding with the region north of the Subantarctic Front in the South Atlantic sector, and in another band extending from the southeastern region of the Indian Ocean toward the Australian west coast.

We project the 500 MEOF modes onto the synthetic N-S-T datasets based on the three sampling strategies in order to estimate the PCs that represent the temporal variations of each MEOF mode (PC′ in Fig. 1, step 3). We treat each MEOF mode equally without performing any weighting to consider only the spatial heterogeneity for each mode. Thus, we are assuming the primary spatial modes of variability of the N-S-T fields are known, and we are using the partially sampled N-S-T anomaly fields to derive their time variability. The

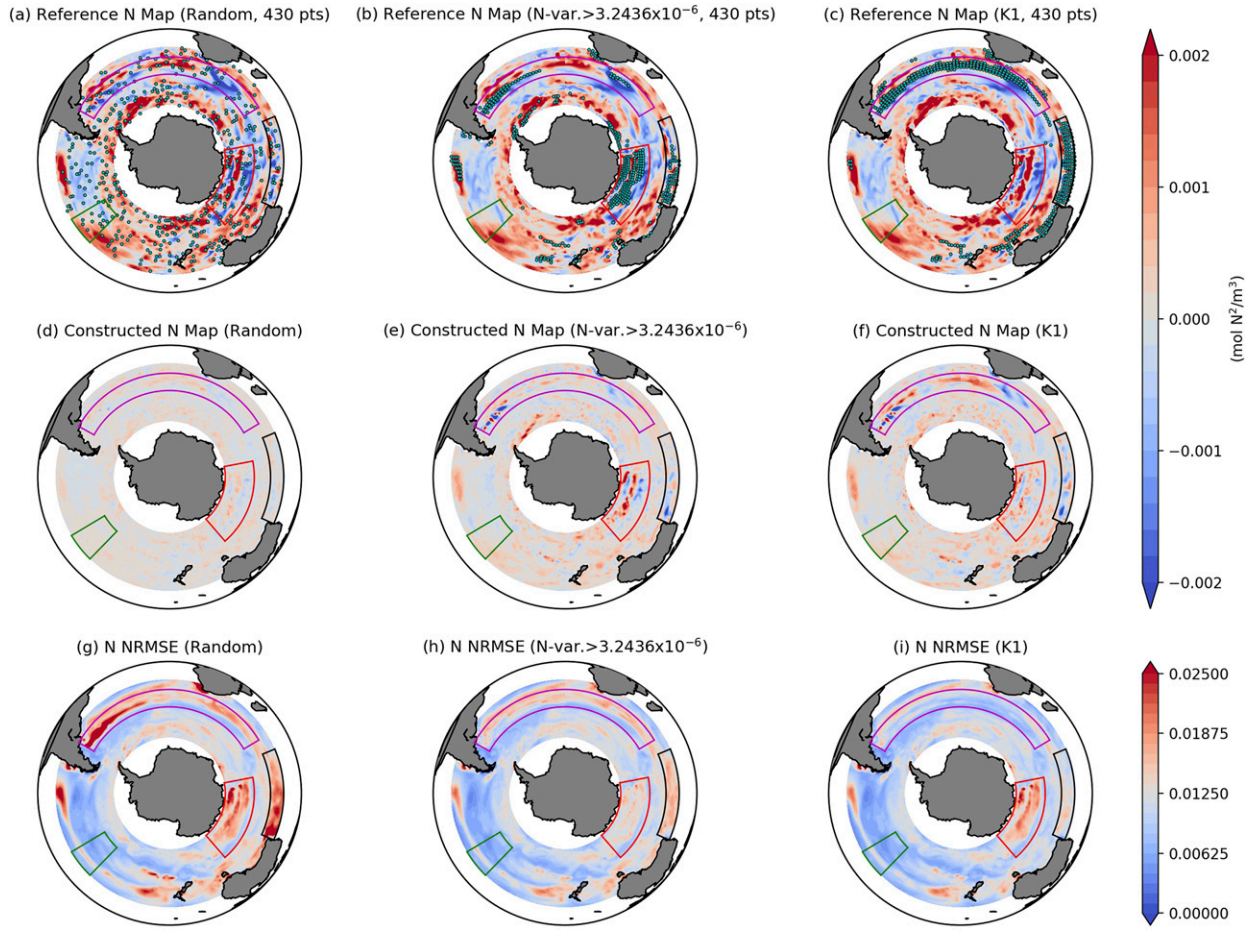

FIG. 6. The cyan dots denote the locations where the reference N-S-T fields are sampled according to (a) random selection, (b) N-var, and (c) *k*-means strategies. The color shadings represent the nitrate anomaly field in August of a random model year. (d)–(f) The constructed nitrate maps for that August. (g)–(i) Maps showing the spatial distribution of nitrate NRMSEs. In each panel the magenta box denotes the confluence of the Brazil and Malvinas Currents and downstream into the South Atlantic and Indian sector regions (60°W–60°E and 49°–38°S); the red box denotes the Davis Sea–Dumont d'Urville Sea regions (80°–140°E and 65°–50°S); the black box denotes the southeastern region of the Indian Ocean (68°–115°E and 40°–30°S); and the green box denotes a South Pacific region (135°–120°W and 55°–30°S) where no sampling data are added by the *k*-means and N-var strategies.

nitrate maps are then constructed by taking the dot product of the estimated PCs and the MEOF modes (see X_{map} in Fig. 1, step 3).

Figure 6d shows the constructed nitrate anomaly map by random selection strategy, which captures little of the structure of the reference nitrate anomaly field. The NRMSEs [calculated following Eq. (4) but at each grid point] are large (Fig. 6g), particularly in the Brazil–Malvinas confluence and downstream regions (magenta box), the Davis Sea–Dumont d'Urville Sea regions (red box), and the southeastern region of the Indian Ocean (black box). The N-var strategy, on the other hand, recovers some details of the reference nitrate field with comparable magnitudes (Fig. 6e) and greatly reduces nitrate NRMSEs (Fig. 6h) in the regions where data are inserted. These regions are again the Brazil–Malvinas confluence and downstream region, the Davis Sea–Dumont d'Urville Sea regions, and the southeastern Indian Ocean, coinciding with where large NRMSEs appear using the random selection strategy. The *k*-means strategy also shows skill in recovering the reference nitrate anomaly field (Fig. 6f), significantly reducing NRMSEs (Fig. 6i) in the South Atlantic and Indian Ocean sectors in particular, but leaves relatively large NRMSEs in high-latitude Davis Sea–Dumont d'Urville Sea regions where the K1 region does not sample (red box in Fig. 6i). However, overall, the *k*-means strategy has the smallest NRMSE with a value of 0.77, while the N-var strategy gives 0.79 and the random sampling strategy gives 0.92.

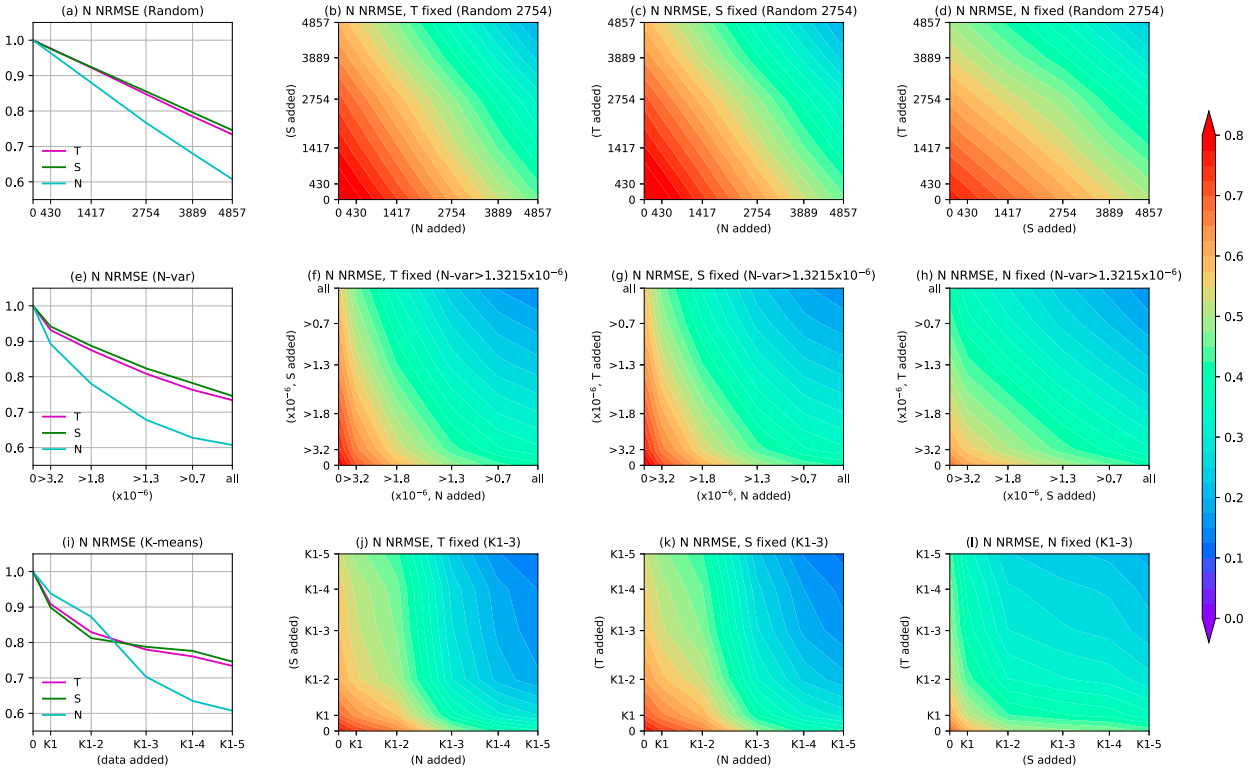

FIG. 7. (a) The nitrate NRMSEs with only temperature (magenta line), salinity (green line), and nitrate (cyan line) data added according to a random selection strategy. (e),(i) As in (a), but according to the N-var and *k*-means strategies, respectively. (b)–(d) The nitrate NRMSEs with fixed amounts of temperature, salinity, and nitrate added, while varied amounts of the other two fields are inserted according to the random selection strategy. (f)–(h),(j)–(l) As in (b)–(d), but according to the N-var and *k*-means strategies, respectively.

We further investigate the capability of the *k*-means and N-var strategies in reconstructing SO nitrate maps in regions where no data are inserted. We calculate the area-averaged NRMSEs over the region marked by the green box in Fig. 6 and find that smaller NRMSEs use the *k*-means strategy (0.086) and the N-var strategy (0.090) than those that use the random sampling strategy (0.095), which has 16 data samples in the region. These findings not only show the mapping strength of the MEOF method that uses sampling data to reconstruct nitrate maps outside sampling areas but also reveal that an organized sampling strategy can better recover the reference nitrate field in regions where no data are inserted than random selection strategy.

To systematically evaluate the nitrate mapping errors and their relationships with adding various amounts of N-S-T fields, we perform a series of nitrate NRMSE analyses (Fig. 7). We first consider adding data to regions following the three sampling strategies but sampling only one field and calculating the corresponding nitrate NRMSEs (Figs. 7a,e,i). Again, the same amount of data is sampled in each set of error calculations for the three sampling strategies to make fair comparisons. In Fig. 7a, adding nitrate data (cyan line) gives rise to smaller nitrate NRMSEs than adding the other two fields (magenta and green lines) as expected. The results also reflect that even with no nitrate data added, the nitrate mapping errors can be reduced by adding temperature or salinity data only because of the capability of the MEOF method to recover one field with information from other fields. We also notice that the nitrate NRMSEs are reduced linearly with an increasing number of data added, implying that the randomness in the sampling processes can be transformed as linear mapping error reduction of the MEOF analyses. However, the linear behavior of nitrate mapping errors disappears when we adopt the N-var strategy (Fig. 7e). NRMSEs are reduced faster at first and then asymptote as more data are inserted. Different from the other sampling strategies, the *k*-means strategy exhibits larger NRMSEs, inserting only nitrate data than only salinity or temperature data into K1 or K1–2 regions, but it results in smaller NRMSEs when adding more regions (Fig. 7i). Such behavior of nitrate mapping errors is possibly caused by the *k*-means cluster method aiming to capture the greatest variance overall, which does not

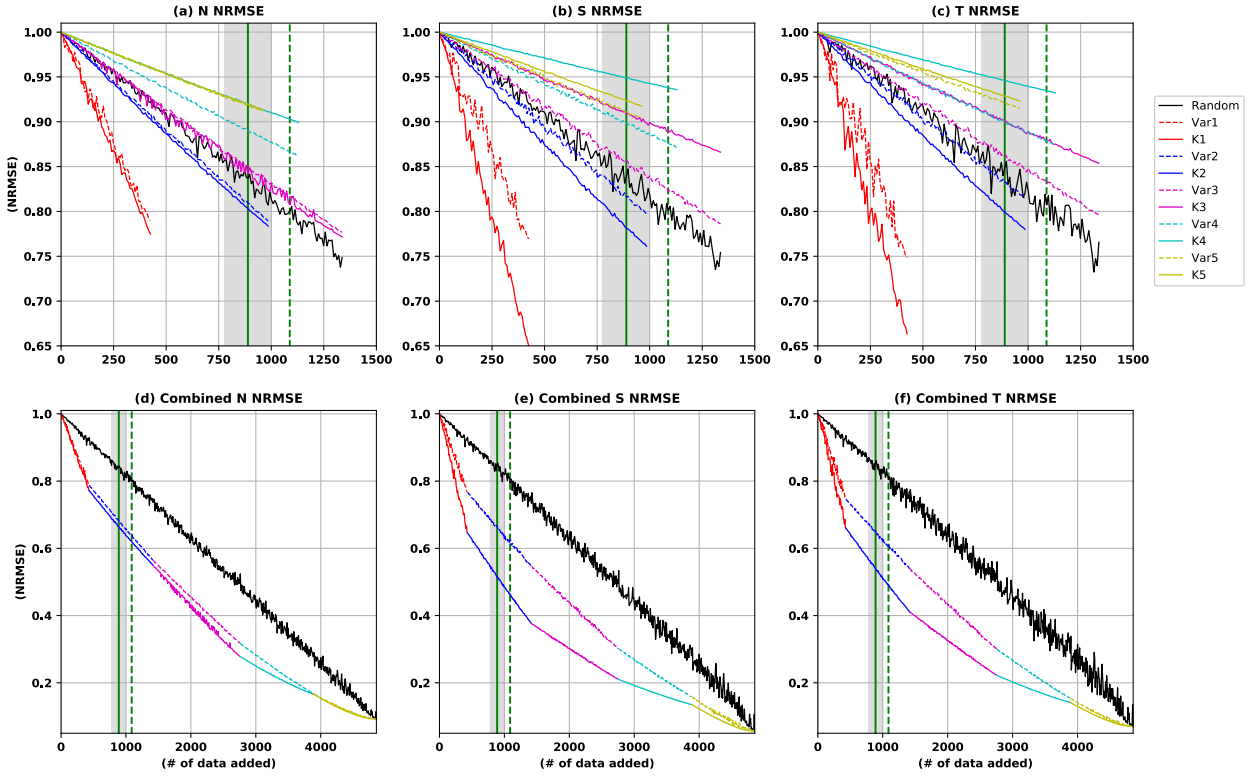

FIG. 8. (a) Nitrate NRMSEs against the number of data added in individual regions determined by the *k*-means (solid lines), N-var (dashed lines), and random selection (black line) strategies. (b),(c) As in (a), but for salinity and temperature NRMSEs, respectively. (d)–(f) As in (a)–(c), but data are cumulatively added over the different regions. All panels show the average Argo float numbers deployed in the SO for 2008–15 (solid green line), with the shading showing the standard deviation; and the number of Argo floats needed to meet the stated goal of one float every 300 km × 300 km (dashed green vertical lines).

necessarily lead to the greatest constraints on the MEOF method when one has only a single data type. These results show that both the *k*-means and N-var strategies perform better in reducing mapping errors than random selection strategy.

Further insights can be gained by generalizing the results in Figs. 7a, 7e, and 7i by coloring NRMSEs calculated by holding the amount of a certain variable constant and varying the other two. Figure 7j, for example, shows the nitrate NRMSE changes with differing amounts of nitrate (signified on the *x* axis) and salinity (signified on the *y* axis) data inserted throughout the K1–5 regions when a fixed amount of temperature data (2754 data points) are inserted into the K1–3 regions. Comparing the NRMSEs in Fig. 7 verifies that the *k*-means and N-var strategies result in smaller NRMSEs and faster error reduction rates relative to random selection strategies. We also find that adding salinity and temperature data tends to reduce nitrate mapping errors more when using the *k*-means strategy than for the other strategies (cf. Figs. 7d,h,l). This reflects that the *k*-means cluster method adopts a sampling strategy that also incorporates the salinity and temperature information to reduce NRMSEs. The abovementioned mapping error analyses indicate that the *k*-means and N-var strategies outperform the random selection strategy, while the *k*-means strategy better utilizes the salinity and temperature information to reduce nitrate mapping errors.

We next examine the relative importance of inserting all N-S-T data in regions determined by the three sampling strategies into the reconstructed maps. The finding that adding data randomly results in an approximately linear NRMSE reduction rate (Figs. 7a–d) implies that we can use this rate (or slope) with data added randomly in one region as a first-order measure for its relative importance. In other words, the faster the error reduction rate (or steeper slope) is in one region with data randomly added, the more important the region is. As such, we add N-S-T data randomly to the K1–5 regions individually and show the corresponding nitrate NRMSEs against the number of added data (Fig. 8a). When adding 0–430 data to the K1 region, we find that the NRMSEs drop fast from 1.0 to about 0.77 (the red solid

line in Fig. 8a). However, adding 430 data to the K2 region drops NRMSE to only about 0.91 (the blue solid line in Fig. 8a). Their different slopes also reveal that in order to obtain the same effect of reducing nitrate NRMSE to about 0.77, approximately 987 data need to be added to the K2 region, but only about 430 data to the K1 region. This result clearly shows the error reduction rate in the K2 region is slower than in the K1 region, and it indicates the K2 region is less important than the K1 region for constraining the mapping. Likewise, as the error reduction rates are much slower for the K3, K4, and K5 regions (magenta, cyan, and dark yellow solid lines in Fig. 8a), they also play less important roles than the K1 and K2 regions in reducing mapping errors. The approximated NRMSE reduction rates in each K-region for the N-S-T fields are summarized in Table 1.

Although adding data to regions determined by the N-var strategy gives rise to similar NRMSE reduction rates as determined by the *k*-means strategy for nitrate field (solid and dashed lines in Fig. 8a), discrepancies become evident for the rates associated with salinity and temperature fields (solid and dashed lines in Figs. 8b,c). The *k*-means strategy outperforms the N-var strategy in reducing salinity and temperature mapping errors, particularly in the high nitrate variance regions (i.e., K1–2 and Var1–2 regions). The results support those shown in Fig. 7 and further indicates that when targeting high nitrate variance regions, the *k*-means strategy not only maintains similar performance as the N-var strategy in reducing nitrate mapping errors but also better reconstructs salinity and temperature maps.

To further examine whether the mapping error reduction rate of one K-region is affected by the data inserted into other regions (i.e., the independence between each K-region), we add N-S-T data sequentially to the K1–K5 regions and calculate NRMSEs (Figs. 8d–f). That is, we randomly add data to the K1 region until 430 data fill it and then randomly add data to the K2 region until 987 data fill it, and so on. We find the mapping error reduction rates of one region do not change when N-S-T data have been added in other regions. For example, the reduction rate of nitrate mapping error added in the K2 region individually (i.e., the solid blue line in Fig. 8a) does not change when the K1 region has first been filled up with 430 data (i.e., the solid blue line segment in Fig. 8d). This result confirms that the effects of inserting data into different regions to reduce NRMSEs are independent of each other. Such independence can also be found in the N-var sampling strategy (dashed lines in Figs. 8d–f). However, the NRMSEs are higher in magnitude and drop slower than those associated with the *k*-means strategy, particularly in high-variance regions. These results are consistent with those shown in Figs. 8a–c.

The abovementioned examinations of N-S-T mapping error reduction rates within different regions determined by the sampling strategies quantify mapping error reduction via adding data and informing the significance of sampling density and distribution. The results highlight that having greater data density in high-variance regions (e.g., the K1 and K2 regions) is more important than in low-variance regions for constructing nitrate (and salinity and temperature) maps.

4. Summary and discussion

This study employs the MEOF method to construct SO nitrate maps using N-S-T combined datasets from a state-of-the-art biogechemical GCM. An assessment of the MEOF method skill in estimating the reference modeled nitrate field suggests that using the first 500 MEOF modes sufficiently recovers 99% of the model signal. To assess an optimal way to sample N-S-T data in the SO, we create synthetic N-S-T datasets via sampling reference N-S-T anomaly fields in regions determined by either random selection, a certain threshold of nitrate variance, or a *k*-means cluster method. The first 500 MEOF modes are then projected onto these synthetic N-S-T datasets to construct the SO nitrate maps. The skill in the constructed maps is systematically examined with a series of error analyses. The examination of mapping error reduction rates of each region determined by these sampling strategies reveals their relative importance and addresses the significance of sampling data density within each region. The results conclude that sampling strategies considering nitrate variance structure yield more mapping skill than unstructured random selection strategy. The *k*-means strategy further utilizes the salinity and temperature information to reduce the mapping errors. The MEOF method together with the *k*-means sampling strategy suggests a framework to prioritize deployment of in situ measurements, such as biogeochemical Argo floats, to sample important nitrate spatiotemporal variations in the SO and to construct accurate nitrate maps.

Our findings indicate that the *k*-means strategy gives rise to better N-S-T maps (Figs. 8d–f), but it does not introduce a fundamentally different sampling strategy from that considering solely nitrate variance. To elaborate this point, we perform the *k*-means cluster method on a nitrate variance field only and obtain similar grouping regions (not shown) as those determined by the N-var strategy. This implies that the *k*-means strategy will perform similarly as the N-var strategy if considering only the nitrate variance. Thus, our findings

emphasize the significance of utilizing as much information as is available, which here includes the N-S-T combined variances, when constructing maps.

One caveat to address is that this study does not consider the resolution effects of the mapping grid. If the grid resolution doubles to approximately $1° \times 0.5°$ in the SO domain, then data redundancy between grid points would likely become an issue and would need to be considered for the optimal sampling strategy. It is worth assessing decorrelation scales (Mazloff et al. 2018) and performing sensitivity analyses of various grid mapping resolutions. These assessments may help identify the optimal number of inserted N-S-T data within specific regions, revealing where increased resolution does not further improve the mapping skill.

It is also noted that this study assumes the statistics of the N-S-T fields are perfectly known. In our analyses there is no mapping error induced by inaccuracies in the MEOF modes. The degree that numerical models can provide accurate variance and covariance information of oceanic N-S-T fields must be assessed. Another caveat is that the analyses are carried out only with N-S-T fields at 100 m. Relationships may differ at other depths, and this must be also assessed. Finally, we note that our calculations assume the N-S-T time series at each observation location is uninterrupted. Floats are limited by battery life, and are affected by severe weather events, variations of prevailing ocean currents and sea ice cover, sensor malfunction, and interventions of signal transmission (Johnson et al. 2017; Briggs et al. 2018). Assessing the impact of spatiotemporal heterogeneity on observational coverage is left for future work.

This study is based on synthetic N-S-T datasets from a biogeochemical GCM simulation. In practice, only a limited number of in situ measurements, such as biogeochemical Argo floats, can be deployed in the SO (e.g., Johnson et al. 2017). We survey Argo float data during the 2008–15 period from the global data centers (ftp://usgodae.org/pub/outgoing/argo; ftp://ftp.ifremer.fr/ifremer/argo) and find that an average of 889 ± 112 Argo floats were deployed in the oceans south of 30°S since 2005. This amount of Argo floats, denoted as solid green vertical lines (with gray shadings) in Fig. 8, can fill out all the K1 region and half of the K2 region, and thus reduce nitrate NRMSE to 0.66, salinity NRMSE to 0.52, and temperature NRMSE to 0.54. We also mark the nominal Argo density goal of 300 km × 300 km resolution (i.e., coverage area per float; Johnson and Claustre 2016) or about 1088 Argo floats deployed in the approximately $35° \times 360°$ SO domain as shown by the dashed vertical green lines in Fig. 8. If this Argo goal is achieved, then the in situ measurement could cover the K1 region and more than half the K2 region, and reduce nitrate NRMSE to 0.62, salinity NRMSE to 0.46, and temperature NRMSE to 0.49. These findings reveal the potential capability of the Argo float array to reconstruct SO N-S-T maps adopting the *k*-means strategy.

The abovementioned comparisons suggest that Argo floats should be deployed densely where the N-S-T combined variance is high and sparsely where the variance is low, rather than just achieving uniform 300 km × 300 km spatial coverage. This study finds that targeting high N-S-T combined variance is key to improving the capability of the MEOF method to construct skilled nitrate maps. However, in practice only 30% of Argo floats (i.e., about 326 floats in the nominal float deployment) are likely to be equipped with biogeochemical sensors (Johnson and Claustre 2016). Such observational limitations stress the importance of using other biogeochemical variables, such as salinity and temperature information, of which Fig. 7 may provide useful insights. In addition, some studies showed global nitrate fields can be estimated using chlorophyll-*a*, sea surface temperatures, and mixed layer depth fields (Switzer et al. 2003; Arteaga et al. 2015). Merging these fields and other observables into the MEOF calculation and *k*-means sampling strategy may also improve our capability to construct more accurate and informative SO nitrate maps.

Acknowledgments. We thank Editor William Emery and one anonymous reviewer, whose comments and suggestions helped to improve this manuscript. This work is supported by NSF's Climate and Large-scale Dynamics Program under Grant AGS-1505145 and the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project, which is supported by the National Science Foundation (NSF) under Award PLR-1425989. The Argo float information during 2008–15 period are obtained from the global data centers (\url{ftp://usgodae.org/pub/outgoing/argo}; \url{ftp://ftp.ifremer.fr/ifremer/argo}). The Subantartic Front information is obtained from NASA's Global Change Master Directory.

REFERENCES

Alvera-Azcárate, A., A. Barth, M. Rixen, and J.-M. Beckers, 2005: Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: Application to the Adriatic Sea surface temperature. *Ocean Modell.*, **9**, 325–346, https://doi.org/10.1016/j.ocemod.2004.08.001.

——, ——, J.-M. Beckers, and R. H. Weisberg, 2007: Multivariate reconstruction of missing data in sea surface temperature, chlorophyll, and wind satellite fields. *J. Geophys. Res.*, **112**, C03008, https://doi.org/10.1029/2006JC003660.

——, ——, D. Sirjacobs, F. Lenartz, and J.-M. Beckers, 2011: Data Interpolating Empirical Orthogonal Functions (DINEOF): A tool for geophysical data analyses. *Mediterr. Mar. Sci.*, **12** (3), 5–11, https://doi.org/10.12681/mms.64.

——, ——, G. Parard, and J.-M. Beckers, 2016: Analysis of SMOS sea surface salinity data using DINEOF. *Remote Sens. Environ.*, **180**, 137–145, https://doi.org/10.1016/j.rse.2016.02.044.

Amante, C., and B. Eakins, 2009: ETOPO1 Global Relief Model converted to PanMap layer format. NOAA National Geophysical Data Center, accessed 7 February 2018, https://doi.org/10.1594/PANGAEA.769615.

Ardyna, M., H. Claustre, J. Sallee, F. D'Ovidio, B. Gentili, G. van Dijken, F. D'Ortenzio, and K. R. Arrigo, 2017: Delineating environmental control of phytoplankton biomass and phenology in the Southern Ocean. *Geophys. Res. Lett.*, **44**, 5016–5024, https://doi.org/10.1002/2016GL072428.

Arrigo, K. R., 2005: Marine microorganisms and global nutrient cycles. *Nature*, **437**, 349–355, https://doi.org/10.1038/nature04159.

Arteaga, L., M. Pahlow, and A. Oschlies, 2015: Global monthly sea surface nitrate fields estimated from remotely sensed sea surface temperature, chlorophyll, and modeled mixed layer depth. *Geophys. Res. Lett.*, **42**, 1130–1138, https://doi.org/10.1002/2014GL062937.

Beckers, J.-M., and M. Rixen, 2003: EOF calculations and data filling from incomplete oceanographic datasets. *J. Atmos. Oceanic Technol.*, **20**, 1839–1856, https://doi.org/10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2.

Briggs, E. M., T. R. Martz, L. D. Talley, M. R. Mazloff, and K. S. Johnson, 2018: Physical and biological drivers of biogeochemical tracers within the seasonal sea ice zone of the Southern Ocean from profiling floats. *J. Geophys. Res. Oceans*, **123**, 746–758, https://doi.org/10.1002/2017JC012846.

Church, M. J., D. A. Hutchins, and H. W. Ducklow, 2000: Limitation of bacterial growth by dissolved organic matter and iron in the Southern Ocean. *Appl. Environ. Microbiol.*, **66**, 455–466, https://doi.org/10.1128/AEM.66.2.455-466.2000.

Dai, A., and K. E. Trenberth, 2002: Estimates of freshwater discharge from continents: Latitudinal and seasonal variations. *J. Hydrometeor.*, **3**, 660–687, https://doi.org/10.1175/1525-7541(2002)003<0660:EOFDFC>2.0.CO;2.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

De Mey, P., and A. R. Robinson, 1987: Assimilation of altimeter eddy fields in a limited-area quasi-geostrophic model. *J. Phys. Oceanogr.*, **17**, 2280–2293, https://doi.org/10.1175/1520-0485(1987)017<2280:AOAAEFI>2.0.CO;2.

Deppeler, S. L., and A. T. Davidson, 2017: Southern Ocean phytoplankton in a changing climate. *Front. Mar. Sci.*, **4**, 40, https://doi.org/10.3389/fmars.2017.00040.

Dormann, C. F., and Coauthors, 2007: Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, **30**, 609–628, https://doi.org/10.1111/j.2007.0906-7590.05171.x.

D'Ortenzio, F., and M. Ribera d'Alcalà, 2009: On the trophic regimes of the Mediterranean Sea: A satellite analysis. *Biogeosciences*, **6**, 139–148, https://doi.org/10.5194/bg-6-139-2009.

——, D. Antoine, E. Martinez, and M. Ribera d'Alcalà, 2012: Phenological changes of oceanic phytoplankton in the 1980s and 2000s as revealed by remotely sensed ocean-color observations. *Global Biogeochem. Cycles*, **26**, GB4003, https://doi.org/10.1029/2011GB004269.

Ducklow, H. W., D. K. Steinberg, and K. O. Buesseler, 2001: Upper ocean carbon export and the biological pump. *Oceanography*, **14** (4), 50–58, https://doi.org/10.5670/oceanog.2001.06.

Dugdale, R. C., and J. J. Goering, 1967: Uptake of new and regenerated forms of nitrogen in primary productivity. *Limnol. Oceanogr.*, **12**, 196–206, https://doi.org/10.4319/lo.1967.12.2.0196.

Elderfield H., Ed., 2006: *The Oceans and Marine Geochemistry*. Treatise on Geochemistry, Vol. 6, Pergamon, 664 pp.

Ferrari, R., C. Artana, M. Saraceno, A. R. Piola, and C. Provost, 2017: Satellite altimetry and current-meter velocities in the Malvinas Current at 41°S: Comparisons and modes of variations. *J. Geophys. Res. Oceans*, **122**, 9572–9590, https://doi.org/10.1002/2017JC013340.

Firdaus, S., and M. A. Uddin, 2015: A survey on clustering algorithms and complexity analysis. *Int. J. Comput. Sci. Issues*, **12** (2), 62–85.

Forgy, E., 1965: Cluster analysis of multivariate data: Efficiency versus interpretability models. *Biometrics*, **21**, 768–769.

Fukumori, I., and C. Wunsch, 1991: Efficient representation of the North Atlantic hydrographic and chemical distributions. *Prog. Oceanogr.*, **27**, 111–195, https://doi.org/10.1016/0079-6611(91)90015-E.

Galbraith, E. D., A. Gnanadesikan, J. P. Dunne, and M. R. Hiscock, 2010: Regional impacts of iron-light colimitation in a global biogeochemical model. *Biogeosciences*, **7**, 1043–1064, https://doi.org/10.5194/bg-7-1043-2010.

Garcia, H. E., and Coauthors, 2013a: *Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation*. Vol. 3, *World Ocean Atlas 2013*, NOAA Atlas NESDIS 75, 27 pp., https://doi.org/10.7289/V5XG9P2W.

——, R. A. Locarnini, T. P. Boyer, J. I. Antonov, O. Baranova, M. M. Zweng, J. R. Reagan, and D. R. Johnson, 2013b: *Dissolved Inorganic Nutrients (Phosphate, Nitrate, Silicate)*. Vol. 4, *World Ocean Atlas 2013*, NOAA Atlas NESDIS 76, 25 pp., https://doi.org/10.7289/V5J67DWD.

Hammond, M. D., and D. C. Jones, 2016: Freshwater flux from ice sheet melting and iceberg calving in the Southern Ocean. *Geosci. Data J.*, **3**, 60–62, https://doi.org/10.1002/gdj3.43.

Hartigan, J. A., and M. A. Wong, 1979: Algorithm AS 136: A k-means clustering algorithm. *Appl. Stat.*, **28**, 100–108, https://doi.org/10.2307/2346830.

Hoppe, C. J. M., and Coauthors, 2015: Controls of primary production in two phytoplankton blooms in the Antarctic Circumpolar Current. *Deep Sea Res. II*, **138**, 63–73, https://doi.org/10.1016/j.dsr2.2015.10.005.

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**, 3586–3594, https://doi.org/10.1002/jgrc.20249.

Johnson, K. S., and H. Claustre, 2016: The scientific rationale, design and implementation plan for a biogeochemical-Argo float array. K. Johnson and H. Claustre, Eds., Biogeochemical-Argo Planning Group, 58 pp., https://doi.org/10.13155/46601.

——, J. N. Plant, J. P. Dunne, L. D. Talley, and J. L. Sarmiento, 2017: Annual nitrate drawdown observed by SOCCOM profiling floats and the relationship to annual net community production. *J. Geophys. Res. Oceans*, **122**, 6668–6683, https://doi.org/10.1002/2017JC012839.

Kaplan, A., Y. Kushnir, M. A. Cane, and M. B. Blumenthal, 1997: Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures. *J. Geophys. Res.*, **102**, 27 835–27 860, https://doi.org/10.1029/97JC01734.

Key, R. M., and Coauthors, 2015: Global Ocean Data Analysis Project, version 2 (GLODAPv2). Oak Ridge National Laboratory Carbon Dioxide Information Analysis Center, accessed 7 February 2018, doi:10.3334/CDIAC/OTG.NDP093_GLODAPv2.

Kondrashov, D., and M. Ghil, 2006: Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes Geophys.*, **13**, 151–159, https://doi.org/10.5194/npg-13-151-2006.

Kruskal, W. H., and W. A. Wallis, 1952: Use of ranks in one-criterion variance analysis. *J. Amer. Stat. Assoc.*, **47**, 583–621, https://doi.org/10.1080/01621459.1952.10483441.

Lacour, L., H. Claustre, L. Prieur, and F. D'Ortenzio, 2015: Phytoplankton biomass cycles in the North Atlantic subpolar gyre: A similar mechanism for two different blooms in the Labrador Sea. *Geophys. Res. Lett.*, **42**, 5403–5410, https://doi.org/10.1002/2015GL064540.

Lauvset, S. K., and Coauthors, 2016: A new global interior ocean mapped climatology: The 1° × 1° GLODAP version 2. *Earth Syst. Sci. Data*, **8**, 325–340, https://doi.org/10.5194/essd-8-325-2016.

Letscher, R. T., F. Primeau, and J. K. Moore, 2016: Nutrient budgets in the subtropical ocean gyres dominated by lateral transport. *Nat. Geosci.*, **9**, 815–819, https://doi.org/10.1038/ngeo2812.

Losch, M., D. Menemenlis, J.-M. Campin, P. Heimbach, and C. Hill, 2010: On the formulation of sea-ice models. Part 1: Effects of different solver implementations and parameterizations. *Ocean Modell.*, **33**, 129–144, https://doi.org/10.1016/j.ocemod.2009.12.008.

MacQueen, J., 1967: Some methods for classification and analysis of multivariate observations. *Statistics*, L. M. Le Cam and J. Neyman, Eds., Vol. 1, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 281–297.

Marshall, J., A. Adcroft, C. Hill, L. Perelman, and C. Heisey, 1997: A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers. *J. Geophys. Res.*, **102**, 5753–5766, https://doi.org/10.1029/96JC02775.

Mayot, N., F. D'Ortenzio, M. R. D'Alcalà, H. Lavigne, and H. Claustre, 2016: Interannual variability of the Mediterranean trophic regimes from ocean color satellites. *Biogeosciences*, **13**, 1901–1917, https://doi.org/10.5194/bg-13-1901-2016.

Mazloff, M. R., P. Heimbach, and C. Wunsch, 2010: An eddy-permitting Southern Ocean state estimate. *J. Phys. Oceanogr.*, **40**, 880–899, https://doi.org/10.1175/2009JPO4236.1.

——, B. Cornuelle, S. Gille, and A. Verdy, 2018: Correlation lengths for estimating the large-scale carbon and heat content of the Southern Ocean. *J. Geophys. Res. Oceans*, **123**, 883–901, https://doi.org/10.1002/2017JC013408.

Moore, C. M., and Coauthors, 2013: Processes and patterns of oceanic nutrient limitation. *Nat. Geosci.*, **6**, 701–710, https://doi.org/10.1038/ngeo1765.

Munro, D. R., and Coauthors, 2015: Estimates of net community production in the Southern Ocean determined from time series observations (2002–2011) of nutrients, dissolved inorganic carbon, and surface ocean pCO_2 in Drake Passage. *Deep-Sea Res. II*, **114**, 49–63, https://doi.org/10.1016/j.dsr2.2014.12.014.

Nikolaidis, A., G. Georgiou, D. Hadjimitsis, and E. Akylas, 2014: Filling in missing sea-surface temperature satellite data over the Eastern Mediterranean Sea using the DINEOF algorithm. *Open Geosci.*, **6**, 27–41, https://doi.org/10.2478/s13533-012-0148-1.

Orsi, A. H., T. Whitworth, and W. D. Nowlin, 1995: On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep-Sea Res. I*, **42**, 641–673, https://doi.org/10.1016/0967-0637(95)00021-W.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Plant, J. N., K. S. Johnson, C. M. Sakamoto, H. W. Jannasch, L. J. Coletti, S. C. Riser, and D. D. Swift, 2016: Net community production at Ocean Station Papa observed with nitrate and oxygen sensors on profiling floats. *Global Biogeochem. Cycles*, **30**, 859–879, https://doi.org/10.1002/2015GB005349.

Reynolds, R. W., and T. M. Smith, 1994: Improved global sea surface temperature analyses using optimum interpolation. *J. Climate*, **7**, 929–948, https://doi.org/10.1175/1520-0442(1994)007<0929:IGSSTA>2.0.CO;2.

Rosso, I., M. R. Mazloff, A. Verdy, and L. D. Talley, 2017: Space and time variability of the Southern Ocean carbon budget. *J. Geophys. Res. Oceans*, **122**, 7407–7432, https://doi.org/10.1002/2016JC012646.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, **14**, 853–871, https://doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2.

Shirkhorshidi, A. S., S. Aghabozorgi, T. Y. Wah, and T. Herawan, 2014: Big data clustering: A review. *Computational Science and Its Applications—ICCSA 2014*, B. Murgante et al., Eds., Lecture Notes in Computer Science, Vol. 8583, Springer, 707–720, https://doi.org/10.1007/978-3-319-09156-3_49.

Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate*, **9**, 1403–1420, https://doi.org/10.1175/1520-0442(1996)009<1403:ROHSST>2.0.CO;2.

Sparnocchia, S., N. Pinardi, and E. Demirov, 2003: Multivariate empirical orthogonal function analysis of the upper thermocline structure of the Mediterranean Sea from observations and model simulations. *Ann. Geophys.*, **21**, 167–187, https://doi.org/10.5194/angeo-21-167-2003.

Stammer, D., and Coauthors, 2002: Global ocean circulation during 1992–1997, estimated from ocean observations and a general circulation model. *J. Geophys. Res.*, **107**, 3118, https://doi.org/10.1029/2001JC000888.

Switzer, A. C., D. Kamykowski, and S.-J. Zentara, 2003: Mapping nitrate in the global ocean using remotely sensed sea surface temperature. *J. Geophys. Res.*, **108**, 3280, https://doi.org/10.1029/2000JC000444.

Verdy, A., and M. R. Mazloff, 2017: A data assimilating model for estimating Southern Ocean biogeochemistry. *J. Geophys. Res. Oceans*, **122**, 6968–6988, https://doi.org/10.1002/2016JC012650.

Wang, J.-F., A. Stein, B.-B. Gao, and Y. Ge, 2012: A review of spatial sampling. *Spat. Stat.*, **2**, 1–14, https://doi.org/10.1016/j.spasta.2012.08.001.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

Williams, R. G., and M. J. Follows, 2003: Physical transport of nutrients and the maintenance of biological production. *Ocean Biogeochemistry*, Global Change—The IGBP Series, Springer, 19–51, https://doi.org/10.1007/978-3-642-55844-3_3.

Wunsch, C., and P. Heimbach, 2007: Practical global oceanic state estimation. *Physica D*, **230**, 197–208, https://doi.org/10.1016/j.physd.2006.09.040.

Xue, Y., A. Leetmaa, and M. Ji, 2000: ENSO prediction with Markov models: The impact of sea level. *J. Climate*, **13**, 849–871, https://doi.org/10.1175/1520-0442(2000)013<0849:EPWMMT>2.0.CO;2.