

UC San Diego

UC San Diego Previously Published Works

Title

Using existing Argo trajectories to statistically predict future float positions with a transition matrix

Permalink

<https://escholarship.org/uc/item/7n3042jq>

Authors

Chamberlain, Paul
Talley, Lynne D
Mazloff, Matthew
et al.

Publication Date

2023-06-12

DOI

10.1175/jtech-d-22-0070.1

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Using existing Argo trajectories to statistically predict future float positions with a transition matrix



Paul Chamberlain,^a Lynne D. Talley,^a Matthew Mazloff,^a Erik van Sebille,^c Sarah T. Gille,^a
Tyler Tucker,^d Megan Scanderbeg,^a Pelle Robbins^e

^a *Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA*

^b *Institute for Marine and Atmospheric Research, Utrecht University, Utrecht, Netherlands*

^c *W. M. Keck Observatory, Waimea, HI*

^d *Woods Hole Oceanographic Institution, Falmouth, MA*

Corresponding author: Paul Chamberlain, pchamber@ucsd.edu

1

Early Online Release: This preliminary version has been accepted for publication in *Journal of Atmospheric and Oceanic Technology*, may be fully cited, and has been assigned DOI 10.1175/JTECH-D-22-0070.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: The Argo array provides nearly 4000 temperature and salinity profiles of the top 2000 meters of the ocean every 10 days. Still, Argo floats will never be able to measure the ocean at all times, everywhere. Optimized Argo float distributions should match the spatial and temporal variability of the many societally important ocean features that they observe. Determining these distributions is challenging because float advection is difficult to predict. Using no external models, transition matrices based on existing Argo trajectories provide statistical inferences about Argo float motion. We use the 24 years of Argo locations to construct an optimal transition matrix that minimizes estimation bias and uncertainty. The optimal array is determined to have a $2^\circ \times 2^\circ$ spatial resolution with a 90 day timestep. We then use the transition matrix to predict the probability of future float locations of the Core Argo array, the Global Biogeochemical Array, and the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) array. A comparison of transition matrices derived from floats using Argos System and Iridium communication methods shows the impact of surface displacements, which is most apparent near the equator. Additionally, we demonstrate the utility of transition matrices for validating models by comparing the matrix derived from Argo floats with that derived from a particle release experiment in the Southern Ocean State Estimate (SOSE).

1. Introduction

The endurance and economy of Argo profiling floats revolutionized how oceanographers practice their craft by providing unprecedented spatially distributed and frequent observations. Since the development of the first Argo prototypes almost 30 years ago (Davis et al. 1992), the global Argo array has grown to include more than 3800 active floats and has collected over 20 years of data (Roemmich et al. 2019). Core Argo floats observe temperature and salinity from 2000 m to the surface every 10 days. The recent integration of miniaturized nitrate, oxygen, pH, and optical sensors has enabled the development of the biogeochemical (BGC) Argo float (Johnson et al. 2017). The Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project (Johnson and Claustre 2016; Talley et al. 2019) has successfully deployed an array of BGC Argo floats in the Southern Ocean, using the same observing protocol as Core Argo, and recently funded projects in multiple countries are deploying them globally as part of OneArgo.

The substantially higher costs of BGC Argo floats relative to Core Argo floats accelerates the need for informed strategic decisions about the optimal deployment positions for the BGC array. Any observing array should be optimized to sample the temporal and spatial scales of the phenomena of interest. Each variable measured by BGC floats has distinct scales of spatial and temporal variability. In the past, Core Argo array deployment locations were selected to optimize a uniform distribution with a spacing of about $3^{\circ} \times 3^{\circ}$ separation in latitude and longitude (Davis 1991; Roemmich et al. 1998). With increased prior knowledge of the ocean state, recent Core Argo array proposals have suggested increasing array density where variability is higher, along the equator and in western boundary regions (Roemmich et al. 2019). The Biogeochemical Argo Implementation Plan (Johnson and Claustre 2016) explored spatial variability of measured BGC variables, concluding that initial deployments should try to achieve a uniform distribution, but finally suggesting that uniform deployments should “be tested as more experience is obtained” (Johnson and Claustre 2016).

This raises the question: should we update recommended deployment strategies now that more BGC Argo floats have been deployed? Past studies have quantified the improvement BGC float observations make to either modeled and calculated fields of an individual BGC variable for both random sampling (Johnson and Claustre 2016; Majkut et al. 2014) and snapshots of past Argo

array distributions (Ford 2020; Kamenkovich et al. 2017). However, optimal design strategies on a global scale for the BGC Argo array deserve dedicated studies.

The work presented here addresses the operational concern of predicting future float locations. It is one piece of an expanded effort of an optimal ship-based float deployment strategy for the integrated Core and BGC Argo arrays. We propose an array design strategy that consists of three innovations:

1. The system will statistically predict the future location of currently deployed instruments to recognize the gaps in coverage at the time of deployment.
2. The array strategy will account for the global inhomogeneities of BGC variables in spatial covariance and temporal variance by putting greater float density in regions of high temporal variability and low spatial covariance.
3. The optimal array strategy will account for the cross-covariance of the full BGC Argo float sensor suite by considering the additional constraint imposed by the prior knowledge of covarying properties.

The latter two developments are left for a companion manuscript (Chamberlain et al. 2023). Several successful pilot projects have created regional BGC float arrays in regions that play an outsized role in global biogeochemistry (Morrison et al. 2015). The largest regional array, the SOCCOM program, deploys floats in unique provinces of the Southern Ocean by analyzing observed float trajectories and numerical particle release experiments (Talley et al. 2019). This approach has shortcomings: visual analysis of previous Argo trajectories is subjective, while particle release experiments do not consistently reproduce actual Argo trajectories (Talley et al. 2019).

Indeed, the ocean is a complex system, and Lagrangian trajectories can be challenging to predict deterministically. Figure 1 shows an example of historic float trajectories passing through a region off Cape Agulhas; there exists a time-varying bifurcation of float trajectories in this complex current system (Boebel et al. 2003; Van Sebille et al. 2010). Argo float trajectories depend on the mesoscale eddy field, which models may not resolve. Eddies result from the intrinsic instabilities of the ocean, and even in eddy-resolving assimilating models the positions and timing of observed eddies can differ from modeled eddies due to constraints on dynamical consistency. The rectified effect of these unresolved or omitted processes is typically stochastically parameterized in Lagrangian

models (Van Sebille et al. 2018). Argo floats also experience ocean shear during their ascent and descent and are advected by winds, currents, and waves at the surface. Some Argo-derived velocity products do not include these processes (Gray and Riser 2014), or attempt to actively remove them (Sevellec et al. 2017; Ollitrault and Rannou 2013; Gille and Romero 2003).

Since Argo floats are not propelled, Argo managers should carefully choose deployment locations to optimize Argo array distribution throughout the lifetime of the float. To address these challenges, we generate a statistical model, known as a transition matrix, from existing Argo array trajectories to predict the probability density function (PDF) of future float locations. The large number of Argo float trajectories can be used to diagnose the probability that a float transitions from one location to the other in a given time step (Fig. 2). Transition matrices represent a potential complement to dynamical models because they contain a probabilistic representation of the complexities of the eddy field, ocean shear, and surface processes that models may miss. Transition matrices are also a way of quantifying the information gained from visually inspecting previous float trajectories.

Transition matrices are an established method (Markov 1906; Van Sebille et al. 2012; Maximenko et al. 2012; Sevellec et al. 2017; Drouin et al. 2022; Miron et al. 2022; Abernathey et al. 2022) to model semi-Lagrangian ocean drifters, and have been primarily applied to surface ocean drifters (Maximenko et al. 2012; Van Sebille et al. 2012) and surface drifter array design (Lumpkin et al. 2016). Sevellec et al. (2017) generated a transition matrix based on processed Argo trajectories obtained from the ANDRO dataset (Ollitrault and Rannou 2013) to study the evolution of deep water masses. The ANDRO dataset estimates the displacements of Argo floats at their drift depth and removes displacements due to ocean shear and surface currents. However, these displacements are important for operational float prediction and should not be removed, so for float deployment analyses a new transition matrix was required.

Here we first broadly explain the relatively simple theory of the construction and use of transition matrices (Section 3). The biases and uncertainties of the estimates that transition matrices produce are sensitive to spatial and temporal spacing and are quantified in Section 4a. In this section, we also define the criteria for the optimal transition matrix that minimizes these biases and uncertainties, based on the available datasets.

To apply this analysis, in Section 4b we use our optimal transition matrix to estimate the future density of the existing Core Argo array and the future array health of the Core Argo and SOCCOM

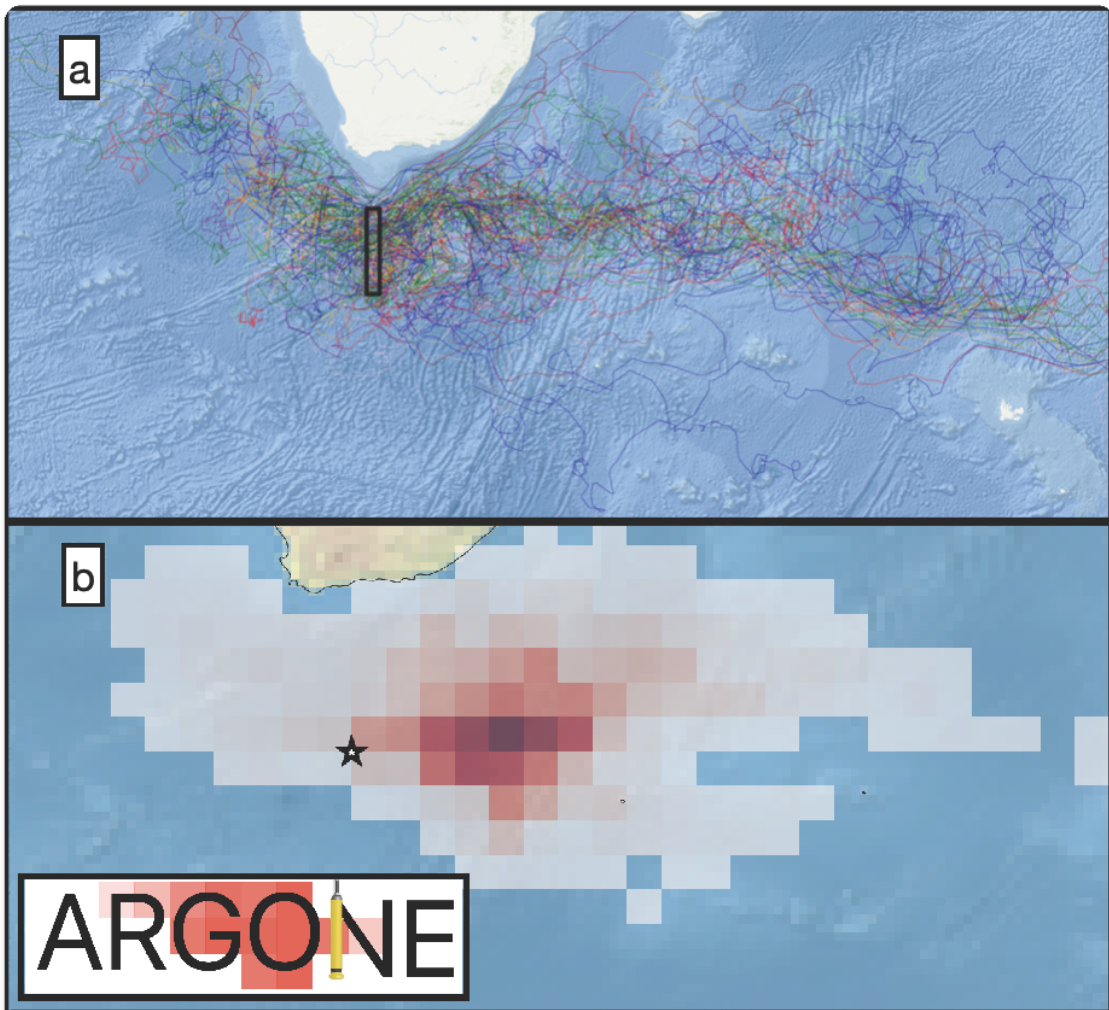


FIG. 1. a) Historical Argo trajectories off Cape Agulhas passing through the black box. White shading indicates land, colored lines indicate individual float tracks. b) Example of float prediction product ARGONE for a float in the same region. This panel was created by accessing the product from the Argovis web app and database (recently upgraded from the app version described in Tucker et al. (2020)), using the Argovis API (https://github.com/argovis/demo_notebooks). The star symbol indicates the deployment location, and colors of the grid cells represent values of the float probability density function after 270 days from the deployment.

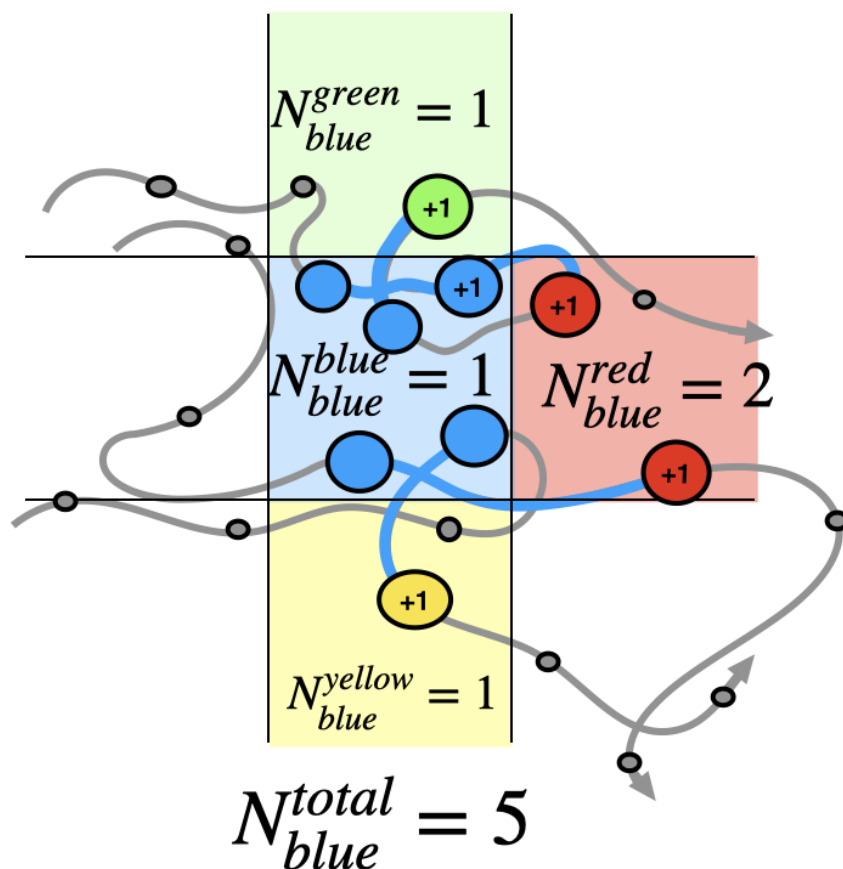


FIG. 2. Cartoon of transition matrix calculation for an arbitrary $Cell_{blue}$. Gray lines represent hypothetical paths of 3 distinct floats, and circles represent independent float profile locations given transition matrix timestep. The timesteps are long enough (e.g. greater than 30 days) that each segment can be considered independent even if they come from the same float. Black squares represent grid cells. Blue lines represent trajectory segments originating from $Cell_{blue}$. Colored circles and “+ 1” text represents trajectory segments that ended in correspondingly colored grid cells. Small gray circles represent float positions that are not included in the transition statistics for $Cell_{blue}$. Total probability for each element of the transition matrix will equal the total number of transitions within a grid cell divided by the total number of transitions in the dataset (N_{blue}^{total}).

Arrays. In Section 4c, we consider the potential future locations of floats deployed from upcoming GO-SHIP cruise tracks over the float lifecycle of 5 years. Then, in Section 4d, we quantify the different drift patterns of Argos System and Iridium-equipped floats using transition matrices derived from these different float trajectories.

BGC Argo floats do not all carry the same sensor suite, and some BGC sensors are more ubiquitous in the ocean than others. Estimating where individual BGC sensors will observe the ocean is important for BGC Argo managers to determine where to deploy BGC floats and where the gaps will be in our BGC observing systems. In Section 4e, we quantify sampling probability by BGC sensor type.

Finally, in Section 4f, the Argo float transition matrix is compared against a transition matrix derived from the modeled particle trajectories derived from the Southern Ocean State Estimate (SOSE) (Mazloff et al. 2010). Modeled particles were programmed to profile to the surface every 10 days, similar to real Argo floats, and were advected by SOSE currents.

2. Data

a. Argo Floats

These results use Argo float trajectories from the October 2022 Argo Snapshot (Argo 2022). The data processing for Argo files excluded trajectories with the following conditions: poor quality flags for position, time, and pressure; floats with problematic file formats; floats that were functioning in a manner outside of Core Argo mission parameters of 1000 db drift depths and 10-day surface intervals; and floats that had unrealistic velocities or traveled over 500 nm in successive positions. After these quality control procedures, the total dataset comprised 2,167,492 positions collected by 14,331 unique floats. These trajectories were measured from May 13, 1998, to October 10, 2022, and spanned the globe from 77.7°S to 89.7°N.

Floats have used two different means for satellite communications: 8,506 floats used the Argos System constellation, and 5,825 floats used Iridium (Fig. 3). The Iridium and Argos constellations of satellites are both low-Earth orbiting. Floats transmit their data more efficiently to Iridium satellites than to Argos; therefore, Iridium-enabled floats typically spend less than 1 hour at the surface compared to the 8 hours that Argos-enabled floats usually spend at the surface.

b. Southern Ocean State Estimate

To validate the results based on Argo float trajectories, we also built an independent transition matrix from a Lagrangian particle release experiment in the Southern Ocean State Estimate (SOSE) (Mazloff et al. 2010). SOSE is an eddy-permitting 0.16° configuration of the Massachusetts

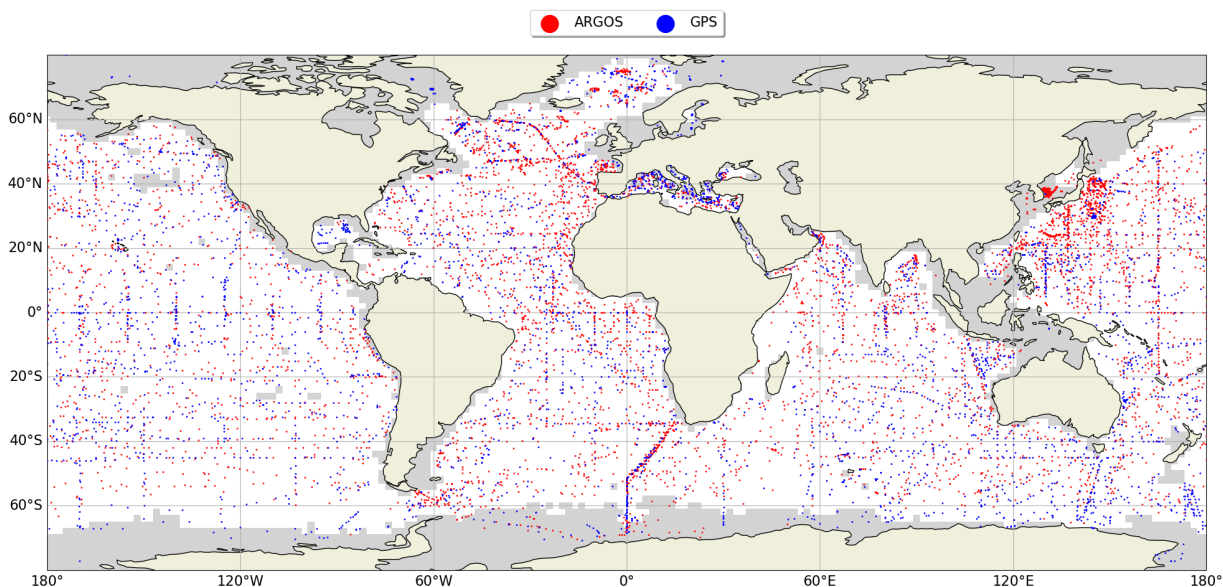


FIG. 3. Initial deployment locations for all Argo floats in the dataset, using both Argos System (red) and Iridium (blue) communications.

Institute of Technology General Circulation Model (MITGCM), which is fit by constrained least-squares to available Southern Ocean satellite and hydrographic observations. The current SOSE version (iteration 100) spans six years (2005-2010) and is calculated from 24.7°S to 78°S. For the particle experiment using Octopus (<http://github.com/jinbow/Octopus>), we randomly released 10,000 particles over the spatial domain and tracked their motion over the full six years of model output. The particles in this release experiment were programmed to drift at 1,000 meters, dive to 2,000 meters, then surface once every ten days with an ascent mission and surface time similar to real Argo floats to simulate the effect of upper-ocean shear on particle trajectories.

3. Methods

A transition matrix is a square matrix used to define the transitions of a discrete-time Markov Chain of a state vector in the state space. For our application, the state space is a discrete spatial grid S . In the real ocean, Argo float position is continuous; one approximation of this method

is that it restricts the locations where Argo floats can be in the ocean to the positional grid cells of S . The state vector is the distribution of the Argo floats that we wish to study. The Markov Chain is the future probability of these originally deployed Argo floats at subsequent locations. The distribution of Argo floats (the state vector) could be a single float or the entire array. This is expressed mathematically as

$$\boldsymbol{\rho}(t+k) = \mathbf{H}\boldsymbol{\rho}(t), \quad \mathbf{H} = \mathbf{Q}^k \quad (1)$$

where \mathbf{Q} is a square $S \times S$ matrix and is the transition matrix calculated at the original timestep; \mathbf{H} is the product of \mathbf{Q} multiplied by itself k times and defines the probability of transition to the k^{th} timestep; and $\boldsymbol{\rho}(t)$ is a column vector in the state space of dimension $S \times 1$ at a given timestep t and represents the distribution of Argo floats at the original timestep. Using equation 1, the probability of an initial array of Argo floats, $\boldsymbol{\rho}(t)$, propagating to a future state at the k^{th} timestep, $\boldsymbol{\rho}(t+k)$, can be quantified. The process for creating a transition matrix from Lagrangian data is well described (Sevellec et al. 2017), so we provide only a brief explanation here. All algorithms used in these calculations are made publicly available in a GitHub repository <https://github.com/Chamberpain/TransitionMatrix> (Chamberlain 2023b)).

First, we spatially and temporally quantize the trajectory data by a defined timestep and spatial grid (Fig. 2). These choices define the dimension of the state space S and the nature of the discrete-time Markov Chain. For our application, these choices determine the spatial grid of latitude and longitude that Argo floats transition between and the timestep in days of these transitions.

A consideration of space-time resolution is the representativeness of positioning: the approximation-induced error of projecting continuous Argo trajectories onto a discrete spatial grid increases when the grid cells are larger. This approximation can lead to biases and uncertainties in estimating future float distribution produced by a transition matrix. For example, if the spatial grid size is too large, some floats may enter along the edge of grid cells and quickly leave; this will cause a bias in transition statistics compared with floats that transit across the entirety of the grid cell. Using large grid cells can also result in underestimating float motion when the grid spacing is larger than coherent structures in the mesoscale eddy field, as floats can recirculate within the grid cell without transitioning to an adjacent grid cell.

Decreasing grid cell size reduces position discrepancies between grid centers and edges. However, it can also lead to fewer transition data per grid cell as smaller grid cells will typically have

fewer floats pass through them. Transition matrices are uninformed by dynamics and need many data to resolve skillful transition statistics. Section 4a explores the trade-offs between grid cell size, timestep, and data density. The distribution of floats relative to the nature of the velocity field is another potential bias; unequally spaced float arrays placed in fields of inhomogeneous diffusivity may infer biased velocity statistics that do not resolve the mean (Freeland 1975; Davis 1991). For example, many floats placed in the middle of a random velocity field with no mean flow may appear divergent due to simple Brownian motion. For these calculations, we assumed the float density to be homogeneous.

Float trajectories incorporated into the transition matrix must be temporally statistically independent. Successful floats carry out extended missions (five years or more), much longer than any transition matrix considered in this analysis. Therefore, we break these longer float trajectories down into shorter trajectories equal to the timestep of the transition matrix. Figure 2 shows the segments corresponding to these shorter trajectories as black circles. Our algorithm tolerated small overlaps between the shorter trajectories to increase data density. The time separation between the start of the smaller trajectories was the greater of either 30 days (Gille and Romero 2003) or one-third of the transition matrix time step; e.g., the time separation for the start of trajectories must be 30 days for a 60-day timestep and 60 days for a 180-day timestep.

We now describe how we quantify the transition probability for an arbitrary spatial grid cell (which we call Cell_{blue} in Figure 2). This process must be repeated for all spatial Grid Cells in the domain. The number of spatial grid cells is equal to S , the dimension of the state space. First, all of the independent Argo profiles that start in Cell_{blue} are identified (colored circles in Figure 2). N_{blue}^{total} defines the total number of starting positions (blue circles in Figure 2). Figure 2 shows the float trajectory segments that start in Cell_{blue} as blue lines. We next count the total number of floats that start in Cell_{blue} and have ended up in the surrounding grid cells. In Figure 2, the grid cells connected to Cell_{blue} via trajectories are Cell_{green} , Cell_{red} , and Cell_{yellow} with number of transitions into each grid cell of N_{blue}^{green} , N_{blue}^{red} , and N_{blue}^{yellow} respectively. The colored circles surrounding the “+1” text in Figure 2 denote a trajectory segment that has ended in a grid cell and are included in the total count of transition statistics. It is possible (and, depending on the location of the ocean and the grid cell size, probable) for floats to stay in the grid cell they started in. The

Grid size (lat × lon)	Timestep (days)
1° × 1°	30
1° × 2°	60
2° × 2°	90
2° × 3°	120
3° × 3°	160
4° × 4°	180
4° × 6°	

TABLE 1. (left) Spatial resolution of all transition matrices calculated. (right) Temporal resolution (timestep) of transition matrices calculated for each spatial resolution.

blue circle surrounding the “+1” text in Figure 2 represents a trajectory segment that started and stayed in Cell_{blue} . Therefore, in this example, $N_{blue}^{blue} = 1$.

The probability of a float transitioning from Cell_{blue} (arbitrarily of index k) to Cell_{red} (arbitrarily of index q) is equal to $N_{blue}^{red} / N_{blue}^{total}$ and is the value of the transition matrix in the index k^{th} column and q^{th} row. The probability of a float staying in the grid cell it started in (the probability of a float transitioning from the Cell_{blue} to Cell_{blue}) is found on the diagonal of the transition matrix and calculated as $N_{blue}^{blue} / N_{blue}^{total}$. Our model does not consider the possibility of float failure: floats can neither be created nor destroyed. Consequently, we impose a conservative tracer constraint such that all columns of the transition matrix sum to one, and the entire column vector is scaled accordingly. For dynamical reasons, the transition matrix is very sparse: floats in the middle of the Atlantic have no chance of transitioning to the Pacific in weeks or months. The number of nonzero rows in each column will equal the number of grid cells into which floats transition during the timestep. To ensure statistically likely transitions over long timesteps, grid cells with a probability less than 4% of the mean transition were set to zero at each step of the Markov Chain, and the transition matrix was rescaled. After performing this calculation for all spatial grid cells, the resultant matrix is $S \times S$.

To quantify the transition matrix’s biases and uncertainties, we considered transition matrix performance at several different spatial and temporal resolutions. Timesteps ranged from 30 to 180 days, and grid cells ranged in size from 1° × 1° to 4° × 6°. Table 1 lists these timesteps and gridsizes.

Low trajectory density or errors in the trajectory dataset can create isolated grid cells disconnected from the rest of the transition matrix. These points have no predictive value and are removed. Eigenvector decomposition and analysis (Miron et al. 2019; Froyland et al. 2014) provides intuition about the connected float distribution modes; regions defined by eigenvectors with eigenvalues close to one tend to have closed circulations and are difficult for floats to leave. To eliminate isolated grid cells, we required all 50 largest eigenvectors to have at least three grid cells and a total number of at least three transitions in each grid cell of the transition matrix. The limiting values for minimum grid cells and transitions were chosen empirically based on the data density.

Ocean dynamics vary in time; therefore, the statistics of where floats are advected must also have seasonal and climate timescale variability. The Argo trajectory dataset is not sufficiently large to resolve seasonal dynamics, and, by necessity, we assumed these statistics are stationary in time. This assumption is a fundamental gap in our analysis, adding uncertainty to our estimates.

Expanding this statistical prediction for two special cases relevant to the Argo array, the probability of a float sampling a given region within a range of timesteps is expressed as

$$\mathbf{P}^{total} = \sum_{k=1}^n \mathbf{Q}^k \boldsymbol{\rho} \quad (2)$$

where \mathbf{P}^{total} is the probability of sampling and n is the number of timesteps considered. And, the probability of a float sampling a given region within all timesteps is expressed as

$$\mathbf{P}_j^{range} = \Pi\{(\mathbf{Q}^k \boldsymbol{\rho})_j : k = 1, \dots, n\} \quad (3)$$

where the product is taken elementwise over the product $\mathbf{Q}^k \boldsymbol{\rho}$. The first statistical moment of a generic spatial grid k can be expressed as

$$\mathbf{R}_j = \sum_{i=1}^M q_i \mathbf{r}_i \quad (4)$$

where \mathbf{r}_i is the relative displacement vector from spatial grid j to spatial grid i , and q_i is the probability of transition from spatial grid j to spatial grid i . The second statistical moment of the

generic grid k can be expressed as the following:

$$\mu_j = \sqrt{\sum_{i=1}^M q_i (\mathbf{R}_j - \mathbf{r}_i)^2} \quad (5)$$

These calculations are the first and second moments of a probability distribution. Equations 4 and 5 can be interpreted as the expected value and variance (the tendency for Lagrangian particles to disperse) for a linear probability distribution. These metrics may not be appropriate in regions where a strong bifurcation in the advection statistics predicts two (or more) distinct and narrowly defined probability distributions. Fortunately for our analysis, most of the ocean does not behave this way.

Standard error relates the second moment of the probability distribution to the number of realizations. It quantifies the quality of our transition matrix as an estimator of the most probable trajectory. The standard error is

$$\mu_{\bar{j}} = \frac{\mu_j}{\sqrt{N_j^{total}}} \quad (6)$$

where μ_j is the second moment from Equation 5, and N_j^{total} is the Transition Density (N_{blue}^{total} in Fig. 2).

Standard error also determines the statistical significance of differences between distributions in a Z test. The Z test is defined as

$$Z_j = \frac{\bar{X}_j - \mathbf{R}_j}{\mu_{\bar{j}}} \quad (7)$$

where \bar{X} is the sample mean of the j^{th} grid cell, \mathbf{R}_j is the first moment of the j^{th} grid cell.

4. Results and Discussion

Previous BGC Argo array design studies have considered both the actual Argo array at snapshots in time and randomly distributed float arrays (Johnson and Claustre 2016; Majkut et al. 2014; Ford 2020; Kamenkovich et al. 2017). The omission of float displacement has been a major limitation. Accounting for this float motion is critical for planning arrays over timespans long enough for instruments to drift significantly, as is the case with Argo. Ocean currents, and the trajectories of floats carried by these currents, follow predictable patterns. Inspired by this, we consider

the construction and assessment of a transition matrix approach for float prediction in several applications: Section 4a quantifies the biases and uncertainties of transition matrices of various spatial and temporal resolutions and presents our justification criteria for the optimal transition matrix; Section 4b predicts the future distribution of the existing Argo array; Section 4c predicts the future distribution of Argo floats deployed from planned GO-SHIP cruises within the next 5 years; Section 4d estimates the regions of convergence and divergence for Argos System and Iridium floats; Section 4e predicts the future sampling of existing BGC Argo floats broken down by sensor class; Section 4f estimates the effective diffusivity of the SOSE model with derived transition matrices.

a. Bias and Uncertainty Quantification

Although computationally straightforward, a limiting assumption of the transition matrix is that it is a linear approximation to a nonlinear process (Lagrangian Argo trajectories in the ocean) (McAdam and van Sebille 2018). Given a dataset of finite size, the choice of resolution and timestep is fundamentally a trade-off of model bias versus model uncertainty. Using short space and long time scales reduces the impact of errors from the linearity assumption (less bias). However, these dimensional choices also reduce the amount of data available to construct the transition matrix (more uncertainty). As such, the timestep and grid spacing, which define the transition matrix, are factors that determine the accuracy of transition matrix prediction.

A representation of the difference in estimates derived from approximating a longer time step transition matrix by multiplying many short timestep transition matrices together is shown for an example grid cell in the Antarctic Circumpolar Current (ACC) in Figure 4 a and b. While this is only one example, it illustrates the uncertainty in the linear approximation made in this method: approximating long-term float behavior with short-term float statistics will result in a smoothed PDF distribution. Indeed, this approximation-induced smoothing can change the first moment (Eq. 4) of the future float PDF. Over long durations, these differences compound (Figure 4c and d); in the aggregated, globally calculated statistics of this figure, misfit in mean transition and standard deviation generally decrease with increasing timestep. The slope of the misfit decrease is similar among two groups of resolution: resolutions of $1^\circ \times 1^\circ$, $1^\circ \times 2^\circ$, $2^\circ \times 2^\circ$, and $2^\circ \times 3^\circ$ of latitude and longitude; and resolutions of $3^\circ \times 3^\circ$, $4^\circ \times 4^\circ$, and $4^\circ \times 6^\circ$ of latitude and longitude. The former

had a lower misfit at the shortest timestep and a shallower slope of misfit decrease with increasing timestep. In comparison, the latter had a higher misfit at the shortest timestep and showed a steeper misfit decrease with increasing timestep. The misfit of the higher resolution group also plateaued at the 90-day timestep.

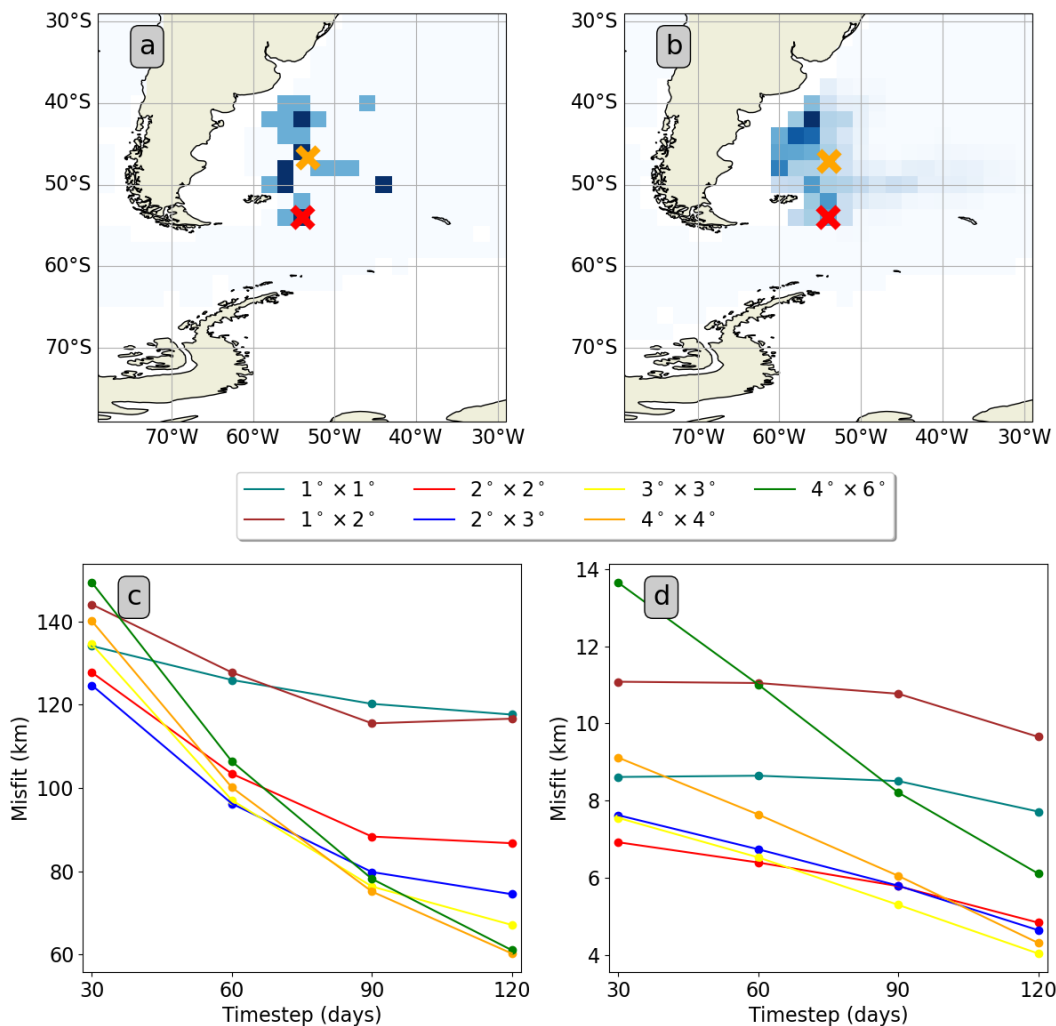


FIG. 4. Examples of $2^{\circ} \times 2^{\circ}$ 180 day transition probabilities starting from red X for (a) a native 180 day transition matrix, and (b) a 30 day transition matrix multiplied by itself 5 times. Blue shading represents transition probability and orange X represents spatial mean of all transition probabilities. Beige shading represents land. (c) Mean difference over all grid cells in the first statistical moment (Eq. 4) of transition matrices of various timesteps and transition matrix with 180 day timestep of same spatial resolution. (d) Mean difference over all grid cells in the second statistical moment (Eq. 5) of transition matrices of various timesteps and transition matrix with 180 day timestep of same spatial resolution. Colored lines represent different resolutions.

An example of the differences introduced by choices in spatial resolution is shown for the same example grid cell in the ACC (Fig. 5a and b). These differences have also been quantified for transition matrices of different grid resolutions and time steps (Fig. 5c). Mean misfit is proportional to timestep and inversely proportional to resolution, and the misfit slope is generally lowest between the 60 and 90-day timesteps.

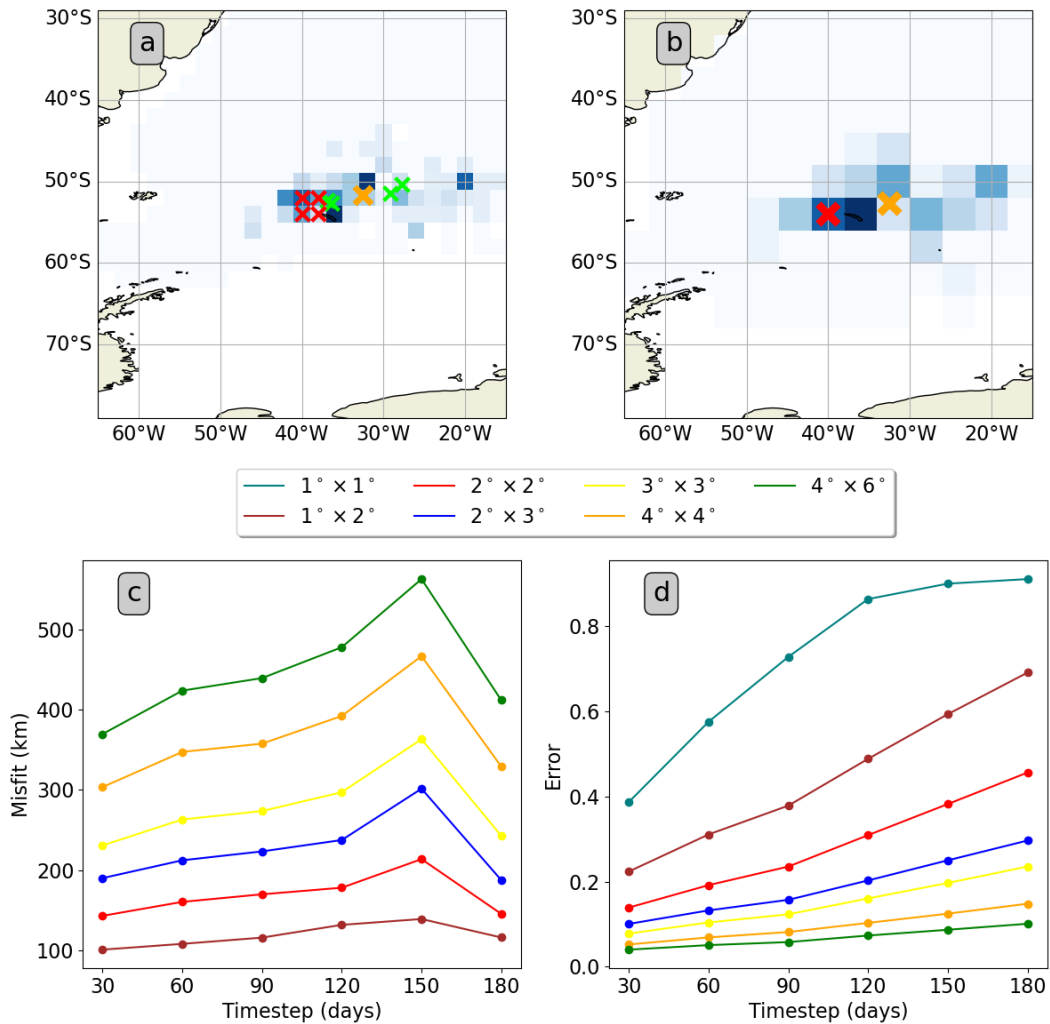


FIG. 5. Examples of 180 day transition probabilities starting from red Xs for (a) $2^\circ \times 2^\circ$ transition matrix and (b) $4^\circ \times 4^\circ$ degree transition matrix. Shading represents transition probability, green Xs represent mean transition from individual grid cells and orange X represents spatial mean of all transition probabilities. (c) Difference in derived mean transition between $1^\circ \times 1^\circ$ derived mean transition and reduced resolution transition matrices derived from $1^\circ \times 1^\circ$ transition statistics. Difference is calculated as the mean difference over all grid cells. (d) Mean of standard error of transition matrices over all grid cells. Colored lines represent different resolutions.

Figure 6 shows standard error (Eq. 6) for two matrices of differing resolution. Spatial area decreases geometrically with increasing resolution, and, broadly speaking, because the Core Argo distribution is homogeneous, the number of Argo trajectories through a grid cell will be proportional to the grid cell size. Unsurprisingly, we see from this example that the lower resolution transition matrix has a lower mean standard error and will produce estimates of the expected value and variance (Equations 4 and 5) with less uncertainty.

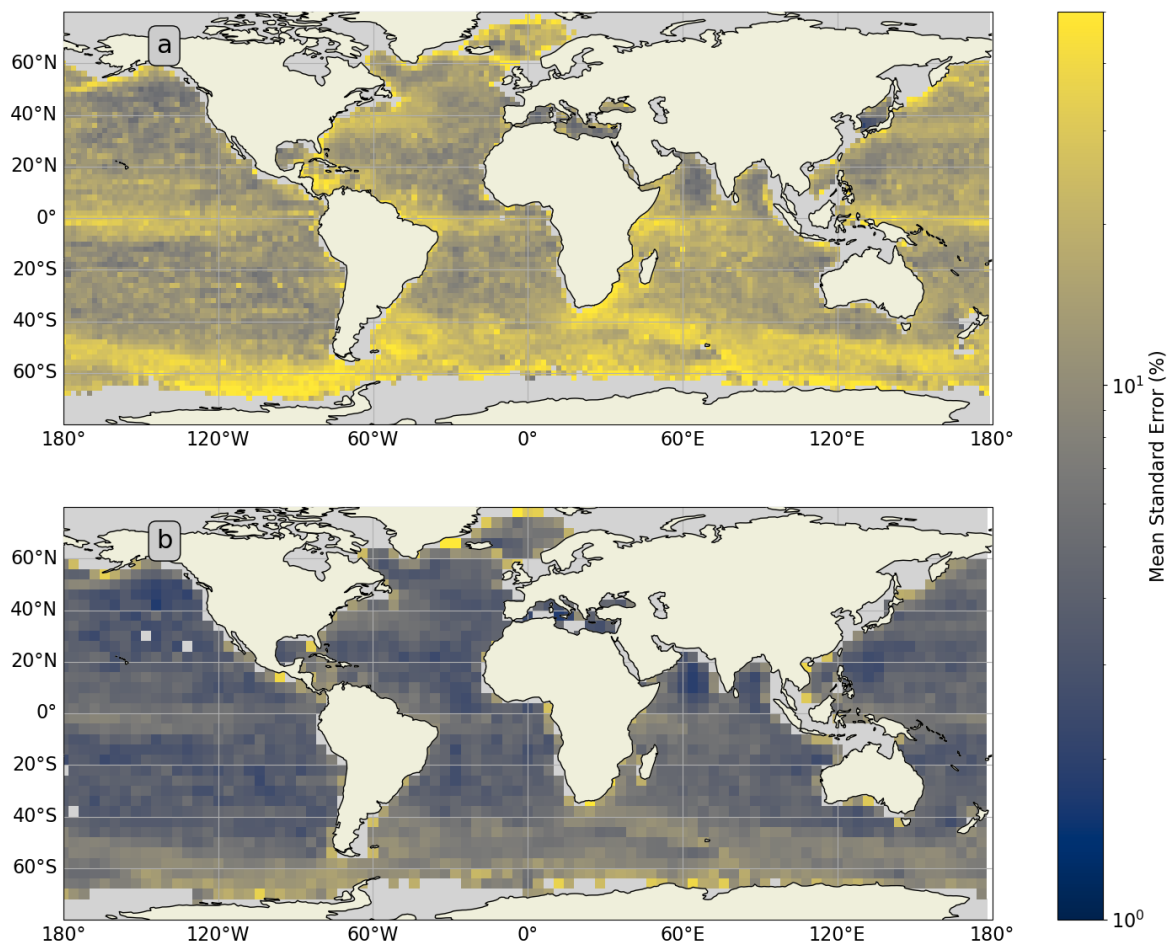


FIG. 6. Standard error for (a) $2^\circ \times 2^\circ$ degree transition matrix and (b) $4^\circ \times 4^\circ$ degree transition matrix at 90 day timestep. Colormap represents standard error; gray shading represents areas outside domain of the transition matrix; beige shading represents land.

For all transition matrices considered, the mean standard error is proportional to resolution and timestep except for the $1^\circ \times 1^\circ$ 30 day transition matrix, which has a lower mean standard error than its $1^\circ \times 2^\circ$ 30 day counterpart (Fig. 5d). This is due to the specific criteria for transition matrix construction that exclude certain high-variance regions from the $1^\circ \times 1^\circ$ transition matrix (namely the southern ACC where trajectory variance is high and data density is low).

Comparing specific matrices and timesteps, we notice that Figure 4c shows substantial misfit improvement between the $1^\circ \times 2^\circ$ and the $2^\circ \times 2^\circ$ resolutions, but slight improvement between the $2^\circ \times 2^\circ$ and the $2^\circ \times 3^\circ$ resolutions, as well as a curvature minimum at the 90-day timestep $2^\circ \times 2^\circ$ transition matrix. Figure 4d shows a misfit minimum for the $2^\circ \times 2^\circ$ resolution transition matrix. Figure 5c shows a misfit plateau between the 60 and 90 timesteps in the $2^\circ \times 2^\circ$ resolution. For these reasons, the $2^\circ \times 2^\circ$ spatial resolution and 90-day timestep is considered the optimal transition matrix configuration and was used in several subsequent calculations. The sensitivity of this transition matrix to data density was tested with a data withholding experiment. Those results are shown in the appendix.

The transition density of each grid cell (N_{blue}^{total} in Figure 2) of the optimal transition matrix dataset (Fig. 7) shows near-global coverage with 8037 grid cells. Transition density suffers primarily in ice-covered regions, where data are sparse, and selection criteria often reject data. The mean number of transitions per grid cell is 76.3 ± 43.9 . The mean transition and variance ellipse of the derived transition matrix (plotted in Fig. 8) resolve the major features of ocean circulation. Regions of large mean transition are also regions of considerable observed kinetic energy. The 3-year mean Lagrangian pathways of each grid cell are computed from the transition matrix (Fig. 9) and show the first statistical moment (Eq. 4) at 12 timesteps of the Markov Chain. The overall pattern of this figure generally resembles the known circulation features in the ocean and closely resembles a composite of the 200 db and 1000 db absolute geostrophic streamfunction maps from Gray and Riser (2014).

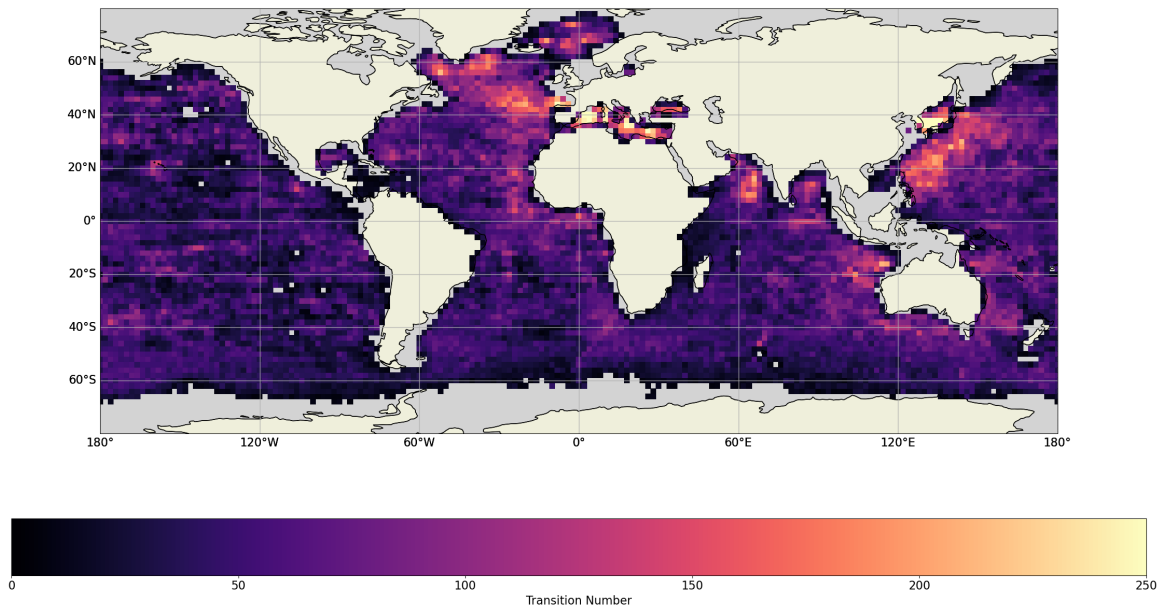


FIG. 7. Transition density based on global Argo trajectories binned in $2^\circ \times 2^\circ$ degree grid cells with a 90 day timestep. Colormap represents transition density of each grid cell; gray shading represents areas outside domain of the transition matrix; beige shading represents land.

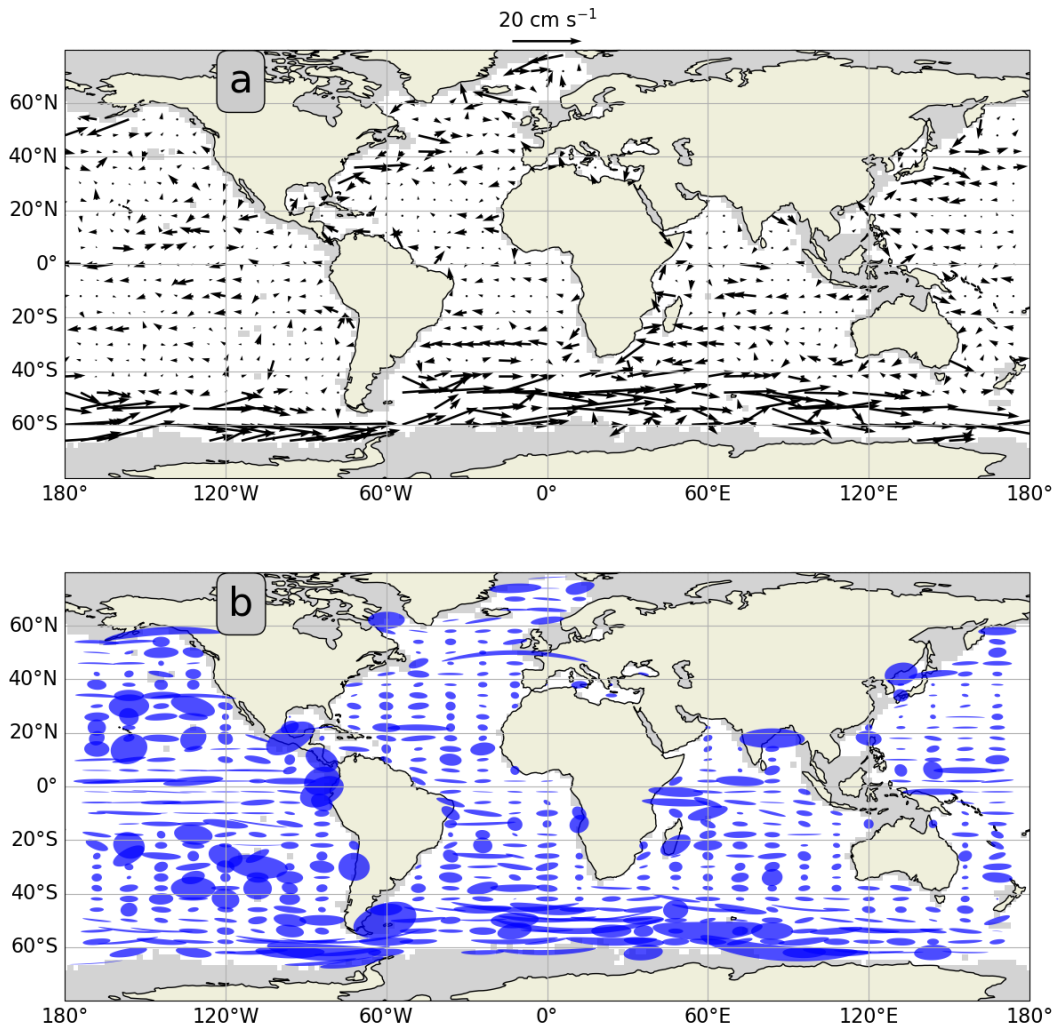


FIG. 8. (a) Mean transition and (b) variance ellipse based on $2^\circ \times 2^\circ$ binned 90 day timestep transition matrix. Arrows represent the mean transition vectors. Blue shapes represent variance ellipses. Gray shading represents areas outside domain of the transition matrix; beige shading represents land.

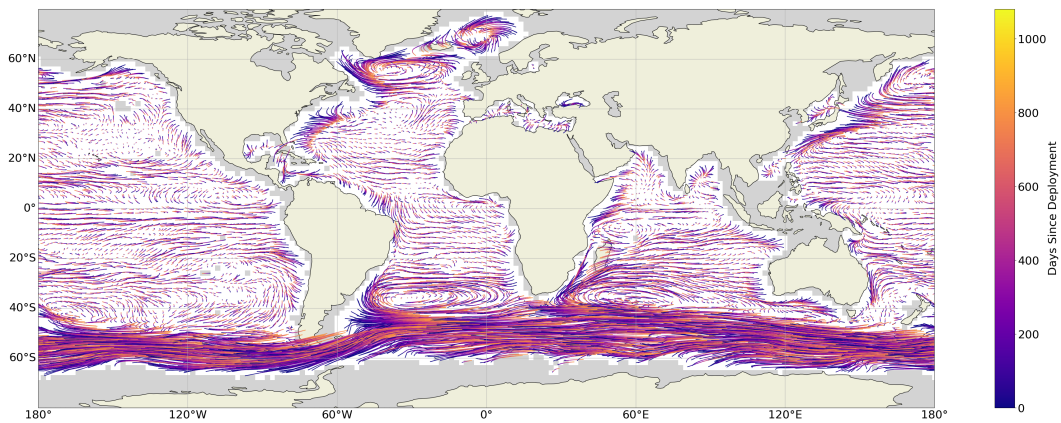


FIG. 9. Mean Lagrangian pathways for Argo floats at each step in 12 step Markov Chain initialized at each grid cell. The Markov Chain is derived from $2^{\circ} \times 2^{\circ}$ grid cell 90 day timestep transition matrix (Eqs. 1 and 4). Colored lines represent time of mean transition. Gray shading represents areas outside domain of the transition matrix; beige shading represents land.

b. Argo Array Prediction

Starting from the actual Argo float distribution of October 10, 2022, with 3262 floats, the transition matrix projected the float array density forward for one and two years. We show the current and projected float spacing of the array in Figure 10 and resulting projected density maps for the Pacific, Atlantic, and Southern Ocean in Figures 11, 12, and 13 respectively. Argo floats older than four years were removed from the array estimate due to the high likelihood of poor sensor performance or float failure, as is a common practice by Argo managers. From these projections: large and growing holes in array distribution exist in the north-central and eastern-equatorial Pacific; sparse distributions exist in the Benguela Current, and the middle of the North Atlantic subtropical gyre; and the Pacific sector of the Southern Ocean will become sparsely observed. Based on this October 2022 example, we would then recommend that Argo deployments prioritize ships transiting or conducting operations in these areas. Such a projection could be performed during each year's Argo planning process.

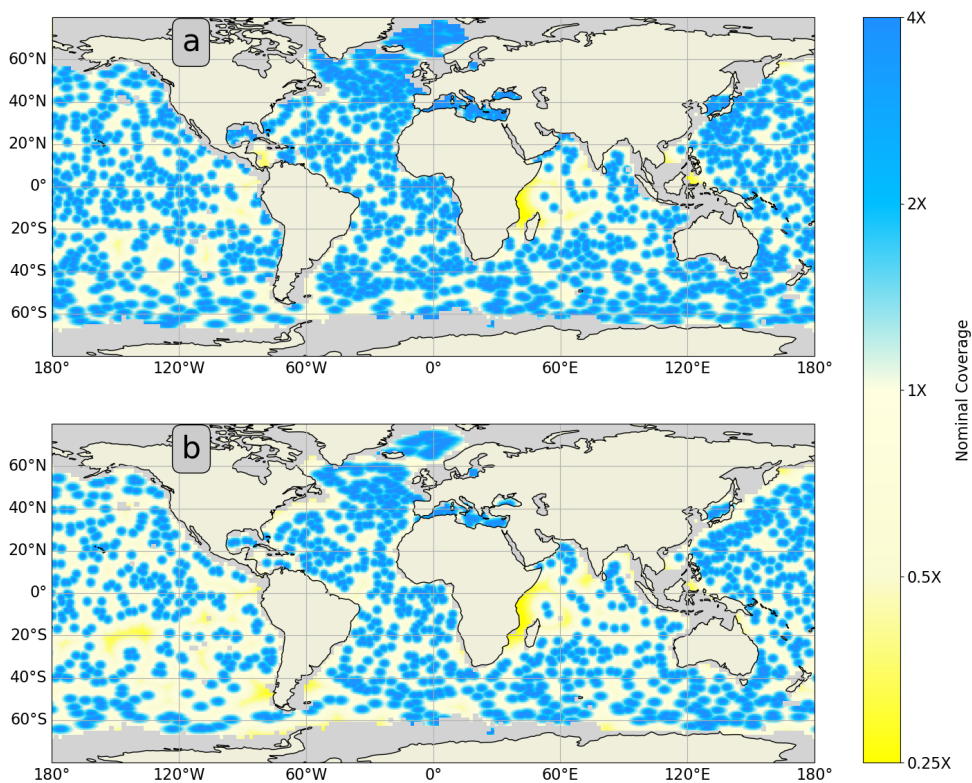


FIG. 10. (a) Current and (b) one year prediction of density of Core Argo array. Prediction based on the Lagrangian pathways of Figure 9 and a float lifetime of 4 years. Colormap is calculated on a $0.5^\circ \times 0.5^\circ$ grid and represents the distance to closest float. 334 km spacing is quantified as "1X" nominal coverage. Beige shading represents land.

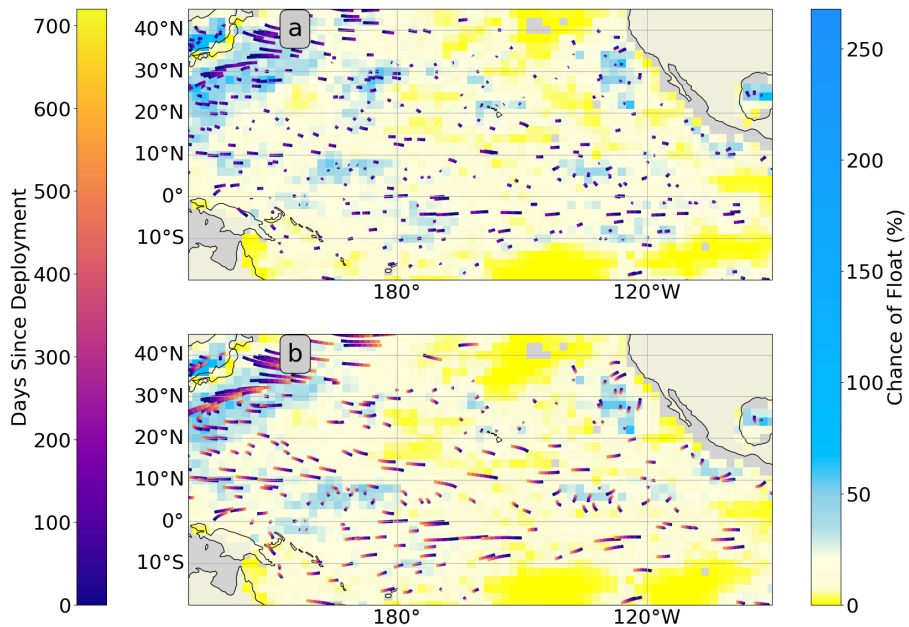


FIG. 11. (a) One year and (b) two year predictions of density of Core Argo array in the Pacific Ocean. Colored lines represent time of mean transition. Grid cell shading represents estimated array density. Gray shading represents areas outside domain of the transition matrix; beige shading represents land.

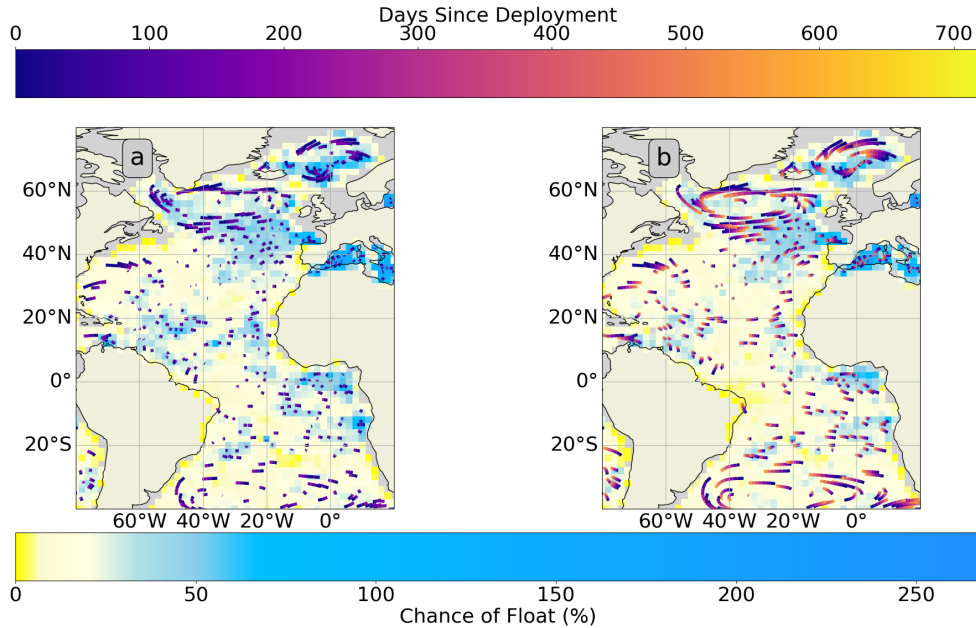


FIG. 12. (a) One year and (b) two year predictions of array density of Core Argo array in the Atlantic Ocean. Colored lines represent time of mean transition. Grid cell shading represents estimated array density. Gray shading represents areas outside domain of the transition matrix; beige shading represents land.

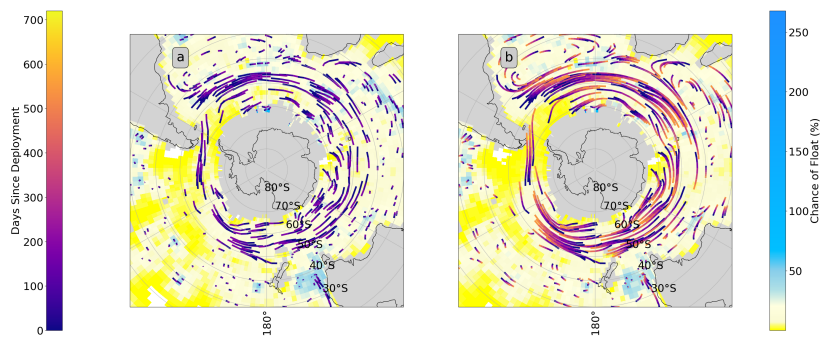


FIG. 13. (a) One year and (b) two year predictions of array density of Core Argo array in the Southern Ocean. Colored lines represent time of mean transition. Grid cell shading represents estimated array density. Gray shading represents areas outside domain of the transition matrix; beige shading represents land.

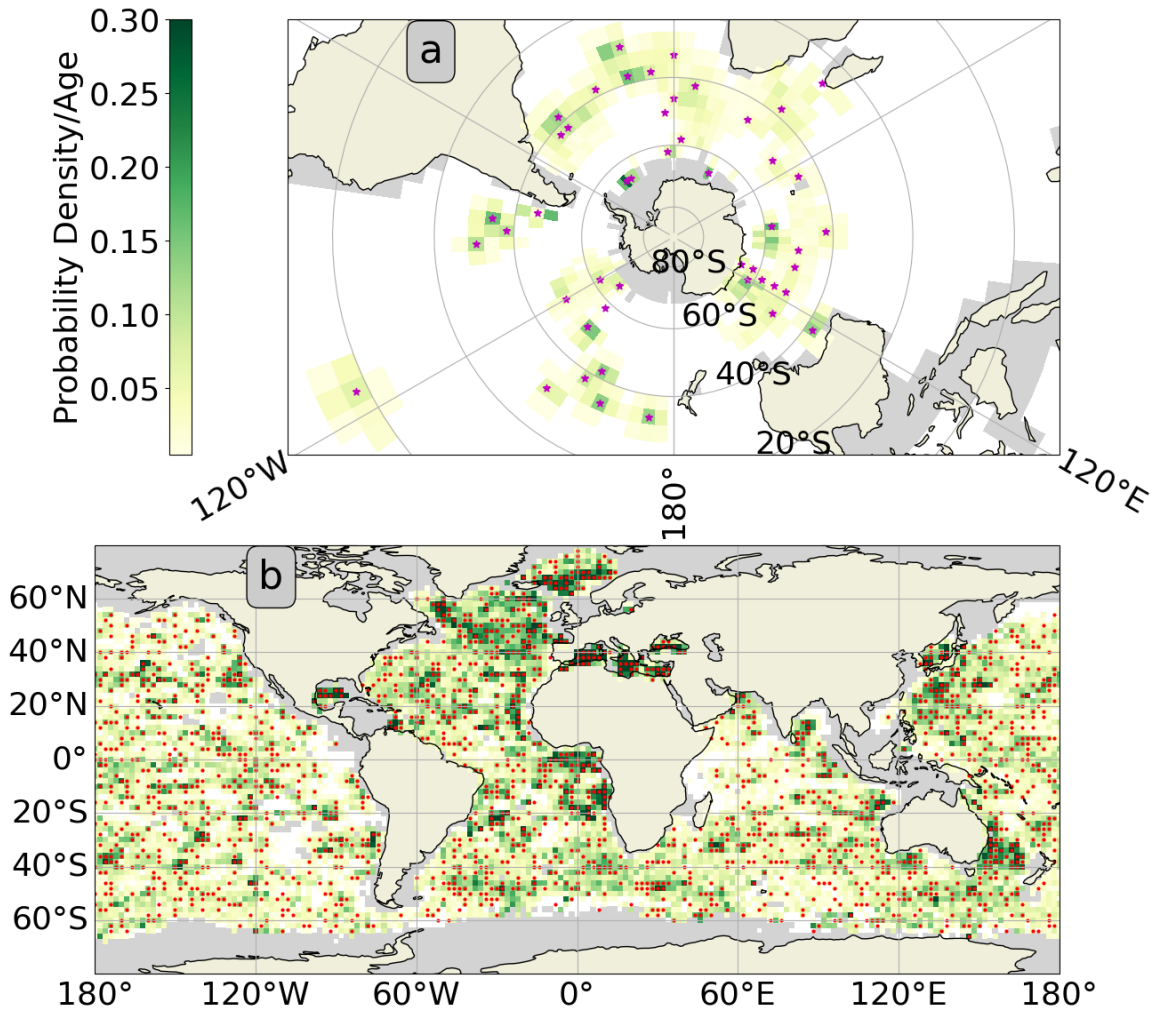


FIG. 14. 180 day projection of (a) SOCCOM and (b) Argo arrays measure of survivability. Colormap represents probability density divided by age. High probability density divided by low age indicates locations where having a surviving float is likely. Magenta stars indicate active SOCCOM floats, red dots indicate active Core Argo floats. Gray shading represents areas outside domain of the transition matrix; beige shading represents land.

As another example of the utility of transition matrices, we demonstrate an improvement of the Density/Age map currently calculated and used by Argo managers (<https://www.ocean->

ops.org/board) as a metric of Core Argo array health. Repopulating old or sparse regions of the network is a goal of Argo managers; the Density/Age map displays the density of Argo floats within a grid cell divided by the average age of those Argo floats. Array health maps use a present snapshot of float distribution and do not estimate the future Density/Age map. Procurement and cruise organization occur many months before putting a float in the water, and estimates of how Argo array health will change in the future could improve planning. The optimal transition matrix can propagate the Density/Age map forward in time to assess the future distribution of the array, which has been done for the Core Argo and SOCCOM arrays (Fig. 14). This example shows Core Argo array health deficits in the Southern Ocean and off the east coast of Africa.

c. Estimating Future Array Density from Float Deployments Along Set Ship Tracks

Research ships of opportunity often deploy Argo floats, with the ship tracks set by other projects. BGC Argo deployments are preferably from projects such as GO-SHIP that provide high-quality biogeochemical data that can be used as a reference to validate BGC sensor calibrations. To determine what fraction of the ocean future GO-SHIP cruises may populate with floats, grid cells containing GO-SHIP cruise tracklines were initialized with floats at the time these GO-SHIP cruises are scheduled to sail (Table 2 and red lines in Figure 15). The optimal transition matrix estimated ocean sampling in the next 5 years based on these deployments. Typically, only a handful of BGC floats are deployed on any cruise. We are showing the greatest possible extent by initializing all grid cells along the trackline.

The resultant sampling densities and mean Lagrangian pathways (Fig. 15) show that in many of the regions where floats could be deployed, they do not travel very far during their lifetimes. Based on these projections, GO-SHIP alone cannot populate the world with floats. Holes will exist in the eastern and western equatorial Pacific, the Gulf of Mexico, the Gulf Stream, and the western tropical Indian Ocean. Moreover, the decadal GO-SHIP transects are not occupied frequently enough to retain optimal sampling density. To achieve uniform distributions of the Core Argo and BGC Argo arrays, these areas will need additional ships of opportunity for deployments.

Cruise Section	Year Planned
I5	2022
A12	2022
SR4	2022
A13.5	2022
P2	2022
P4	2022
I9	2023
I8	2023
A16	2023
P16	2024
P15	2024
S4P	2025
ARC1	2025
P6	2026

TABLE 2. Planned GO-SHIP Cruises

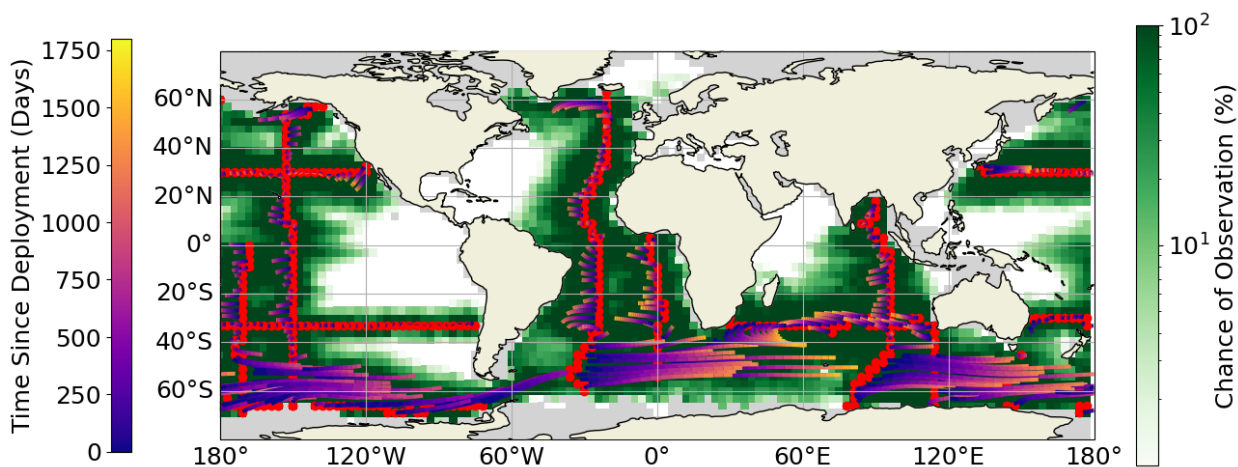


FIG. 15. Chance of observation for floats deployed in each gridcell along ship tracks of GO-SHIP cruises organized for the next 5 years. Colormap represents chance of observation; white shading indicates less than 1 % chance of observation; Colored lines represent time of mean transition. lines represent gray shading represents areas outside domain of the transition matrix; beige shading represents land.

d. Iridium vs. Argos System Communications

Long-term differences in Argos System and Iridium equipped float trajectories are well known due to the difference in surface transmission times (Wong et al. 2020). For the first time, the transition matrix methodology allows us to quantify the implications of the increased surface time of Argos-enabled floats distributed over many profiles. Argo floats transmit data through 2 distinct satellite constellations: Iridium and the Argos System. Floats have the hardware to transmit via one system or the other, but not both. Data transmission is much faster via the Iridium constellation. Consequently, Iridium-enabled floats spend about 15 minutes at the surface compared to their Argos counterparts, which can take up to 12 hours. Surface velocities are also different from velocities at 1000m depth (the typical Argo drift depth), and Argo floats are undrogued and advected by winds and waves while transmitting.

Dividing the full trajectory dataset into only Argos or Iridium enabled trajectories results in significantly less data density for both; we accommodate this by reducing spatial resolution. Transition matrices were constructed using $2^{\circ} \times 3^{\circ}$ grid cells of latitude and longitude and a 180 day timestep. The statistical difference between the Argos and Iridium enabled transition matrices is subtle and could not be distinguished from the null hypothesis by a Z test (Eq. 7).

Transition matrices were multiplied by themselves 15 times to estimate the transition statistics after 8 years - the upper range of current Argo float lifetimes - to highlight the differences in transition statistics. We then uniformly seeded the world ocean with theoretical Argo floats and considered the differences in resultant future float densities predicted with the transition matrix (Fig. 16). In the long-term estimates, the relative density of Iridium-enabled floats stays relatively uniform, and the regional differences in float density do not have a spatial structure consistent with known circulation. In contrast, the Argos System-derived prediction shows strong aggregation in the middle of the subtropical gyres and relative divergence of floats along the equator. This corresponds to divergence in surface currents similar to transition matrices derived from surface drifters (Van Sebille et al. 2012).

This analysis has several implications. Firstly, Argos-enabled floats do not stay on the equator because they are more susceptible to the divergent Ekman transport caused by easterly trade winds. In the Roemmich et al. (2019) vision of the future Argo array, the equator is a prioritized region for increased float density. This analysis supports the decision made by Argo managers that floats

deployed near the equator be equipped with Iridium communications to prevent them from being advected off the equator during their time at the surface. Secondly, Argos- and Iridium-enabled floats move differently, especially on long timescales: the performance of hybrid transition matrices derived from both types of floats may be degraded, primarily in equatorial regions.

For this reason, our data products provide Argos System and Iridium transition matrices separately. However, improved resolution and data density in hybrid transition matrices may offer enhanced performance for shorter-duration predictions. Because the current Argo fleet is composed of Iridium- and Argos-enabled floats, the hybrid matrix has general utility for predicting Argo fleet dynamics. Further, the innovation of new sensors has increased the quantity of data that BGC floats transmit; the increased data requires an average time of an hour at the surface (Riser 2022), and, depending on the sensor suite, traditional Core Argo Iridium-enabled float statistics may underrepresent surface advection.

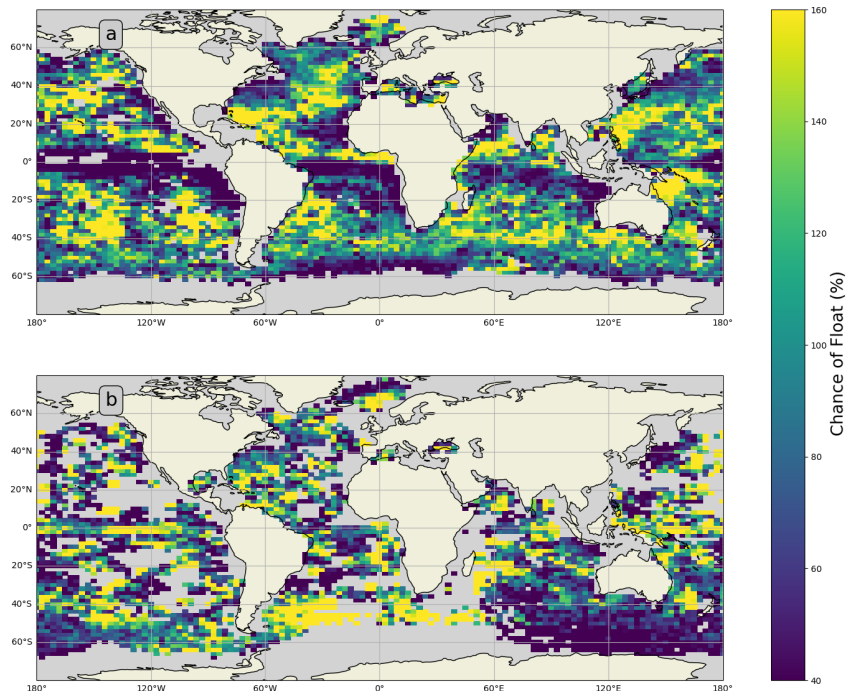


FIG. 16. Comparison of hypothetical future array density for (a) Argos System and (b) Iridium communications derived transition matrices at $2^\circ \times 3^\circ$ grid resolution and 180 day timestep. Transition matrices were propagated forward 8 years with initial array containing one float in each gridcell. Colormap represents future array density. Gray shading represents areas outside domain of the transition matrix; beige shading represents land.

e. BGC Argo Sampling Predictions

Temperature and salinity sensors are ubiquitous within the Argo fleet, but recently developed and more expensive BGC sensors are not. BGC float managers need to know where specific BGC sensors will be when planning deployment cruises.

Motivated by the spatial inhomogeneities of BGC sensors, we estimate the future probability of sampling by sensor class. For these calculations, we used a transition matrix with $3^\circ \times 3^\circ$ grid cells of latitude and longitude and a 90-day timestep to match the designed separation of the Core Argo array (Roemmich et al. 1998). This is a departure from the optimal transition matrix of Section 4a and is used to match maps like Figure 10 commonly used by Argo managers.

Using equation 2, the probability of current Argo sensors sampling in the next year is inhomogeneously distributed by various sensor classes (Figs. 17–18). Temperature and salinity sensors achieve global sampling over the course of a year (Fig 17). Oxygen sensors are the second most widely deployed, with a mean chance of any annual sampling of 35.5% over the spatial domain and no ocean regions omitted (Fig. 18a). Chlorophyll is the third most widely deployed sensor, with a mean chance of any annual sampling of 21.8% over the spatial domain and a potential hole in the northeast Atlantic Ocean (Fig. 18b). Finally, pH is the most sparsely deployed sensor, with a mean chance of sampling any annual sampling of 15.8% over the spatial domain and holes in the Indian and northwest Pacific Oceans (Fig. 18c).

Another important metric to consider is the regions of the ocean that will be sampled year-round. Historically, BGC variables have only been sampled during hydrographic cruises with follow-up cruises years or decades later. Indeed, BGC float observations in the Southern Ocean have led to discoveries about the seasonal variability of BGC variables following fully resolved seasonal observations (Gray et al. 2018). Equation 3 was used to calculate the chance of year-round observation for temperature and salinity observations (Fig 17). Temperature and salinity have the highest probability of year-round sampling, with a mean chance of 44.9% of the domain covered. The BGC array is not yet a fully developed network, and year-round sampling has thus far rarely been achieved. The year-round oxygen sampling has a chance of 1.2% of the domain covered. Chlorophyll and pH have a substantially smaller than 1% chance of year-round sampling in the domain.

Strong currents, such as the ACC, require a uniform density of float coverage to achieve year-round sampling (Davis 1991), and creative methods such as creating regional composites of observations to resolve seasonal signals (Gray et al. 2018) may be necessary for some time.

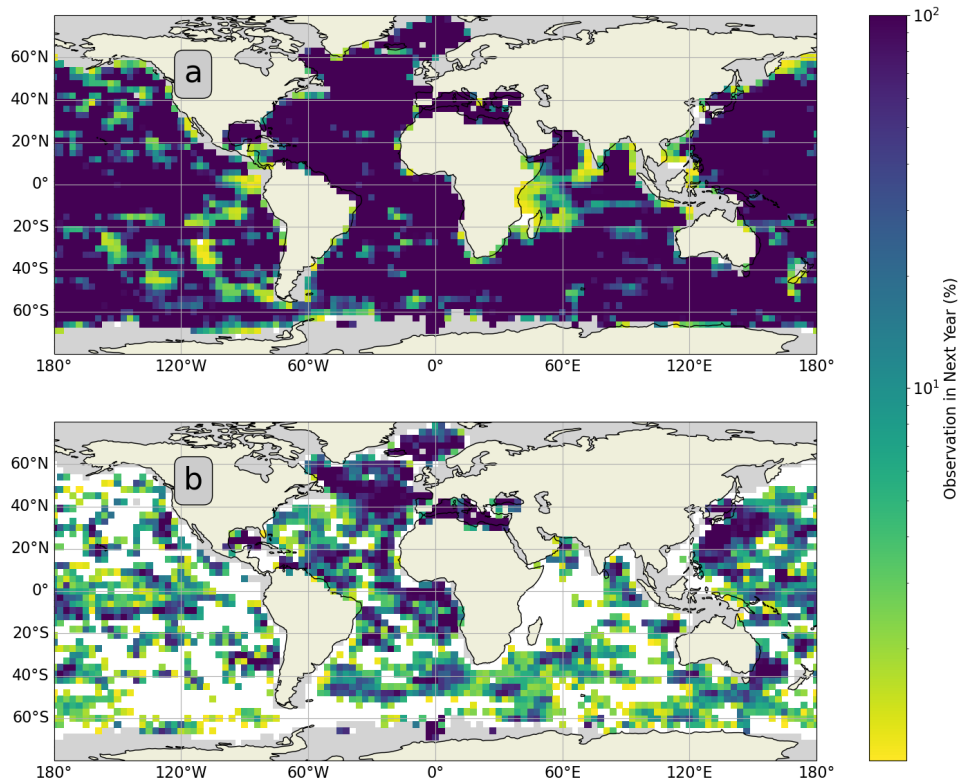


FIG. 17. Chance of (a) any observation within the next year (Eq. 2) and (b) seasonal observation within the next year (Eq. 3) of temperature and salinity sensor equipped Argo floats. Argo distribution is taken from the October 10, 2022 snapshot and is comprised of 3262 floats. Colormap represents chance of observation; gray shading represents areas outside domain of the transition matrix; white shading indicates less than 1 % chance of observation; beige shading represents land. All projections are based on multiples of $3^\circ \times 3^\circ$ 90 day transition matrix.

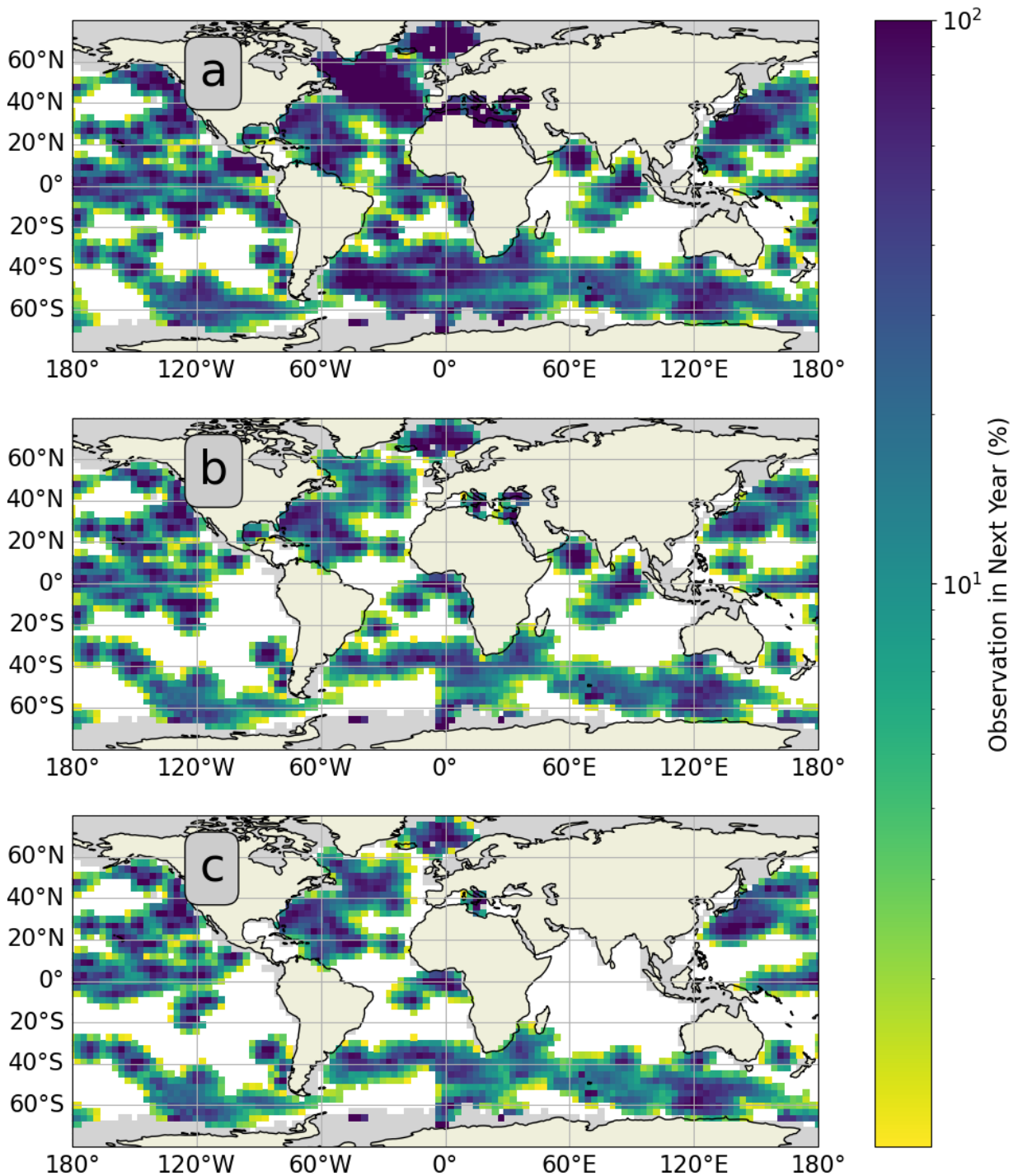


FIG. 18. Chance of any observation within the next year (Eq. 2) of (a) oxygen sensor equipped Argo floats, (b) chlorophyll sensor equipped Argo floats, (c) pH sensor equipped Argo floats. Colormap represents chance of observation; gray shading represents areas outside domain of the transition matrix; white shading indicates less than 1 % chance of observation; beige shading represents land. All projections are based on multiples of $3^\circ \times 3^\circ$ 90 day transition matrix.

f. SOSE Comparison

Models, including the Southern Ocean State Estimate (SOSE), have been used to predict Lagrangian trajectories for both operational (Talley et al. 2019) and scientific (Tamsitt et al. 2017) applications. However, models generally do not reproduce Argo float dispersion well, even when the Lagrangian particles simulate the full Argo 10-day cycle (Talley et al. 2019). As a validation for both the SOSE model and the transition matrices, we compared transition matrices derived from SOSE model-based trajectories with the Argo-derived transition matrices.

Transition matrices composed of $1^\circ \times 2^\circ$ and $4^\circ \times 6^\circ$ grid cells of latitude and longitude with a 180-day timestep recreated the upper and lower limits of available grid resolution for this region. Across both resolution levels, the zonal mean transitions in the ACC were greater in the SOSE-derived matrices. The $1^\circ \times 2^\circ$ matrix comparison had a 1 cm s^{-1} increase in the ACC while the 4×6 comparison had a 0.5 cm s^{-1} increase. The second moment was also compared (Eq. 5, Fig. 19) and shows that SOSE consistently under-represents ACC Lagrangian diffusion in the high-resolution case with a mean difference of $-6.8 \times 10^{-1} \pm 6.2 \times 10^{-1} \text{ cm}^2 \text{ s}^{-2}$, and well resolves ACC diffusion in the low-resolution case with a mean difference of $0.00 \pm 5.8 \times 10^{-1} \text{ cm}^2 \text{ s}^{-2}$.

SOSE is an eddy-permitting model but does not have sufficient resolution to fully resolve high-latitude eddies. This analysis suggests that SOSE ACC kinetic energy, at high resolution, is concentrated in the mean flow and does not sufficiently cascade into smaller-scale features; this manifests in lower diffusivity. The low-resolution case seems to have the effect of smoothing these differences. From this analysis, scientific conclusions derived from SOSE Lagrangian particle statistics should only be considered accurate for coarse-resolution studies. Changes in parameterized diffusivities in the offline Lagrangian model Octopus could potentially address these problems, but these changes have not been studied. A higher resolution $1/12^\circ$ grid cell solution is now also available and may improve mesoscale statistics, but this has not been tested.

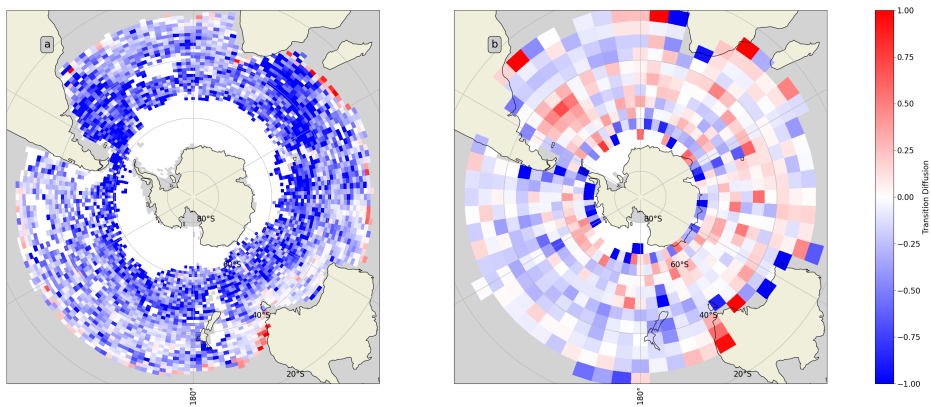


FIG. 19. Difference of standard deviation around the mean of SOSE and Argo derived matrices for (a) $1^\circ \times 2^\circ$ and (b) $4^\circ \times 6^\circ$ grid cells with 180 day timestep. Shading represents SOSE - Argo standard deviation of transition.

5. Conclusion

In this paper, we have explained, justified, and tested the construction of a transition matrix for Argo float location prediction, following the surface drifter work by Van Sebille et al. (2012). Our work is in the broader context of BGC Argo global array design, and a companion paper will describe an optimal float deployment algorithm.

After quantifying a wide range of temporal and spatial biases and uncertainties, we have concluded that using the available Argo trajectory data, the transition matrix constructed from a $2^\circ \times 2^\circ$ spatial resolution at a 90-day timestep is optimal. This transition matrix is used for: Core Argo predictions, GO-SHIP deployment predictions, and array health products. A description of publicly available web applications and code repositories to predict future Argo float locations with the transition matrix can be found in the supplement.

We will update the transition matrix as more trajectory data are made available. The present transition matrix is available as supplemental material for this paper. This transition matrix is a hybrid of Argos- and Iridium-enabled float trajectories.

We have shown a significant difference in transition matrices derived from floats equipped with different communication systems. We recommend that floats deployed in equatorial waters use Iridium communications. We additionally provide supplementary material for the Argos and Iridium transition matrices for investigators who wish to make specific predictions based on communication type.

Finally, we compared the Argo transition matrices to transition matrices derived from a particle release experiment in the SOSE model. We found that the overall mean particle transition in the ACC was greater in the SOSE transition matrix, and particle diffusion was too low in the SOSE transition matrix at high resolution but consistent with the Argo-derived transition matrix at low resolution. We hypothesize that SOSE does not fully resolve the mesoscale eddy field.

The ever-growing Argo float dataset will continue to improve both the statistical accuracy and resolution of transition matrices. However, we find that the array is already of sufficient size for transition matrix construction, enabling significant insights into difficult questions that BGC Argo managers face now. BGC Argo floats offer new technology to answer questions of critical societal importance. We hope the transition matrix tools presented here will contribute to the ongoing community conversation regarding optimal array design.

Acknowledgments. This work was supported by the SOCCOM project under NSF Award PLR-1425989, the Global Ocean Biogeochemical Array (GO-BGC) under NSF Award OCE-1946578, and the Hypernav project under NASA Award 80GSFC20C0101. Co-authors acknowledge NSF EarthCube Award 1928305. EvS was supported by the Netherlands Organization for Scientific Research (NWO), Earth and Life Sciences, through project OCENW.KLEIN.085. We thank Dr. Isa Rosso for providing SOSE Lagrangian particle trajectories. Prof. Donata Giglio (PI of the Argovis project) and Dr. William Mills (Argovis software engineer) at University of Colorado Boulder worked on including the product ARGONE (described in this paper) on the web app and database Argovis, including designing a new schema and API for the data and creating a demo notebook (https://github.com/argovis/demo_notebooks) to create panel b in Figure 1 of the paper. Prof. Donata Giglio, Dr. William Mills, and Tyler Tucker were supported by NSF awards 1928305 and 2026954. We also thank two anonymous reviewers for their insightful comments and suggestions.

Data availability statement. The Argo Program (<https://argo.ucsd.edu>, <http://argo.jcommops.org>) is part of the Global Ocean Observing System. 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. GO-SHIP cruise waypoints were collected from the CCHDO website (<http://cchdo.ucsd.edu/>). All code used in this publication is publicly available ([https://github.com/Chamberpain/TransitionMatrix\(Chamberlain2023b\)](https://github.com/Chamberpain/TransitionMatrix(Chamberlain2023b))).

APPENDIX

a. transition matrix Web Applications and Repositories

Argo float location prediction is an observing system priority. Because of this, we worked with the Argovis team at University of Colorado Boulder to add the product ARGONE to the Argovis web app and database (https://github.com/argovis/demo_notebooks, recently upgraded from the app version described in Tucker et al. 2020). A demo notebook leveraging the new Argovis API to access the product ARGONE (described in this paper) and predict Argo float

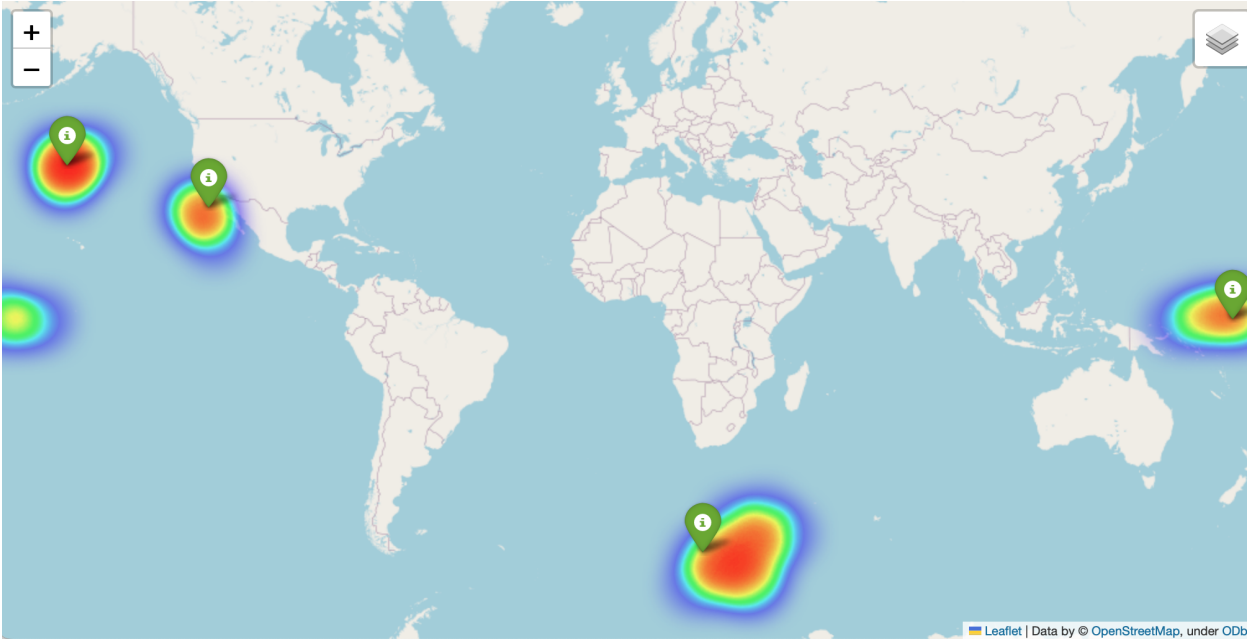


FIG. A1. Example of ARGONE repository output (<https://github.com/Chamberpain/ARGONE> (Chamberlain 2023a)) for example float array. Green markers represent initial Argo float array locations, colored grid cells represents Argo float probability density function after 180 days with initial deployment at float icon.

locations is available at https://github.com/argovis/demo_notebooks and the product will also be featured on the web app front end in the future.

The Argovis web app serves the statistical prediction of Argo float locations (using ARGONE) up to about 5 years in the future (the target float lifetime). Figure 1 shows an example of accessing ARGONE through Argovis.

The ARGONE GitHub repository (<https://github.com/Chamberpain/ARGONE>(Chamberlain 2023a)) is publicly available and produces future probabilities of a float array. Figure A1 shows an example of these results.

b. Data Withholding Experiment

To test the data sensitivity of the recommended $2^\circ \times 2^\circ$ grid cell at 90 day timestep transition matrix, we performed a data withholding experiment. The data withholding experiment compared transition matrices created from subsets of the Argo trajectory dataset with the transition matrix made from the full Argo trajectory dataset. The subsets of the Argo trajectory database were

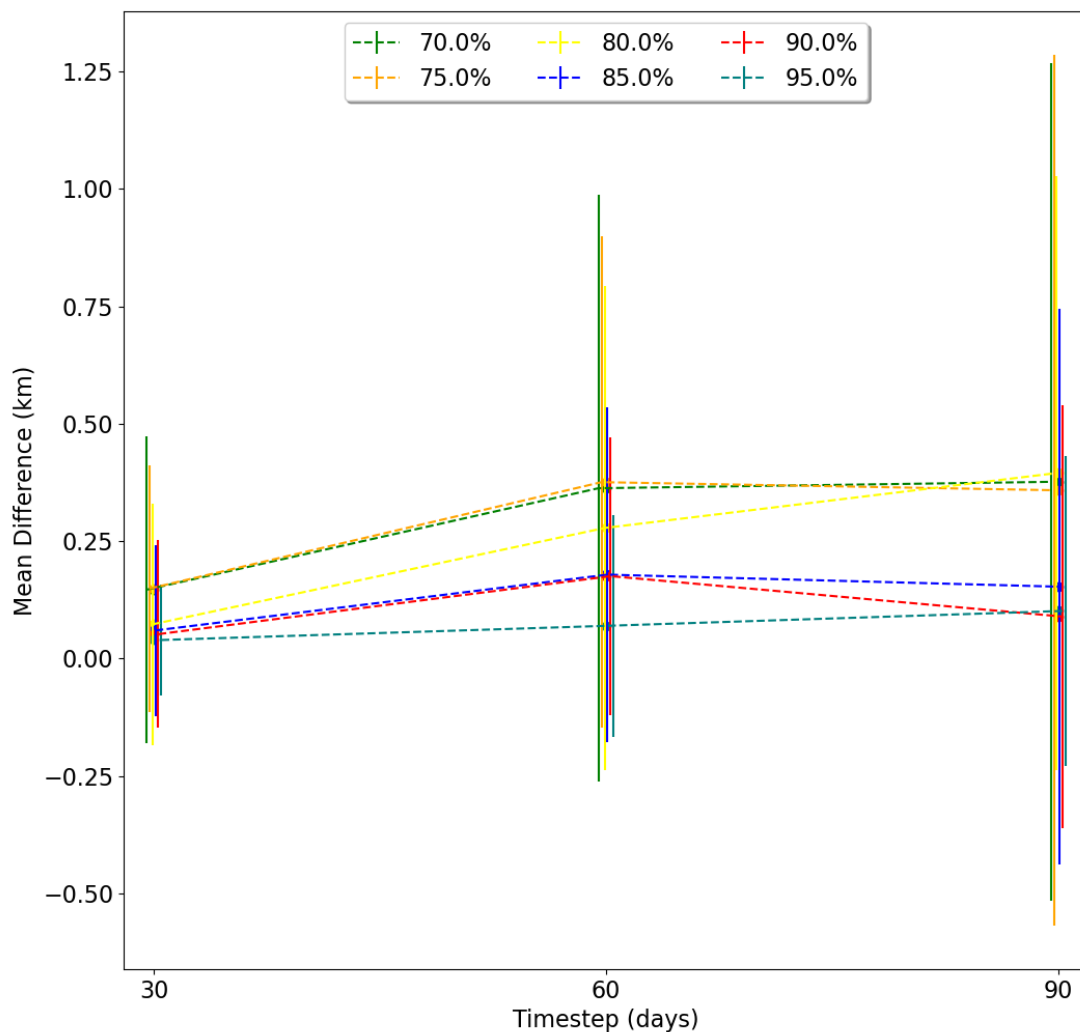


FIG. A2. Results of data withholding experiment. Transition matrices were calculated from subsets of the Argo trajectory data and compared to the original transition matrix. The mean and standard deviation of the differences are shown by the dashed and solid lines respectively. Colored lines represent different data densities.

generated by randomly withholding floats from the full Argo database. The dependence on data density was tested by increasing the number of withheld floats from 5% to 30% of the total number of Argo floats in 5% increments. 30, 60, and 90 day transition matrices generated from these

randomly generated subsets were compared to the original transition matrices 50 times at each data density. The difference in mean transition (Eqn 4) between each data withheld matrix and the original transition matrix was calculated for every grid cell (Fig. A2)

There exist 2 distributions in the mean differences at 5-15% withheld and 15-30% withheld. These distributions have a mean difference of 0.37 km and 0.11 km, respectively for the 90-day transition matrix. The small difference in the 85-95 % distribution suggests that the mean of the full transition matrix may not be significantly sensitive to new data. The standard deviation of the difference is inversely proportional to the data density and larger than the mean difference for all cases considered.

References

- Abernathey, R., C. Bladwell, G. Froyland, and K. Sakellariou, 2022: Deep Lagrangian Connectivity in the Global Ocean Inferred from Argo Floats. *Journal of Physical Oceanography*, **52** (5), 951–963.
- Argo, 2022: Argo Float Data and Metadata from Global Data Assembly Centre (Argo GDAC) - Snapshot of Argo GDAC of October 10st 2022. *SEANOE*, <https://doi.org/10.17882/42182#96550>.
- Boebel, O., J. Lutjeharms, C. Schmid, W. Zenk, T. Rossby, and C. Barron, 2003: The Cape Cauldron: A Regime of Turbulent Inter-Ocean Exchange. *Deep Sea Research Part II: Topical Studies in Oceanography*, **50** (1), 57–86.
- Chamberlain, P., 2023a: Chamberpain/argone: v1.0.0. Zenodo, URL <https://doi.org/10.5281/zenodo.7623074>, <https://doi.org/10.5281/zenodo.7623074>.
- Chamberlain, P., 2023b: Chamberpain/transitionmatrix: v1.0.0. Zenodo, URL <https://doi.org/10.5281/zenodo.7623067>, <https://doi.org/10.5281/zenodo.7623067>.
- Chamberlain, P., L. Talley, B. Cornuelle, M. Mazloff, and S. Gille, 2023: Optimizing The Biogeochemical Argo Float Distribution. *Journal of Atmospheric and Oceanic Technology*.
- Davis, R., L. Regier, J. Dufour, and D. Webb, 1992: The Autonomous Lagrangian Circulation Explorer (ALACE). *Journal of atmospheric and oceanic technology*, **9** (3), 264–285.

- Davis, R. E., 1991: Observing the General Circulation with Floats. *Deep Sea Research Part A. Oceanographic Research Papers*, **38**, S531–S571.
- Drouin, K. L., M. S. Lozier, F. J. Beron-Vera, P. Miron, and M. J. Olascoaga, 2022: Surface Pathways Connecting the South and North Atlantic Oceans. *Geophysical Research Letters*, **49** (1), e2021GL096646.
- Ford, D., 2020: Assimilating Synthetic Biogeochemical-Argo and Ocean Colour Observations into a Global Ocean Model to Inform Observing System Design. *Biogeosciences Discussions*, **2020**, 1–34, <https://doi.org/10.5194/bg-2020-152>, URL <https://www.biogeosciences-discuss.net/bg-2020-152/>.
- Freeland, H., 1975: Statistical Observations of the Trajectories of Neutrally Buoyant Floats in the North Atlantic. *J. Mar. Res.*, **33**, 383–404.
- Froyland, G., R. M. Stuart, and E. van Sebille, 2014: How Well-Connected is the Surface of the Global Ocean? *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **24** (3), 033126.
- Gille, S. T., and L. Romero, 2003: Statistical Behavior of ALACE Floats at the Surface of the Southern Ocean. *Journal of Atmospheric and Oceanic Technology*, **20** (11), 1633–1640.
- Gray, A. R., and S. C. Riser, 2014: A Global Analysis of Sverdrup Balance Using Absolute Geostrophic Velocities from Argo. *Journal of Physical Oceanography*, **44** (4), 1213–1229.
- Gray, A. R., and Coauthors, 2018: Autonomous Biogeochemical Floats Detect Significant Carbon Dioxide Outgassing in the High-Latitude Southern Ocean. *Geophysical Research Letters*, **45** (17), 9049–9057.
- Johnson, K., and H. Claustre, 2016: The Scientific Rationale, Design, and Implementation Plan for a Biogeochemical-Argo Float Array. *Biogeochem.-Argo Plann. Group*, **58**.
- Johnson, K. S., and Coauthors, 2017: Biogeochemical Sensor Performance in the SOCCOM Profiling Float Array. *Journal of Geophysical Research: Oceans*.
- Kamenkovich, I., A. Haza, A. R. Gray, C. O. Dufour, and Z. Garraffo, 2017: Observing System Simulation Experiments for an Array of Autonomous Biogeochemical Profiling Floats in the Southern Ocean. *Journal of Geophysical Research: Oceans*, **122** (9), 7595–7611.

- Lumpkin, R., L. Centurioni, and R. C. Perez, 2016: Fulfilling Observing System Implementation Requirements with the Global Drifter Array. *Journal of Atmospheric and Oceanic Technology*, **33** (4), 685–695.
- Majkut, J. D., B. R. Carter, T. L. Frölicher, C. O. Dufour, K. B. Rodgers, and J. L. Sarmiento, 2014: An Observing System Simulation for Southern Ocean Carbon Dioxide Uptake. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **372** (2019), 20130046.
- Markov, A. A., 1906: Rasprostranenie Zakona Bol'shikh Chisel na Velichiny, Zavisyaschie Drug ot Druga. *Izvestiya Fiziko-matematicheskogo obshchestva pri Kazanskom universitete*, **15** (135-156), 18.
- Maximenko, N., J. Hafner, and P. Niiler, 2012: Pathways of Marine Debris Derived from Trajectories of Lagrangian Drifters. *Marine pollution bulletin*, **65** (1-3), 51–62.
- Mazloff, M. R., P. Heimbach, and C. Wunsch, 2010: An Eddy-Permitting Southern Ocean State Estimate. *Journal of Physical Oceanography*, **40** (5), 880–899.
- McAdam, R., and E. van Sebille, 2018: Surface Connectivity and Inter-Ocean Exchanges From Drifter-Based Transition Matrices. *Journal of Geophysical Research: Oceans*.
- Miron, P., F. Beron-Vera, and M. Olascoaga, 2022: Transition Paths of North Atlantic Deep Water. *Journal of Atmospheric and Oceanic Technology*.
- Miron, P., F. J. Beron-Vera, M. J. Olascoaga, G. Froyland, P. Pérez-Brunius, and J. Sheinbaum, 2019: Lagrangian Geography of the Deep Gulf of Mexico. *Journal of Physical Oceanography*, **49** (1), 269–290.
- Morrison, A. K., T. L. Frölicher, and J. L. Sarmiento, 2015: Upwelling in the Southern Ocean. *Physics Today*, **68** (1), 27.
- Ollitrault, M., and J.-P. Rannou, 2013: ANDRO: An Argo-Based Deep Displacement Dataset. *Journal of Atmospheric and Oceanic Technology*, **30** (4), 759–788.
- Riser, S., 2022: personal communication.

- Roemmich, D., and Coauthors, 1998: On the Design and Implementation of Argo: an Initial Plan for a Global Array of Profiling Floats. *International CLIVAR Project Office Report*, **21**.
- Roemmich, D., and Coauthors, 2019: On the Future of Argo: A Global, Full-Depth, Multi-Disciplinary Array. *Frontiers in Marine Science*, **6**, 439.
- Sevellec, F., A. Colin de Verdière, and M. Ollitrault, 2017: Evolution of Intermediate Water Masses Based on Argo Float Displacements. *Journal of Physical Oceanography*, **47** (7), 1569–1586.
- Talley, L., and Coauthors, 2019: Southern Ocean Biogeochemical Float Deployment Strategy, With Example From the Greenwich Meridian Line (GO-SHIP A12). *Journal of Geophysical Research: Oceans*, **124** (1), 403–431.
- Tamsitt, V., and Coauthors, 2017: Spiraling Pathways of Global Deep Waters to the Surface of the Southern Ocean. *Nature communications*, **8** (1), 1–10.
- Tucker, T., D. Giglio, M. Scanderbeg, and S. S. Shen, 2020: Argovis: a Web Application for Fast Delivery, Visualization, and Analysis of Argo Data. *Journal of Atmospheric and Oceanic Technology*, **37** (3), 401–416.
- Van Sebille, E., M. H. England, and G. Froyland, 2012: Origin, Dynamics and Evolution of Ocean Garbage Patches from Observed Surface Drifters. *Environmental Research Letters*, **7** (4), 044 040.
- Van Sebille, E., P. J. Van Leeuwen, A. Biastoch, and W. P. de Ruijter, 2010: On The Fast Decay of Agulhas Rings. *Journal of Geophysical Research: Oceans*, **115** (C3).
- Van Sebille, E., and Coauthors, 2018: Lagrangian Ocean Analysis: Fundamentals and Practices. *Ocean Modelling*, **121**, 49–75.
- Wong, A. P., and Coauthors, 2020: Argo Data 1999–2019: Two Million Temperature-Salinity Profiles and Subsurface Velocity Observations from a Global Array of Profiling Floats. *Frontiers in Marine Science*, **7**, 700.