Spatiotemporal Topic Modeling Reveals Storm-Driven Advection and Stirring Control Plankton Community Variability in an Open Ocean Eddy

John Edward San Soucie[1], Heidi M Sosik[2], Yogesh Girdhar[1], Alexi Shalapyonok[2], Emily Peacock[1], and Leah Johnson[3]

[1]Woods Hole Oceanographic Institution
[2]WHOI
[3]Brown University

April 26, 2024

Abstract

Phytoplankton communities in the open ocean are high-dimensional, sparse, and spatiotemporally heterogeneous. The advent of automated imaging systems has enabled high-resolution observation of these communities, but the amounts of data and their statistical properties make analysis with traditional approaches challenging. Spatiotemporal topic models offer an unsupervised and interpretable approach to dimensionality reduction of sparse, high-dimensional categorical data. Here we use topic modeling to analyze neural-network-classified phytoplankton imagery taken in and around a retentive eddy during the 2021 North Atlantic EXport Processes in the Ocean from Remote Sensing (EXPORTS) field campaign. We investigate the role physical-biological interactions play in altering plankton community composition within the eddy. Analysis of a water mass mixing framework suggests that storm-driven surface advection and stirring were major drivers of the progression of the eddy plankton community away from a diatom bloom over the course of the cruise.

Spatiotemporal Topic Modeling Reveals Storm-Driven Advection and Stirring Control Plankton Community Variability in an Open Ocean Eddy

John E. San Soucie[1,3], Yogesh Girdhar[1], Leah Johnson[4], Emily E. Peacock[2], Alexei Shalapyonok[2], Heidi M. Sosik[2]

[1]Woods Hole Oceanographic Institution, Applied Ocean Physics and Engineering Department
[2]Woods Hole Oceanographic Institution, Biology Department
[3]Massachusetts Institute of Technology, Department of Mechanical Engineering
[4]University of Washington, Applied Physics Laboratory

Key Points:

- Topic models provide a robust alternative to traditional statistical techniques for analysis of sparse, high-dimensional categorical data.
- We perform topic model analyses of machine-classified plankton images taken near a retentive eddy during the 2021 EXPORTS NA field campaign.
- Surface advection and stirring during storms controlled the surface plankton community of the eddy as it transitioned out of a diatom bloom.

Corresponding author: John E. San Soucie, `jsansoucie@whoi.edu`

Abstract

Phytoplankton communities in the open ocean are high-dimensional, sparse, and spatiotemporally heterogeneous. The advent of automated imaging systems has enabled high-resolution observation of these communities, but the amounts of data and their statistical properties make analysis with traditional approaches challenging. Spatiotemporal topic models offer an unsupervised and interpretable approach to dimensionality reduction of sparse, high-dimensional categorical data. Here we use topic modeling to analyze neural-network-classified phytoplankton imagery taken in and around a retentive eddy during the 2021 North Atlantic EXport Processes in the Ocean from Remote Sensing (EXPORTS) field campaign. We investigate the role physical-biological interactions play in altering plankton community composition within the eddy. Analysis of a water mass mixing framework suggests that storm-driven surface advection and stirring were major drivers of the progression of the eddy plankton community away from a diatom bloom over the course of the cruise.

Plain Language Summary

Plankton communities in the ocean can have many different species, with large differences in their abundance and patchy distributions in space. Automated imaging systems allow for high-resolution observation of these plankton communities, but many traditional statistical techniques fail to capture their full complexity. Spatiotemporal topic models, a kind of statistical model designed to work directly with categorical data, can effectively represent this kind of information. In this work, we use topic models to analyze plankton images taken near an eddy in the spring of 2021 and classified into 50 different kinds of plankton with an automated algorithm. We investigate how interactions between ocean physics and biology can change the plankton community inside the eddy. Analysis suggests that storms in the area moved surface water carrying a different plankton community into the eddy.

1 Introduction and background

Marine plankton communities are highly dynamic (Ryther, 1969), with impacts from short- (Mahadevan, 2016) and long-scale (Raitsos et al., 2014) ocean physics, weather (Fiorendino et al., 2021) and climate (Henson et al., 2021), chemical presence (Ianora et al., 2011) and nutrient availability (Barcelos e Ramos et al., 2017), and biological interactions (Banse, 1994). In turn, plankton populations have major impacts on the entire marine food web (Frederiksen et al., 2006), commercial fishing and aquaculture (Brown et al., 2020), and ocean carbon fluxes (Guidi et al., 2016). Understanding how plankton communities respond to external disturbance is therefore crucial for economic and climate forecasting efforts.

In the Northeast Atlantic, which has a strong and energetic eddy field and experiences vigorous wintertime convection, ocean physics plays an important role in mediating phytoplankton community dynamics on a wide range of spatiotemporal scales. Interannually, the North Atlantic Oscillation may impact the community balance between diatoms and dinoflagellates (Henson et al., 2012; Allen et al., 2020). Seasonally, the onset of spring diatom blooms has been linked to a shutdown of wintertime convection (Taylor & Ferrari, 2011) along with solar- (Sverdrup, 1953) and eddy-induced (Mahadevan et al., 2010, 2012) restratification. In addition to temporal changes, the existence of an energetic eddy field creates horizontal heterogeneity and patchiness in phytoplankton productivity and type (Martin, 2003; Lévy & Martin, 2013).

Many approaches for characterizing plankton communities are too low-resolution — either spatiotemporally or in terms of the compositional information acquired —

to fully resolve important internal and external dynamics in marine ecosystems. Genomic data from seawater sampled via bottle casts on a ship is limited by the deployment frequency of the sampling rosette (hours). Conversely, bulk property sensor deployments such as fluorometers on profiling moorings can provide high-frequency measurements but lack fine plankton composition resolution.

In contrast, automated imaging techniques can sample at high temporal resolution, with enough detail to resolve relevant taxonomic distinctions. The Imaging FlowCytobot (IFCB) (Olson & Sosik, 2007) uses flow cytometry integrated with video imaging to detect phytoplankton cells in seawater samples. The IFCB typically samples automatically two to three times per hour, generating thousands of plankton images per sample. Due to the high temporal resolution and information density, full manual review of IFCB datasets is impractical. Instead, classification typically proceeds with machine learning-based classifiers. Ecologically relevant classification of IFCB images with machine learning algorithms such as convolutional neural networks (CNNs) has been well documented (Olson & Sosik, 2007; Olson et al., 2017; Campbell et al., 2010; Catlett et al., 2023; Peacock et al., 2014).

The plankton community composition dynamics which generate image time series are nonlinear, with high-dimensional and spatially heterogeneous (patchy) communities. These properties make data analysis challenging. Statistical tools such as Principal Component Analysis (PCA) and (non-)metric Multidimensional Scaling (NMDS and MDS) greatly reduce the dimensionality of the data while preserving part of the higher-dimensional structures and patterns. But some of these tools make unrealistic assumptions about how data are generated. For example, PCA assumes that observations decompose into real-valued weightings of orthogonal eigenvectors, but actual underlying trends in communities need not be orthogonal. Other tools, like (N)MDS, may not make any generative assumptions at all, and provide a purely descriptive approach to dimensionality reduction.

Topic models offer an approximate but robust and interpretable alternative to classical dimensionality reduction approaches. Topic models are a class of Bayesian graphical model that factor the distribution of categorical observations with latent "topics", which themselves represent distributions over observation categories. A key early topic model, the Latent Dirichlet Allocation model (David M. Blei et al., 2003), was originally used to model text documents. With a Bayesian inference algorithm, the Latent Dirichlet Allocation model converges on topics with semantic meaning, organized by co-occuring clusters of words. The Real-time Online Spatiotemporal Topic (ROST) model extends the Latent Dirichlet Allocation model to operate on data with an associated spatiotemporal context (Girdhar et al., 2014). ROST alters inference so that the topic distribution at a particular point in spacetime incorporates information from nearby points (Girdhar & Dudek, 2015). This allows learned models to generate realistic spatiotemporal distributions for topics. The ROST model has been used to model distributions of corals and seafloor types from robotic surveys of coral reefs (Jamieson et al., 2021), and topics learned from a ROST model have been previously shown to capture meaningful co-occurrence relationships from phytoplankton observation data (Kalmbach et al., 2017).

Compared to standard dimensionality reduction based community modeling approaches such as PCA and NMDS, topic models are more directly interpretable. PCA components are eigenvectors of the covariance matrix, and loadings for a given variable and component represent the correlation between them. But component weights for each observation may be arbitrary positive or negative real numbers. In fact, the location of data in the lower-dimensional space will only be a rotation and flattening of the high-dimensional data. NMDS embeddings are even less directly interpretable than PCA components. NMDS embedding dimensions do not directly correspond to any variables, and the values produced are non-quantitative. Further clustering analysis on NMDS embeddings can identify similar data points, but relationships between observed variables are still not directly encoded and must be inferred. In contrast, topic models produce both a distribution of topics over (space-)time, and a distribution of variables within each topic. The distribution of variables within each topic is a valid categorical probability distribution, and the probabilities can be understood as relative abundances of a particular variable within a given community.

In this paper, we use a Bayesian topic modeling approach to characterize surface plankton community variability, and uncover mechanisms by which disturbance influences that variability. We highlight how topic modeling augments a more traditional NMDS-based approach to link specific co-occurrence patterns to observed similarities in data. With a pseudo passive tracer approach, we show that the learned topic model agrees with a storm-driven surface advection hypothesis for explaining plankton community variability inside a coherent North Atlantic eddy.

2 Methods

2.1 Survey site and timeline

The Porcupine Abyssal Plain (PAP) sits near the transition zone between the North Atlantic subpolar and subtropical gyres (Henson et al., 2012; Chaudhuri et al., 2011; Eden & Willebrand, 2001). The presence of a long-term observatory at PAP, as well as continuous plankton recorder surveys across the region, provide a long history of community level plankton data. This site was chosen for study in the EXport Processes in the Ocean from Remote Sensing (EXPORTS) 2021 spring campaign, which was focused on characterizing the processes controlling carbon flux in the vicinity of a mesoscale eddy (Johnson et al., 2023). An extensive eddy tracking campaign preceded a three ship adaptive sampling effort, allowing for coordinated deployments of instruments and resolution of $\mathcal{O}(100\text{ m})$ spatial variability.

2.2 Data collection

From May 5-21 2021, the R/V *Sarmiento de Gamboa* conducted sampling of a targeted retentive eddy (Figure 1a). With an Imaging FlowCytobot (McLane Reseearch Laboratories, Inc.) plankton imaging system sampling from the Sarmiento's underway seawater pump, images of surface plankton were taken approximately every 20 minutes. These images were classified with a CNN to produce a time series of 50 different plankton taxa concentrations. Niskin bottle samples from CTD casts were also run through an IFCB imaging system, but as the temporal resolution of these data is coarse (about 1 to 2 profiles per day), we do not analyze them in this paper.

The EXPORTS field program targeted sampling within and around a single mesoscale eddy east of the PAP observatory (Johnson et al., 2023). Sophisticated real-time eddy tracking (Erickson et al., 2023) allowed data to be collected in an "eddy center" reference frame, with multiple vessels and assets aimed at characterizing both the eddy center and the variability across the eddy.

Temporal sampling was designed around three epochs of 7 to 10 day duration. These epochs were punctuated by four major storms that passed through the study site. This work will focus on data collected while the R/V *Sarmiento de Gamboa* was on site, which include epoch 1 and 2 and storms on May 7-11 and May 14-16. These two storms limited the ability of the three ships to sample near the target eddy at those times. Major analyses of temporal trends in community composition around the eddy are therefore structured around the impacts of these storms (Figure 1c).

The North Atlantic is characterized by warm salty waters from the south and cold fresh waters from the north. The energetic eddy field stirs these waters, creating sharp variations in temperature and salinity around the eddy edges. Three surface water masses were identified near the survey site, distinguished primarily by spice (e.g. a measure of the temperature and salinity along density surfaces (McDougall et al., 2021)) and density. A surface core water mass is defined as water within 15km of the eddy center (hereafter referred to as core waters). For surface waters outside of the eddy, cold-fresh and warm-salty water masses are distinguished by a spice threshold of 2.1. A particularly relevant source of horizontal variability is a warm/salty (high spice) filament to the south east that is wrapped around the eddy periphery by the geostrophic flow; hereafter referred to as the "filament" (see Figure 1b). Further details about water mass classification are given in Johnson et al. (2023). Johnson et al. showed that storm driven Ekman currents caused exchange between core water and surrounding water classes. In this work we focus on how that exchange impacted phytoplankton communities in the core waters of the eddy.

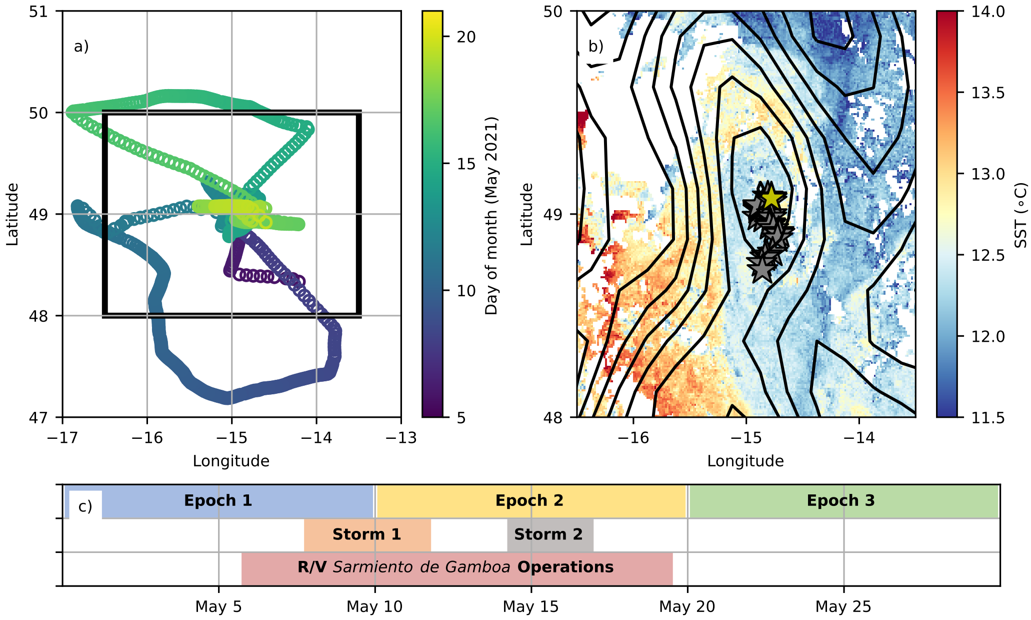

Figure 1. As part of the May 2021 EXPORTS North Atlantic field program, the R/V *Sarmiento de Gamboa* performed extensive oceanographic sampling in and around a retentive eddy in the northeast Atlantic ocean. (a) R/V *Sarmiento de Gamboa* cruise track, with date indicated in color. Two clear deviations from the sampling plan reflect the ship's avoiding of a pair of storms during the cruise. The rectangle indicates the region pictured in Figure 1b. (b) Satellite-derived sea surface temperature and sea surface height in the vicinity of the quasi-retentive eddy, 13 May 2021. Gray stars represent all daily post-processed eddy centers during May 2021, while the yellow star represents the eddy center on May 13. (c) The two main sampling periods planned in advance of the cruise (Epoch 1, May 1-10, and Epoch 2, May 11-20) were interrupted by storm activity.

2.3 Plankton images and classification

Regions of Interest (ROIs) extracted from IFCB images were classified with a CNN-based classifier. The CNN sorted each ROI into one of 50 different classes distinguished morphologically (Orenstein et al., 2015). Of these taxa, two (*bead* and *bubble*) are grouped into the "artifacts" category, and five (*detritus*, *detritus_transparent*, *detritus_theca_fragment*, *fecal_pellet*, and *fiber*) are grouped into "Other not alive". ROIs classified into these categories were removed from the data prior to analysis. An additional taxon, *nanoplankton_mix*, contained ROIs of miscellaneous nanoplankton. Topic models learned with miscellaneous nanoplankton excluded tended to infer more distinguishable topics. In the interest of improving community composition analysis with topic models, these data were also excluded from further analysis.

2.4 NMDS embeddings

NMDS analyses were run with three different dissimilarity matrices. Bray-Curtis dissimilarity was used with direct plankton count data, and KL divergence was used on plankton relative abundance data and ROST model topic proportions. The `scikit-learn python` package's NMDS ordination algorithm was used to calculate lower-dimensional embeddings. Three initialization strategies were compared: random initialization, geographic initialization, PCA initialization, and higher-dimensional NMDS initialization. Initialization of the embeddings with principle coordinates from a PCA analysis resulted in the lowest stress of all strategies, and was used for all further analyses in this paper. NMDS ordinations were used to generate 2D embeddings, to facilitate visualization and further analysis.

2.5 Topic modeling

A ROST model was trained to produce four topics from plankton relative abundance data. Training was done using the `rost-cli` command line program for 1000 epochs, with Dirichlet hyperparameters $\alpha = 0.001$ and $\beta = 0.001$. Most values of Dirichlet hyperparameters below 1.0 produced qualitatively similar results, so no rigorous hyperparameter search was performed. The most important hyperparameter for model quality was the number of topics, K. We chose four topics for the analysis in the rest of the paper, as it effectively captures much of the increase in model accuracy over the bulk of the cruise without including too many negligible communities. Specifically, four topics is the largest number for which each topic has a distinct dominant taxon (plurality relative abundance). With 5 or more topics, the ROST model consistently identified at least 2 topics with a shared dominant taxon. By choosing four topics, the ROST model is forced to identify the primary co-occurrence pattern associated with each of the most common taxa, instead of spreading co-occurrence patterns among multiple topics which causes identifiability issues.

The ROST model (Figure 2) assumes data are produced by a generative process linking each categorical observation of a single plankton to latent (unobserved) assemblages or communities of taxa. Every location in space-time is associated with a particular distribution over communities. To generate an observation, first a community is randomly chosen for that observation. Then, that community's relative abundances are used as probabilities to choose the observed taxon. As both community relative abundances and the spatiotemporal distributions of communities are jointly learned by the model, we infer an effective relative abundance of all plankton taxa at every location containing observations. By comparing this inferred relative abundance to the actual relative abundances, we can quantify the accuracy of a set of learned communities. The Kullbach-Liebler (KL) divergence measures the difference between two probability distributions. Formally, the KL divergence is the expected log likelihood ratio between two distributions P and Q, if an observation is actually drawn from P:

$$D_{KL}(P||Q) = \mathbf{E}_{x\sim P}\left[\log\left(\frac{P(x)}{Q(x)}\right)\right]$$

For absolute and relative abundance data, Bray-Curtis dissimilarity provides a quantifiable measure of the difference between observations. In terms of the relative abun-

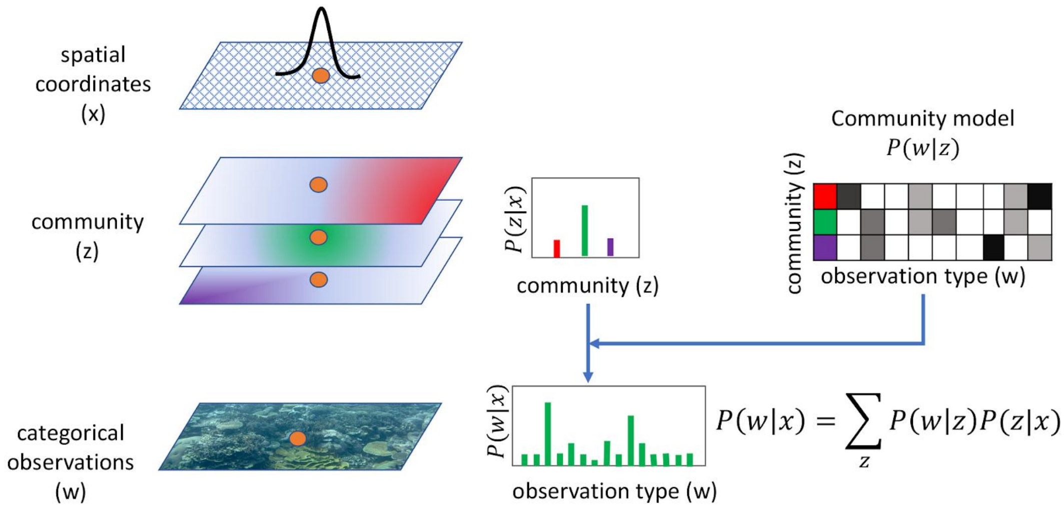

Figure 2. A spatiotemporal topic model factors the distribution of categorical observations in spacetime into a pair of latent distributions. One latent distribution, the "community model", represents a series of communities or topics, each of which is itself a distribution over observation types. The other latent distribution, the spatiotemporal model, represents the probability of finding each topic at any point in spacetime. By multiplying the spatiotemporal model by the matrix community model, we can recover a spatiotemporal distribution over observation types with significantly fewer parameters and desirable structural properties such as sparseness and robustness to rare observation types.

dance, we can calculate it as follows:

$$D_{BC}(P||Q) = 1 - \sum_x \min\left(P(x), Q(x)\right)$$

Bray-Curtis dissimilarity is bounded to be between zero and one, and unlike D_{KL} it is symmetric. We primarily use D_{KL} to compare probability distributions. However, for calculation of dissimilarity matrices as an initial step in other analyses, we also use D_{BC}.

2.6 Water mass mixing

To quantify the extent to which storm events caused surface water mass mixing, we consider the observed plankton concentrations to be ideal (passive) tracers and calculate how close a given sample is to a sample from a mixture of the water masses. First, we take the mean concentration of each taxon in each water mass sampled before the first storm. These are mixed in varying ratios, and normalized to produce a mean mixture relative abundance, representing the hypothetical community of a mixture of the mean of each water mass. At each point between the two storms when the eddy water mass is sampled, we calculate the mean mixture relative abundance with the lowest KL divergence to the observed relative abundance. Zero KL divergence implies that a sampled point's community can be perfectly represented as a mixture of the mean communities seen before the storm. High KL divergence implies that a mixture model is a poor fit for the data, and the observed community variability likely has another mechanism (such as vertical mixing or biological dynamics).

3 Results

3.1 NMDS embedding of observed plankton taxa

NMDS embeddings from a Bray-Curtis dissimilarity matrix calculated with plankton taxon relative abundance data (Figure 3) highlight how the eddy becomes more similar to the filament water mass after the first storm. Separating the observations by epoch and water mass (Figure 3a) identifies a tight cluster of observations for the core water mass in epoch 1. From epoch 1 to epoch 2, the core water mass cluster centroid becomes more negative along the x component, and more positive along the y component. Additionally, the core surface waters get saltier over the same timespan (Figure 3b). This also represents a mean shift of the eddy towards the filament. These results support a mixing/advective source of plankton variability in the core. This is consistent with results from Johnson et al. (2023), which suggests wind driven Ekman transport advected warm salty water from the filament into the 15 km radius around the eddy center.

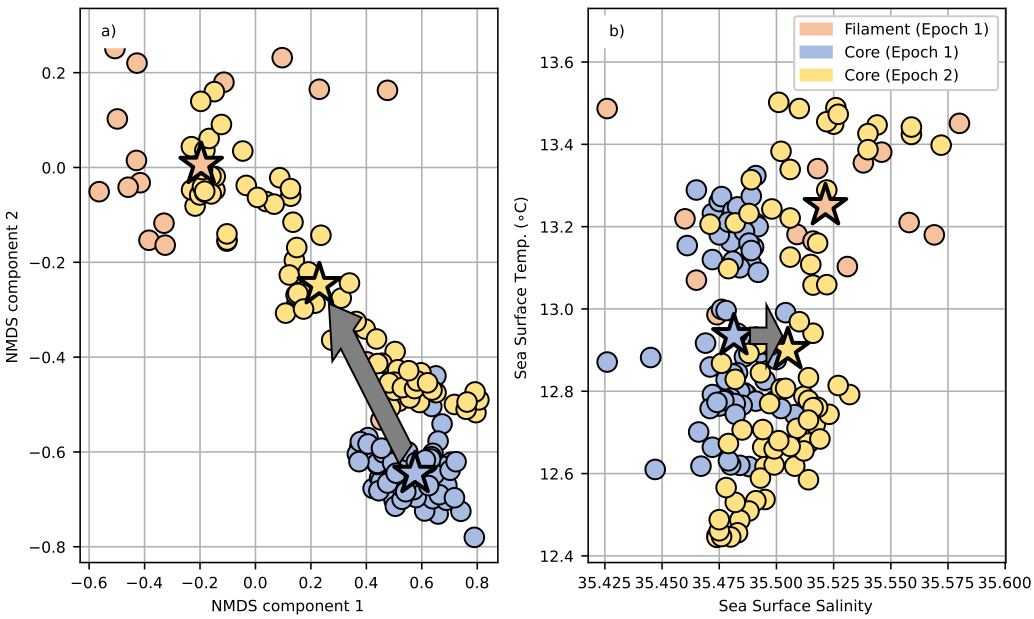

Figure 3. (a) Two-dimensional NMDS embedding of plankton relative abundance data. (b) Temperature-salinity diagram. Stars mark group means for the water mass/epoch combinations listed in the legend. The grey arrow indicates the change in the eddy water mass mean from epoch 1 to epoch 2.

3.2 Community variability inferred by ROST model

Topic models represent co-occurrence patterns as a topic, i.e. a probability distribution over observation categories. These topics are directly interpretable as representing a hypothetical relative abundance matching the co-occurrence pattern, with real observations being drawn from a mixture of these hypothetical abundances. Looking at these patterns temporally (Figure 4) highlights the high variability of community 1. Community 1's relative abundance varies from more than 60% at the peak during epoch 1, to completely absent a few days later during the first storm. This corresponds to a *Pseudo-nitzschia*-dominated community (Figure 5) highly present inside the eddy, especially during epoch 1, but relatively low-proportion far away from it. We proceed by breaking out the community relative abundances in both space and time (Figure 6 and Table 1), in order to characterize broad patterns in community distributions during the cruise.

Community 1 can be identified with initial *Pseudo-nitzschia* bloom conditions inside the eddy. From the initial high proportion in the eddy, community 1 proportions decrease with distance before the first storm and sharply decreases with time after the first storm inside the eddy.

Community 2, dominated by *Dinophyceae_morphotype3*, starts at a low 4% relative abundance inside the eddy, and increases throughout the cruise, ending at a mean eddy relative abundance of 13%. The concentration of this community peaks at 81% far away from the eddy during the first storm excursion.

Community 3 contains a plurality of *Dinophyceae*. Its relative abundance inside the eddy increases throughout the cruise, from about 10% at the start to about 45% at the end. However, these are strictly lower than the abundances seen far from the eddy. The highest abundances of this community are seen far from the eddy at all times.

Community 4 has the highest relative abundance of the *pennate* taxon. Inside the eddy, this community has three distinct relative abundances before, between, and after the two storms, with the mean peaking between the two storms. The distribution is similar just outside the eddy, with lower proportions than inside the eddy. The highest relative abundances of this community are seen during the first storm excursion, as well as inside the eddy between the storms. These peaks are just under 20%, however.

These communities learned by the topic model are highly informative about the water masses sampled, but do not match the water masses exactly. This suggests that water mass variability is linked to (but not the only driver of) plankton community variability in the region surveyed.

Overall, the communities seen in the eddy shift markedly over time, transitioning from a *Pseudo-nitzschia* dominated diatom bloom to a more mixed community. Community 1 (the only community with a significant proportion of *Pseudo-nitzschia*) makes up 84% of the mean community proportions seen in the core in epoch 1, but in epoch 2 it decreases to 42%. All other communities increase in the core from epoch 1 to epoch 2, with communities 2 and 4 reaching a maximum between the storms and community 3 increasing throughout the cruise.

3.3 Topic models decompose compositonal impacts of water mass mixing

The ROST communities inferred in the eddy after the first storm resemble a mixture of the communities in the eddy and filament before the first storm (Figure 3a), further supporting the notion that mixing and/or advection during the storm are primary drivers of plankton variability in the eddy. In epoch 1, the eddy is dominated by the *Pseudo-nitzschia* bloom of community 1 (Figure 7b), while the filament is dominated by community 3, which is primarily *Dinophyceae_morphotype3* (Figure 7a). Later in the cruise, the community distribution in the eddy shifts to be less dominated by community 1 (the bloom community). Instead, the community distribution represents more of a mixture of the community distribution in the eddy and the filament from epoch 1 (Figure 7c). Johnson et al. (2023) showed that Ekman currents during storm 1 flushed approximately 73% of the surface core waters that were replaced with warm/salty waters outside the eddy.

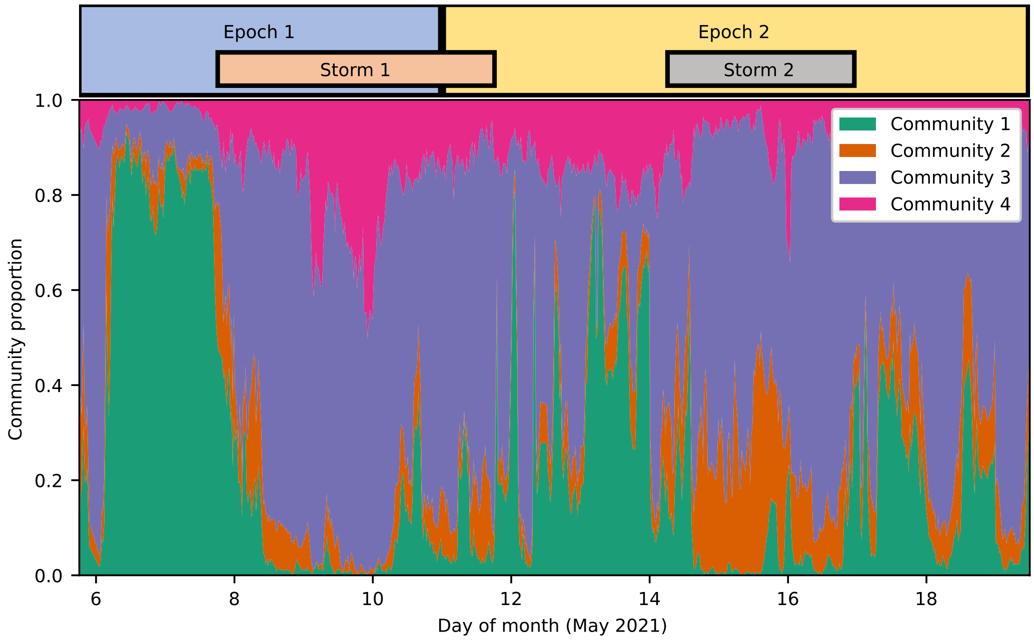

Figure 4. The relative abundance of topics ("communities") inferred by the ROST model versus time over the cruise.

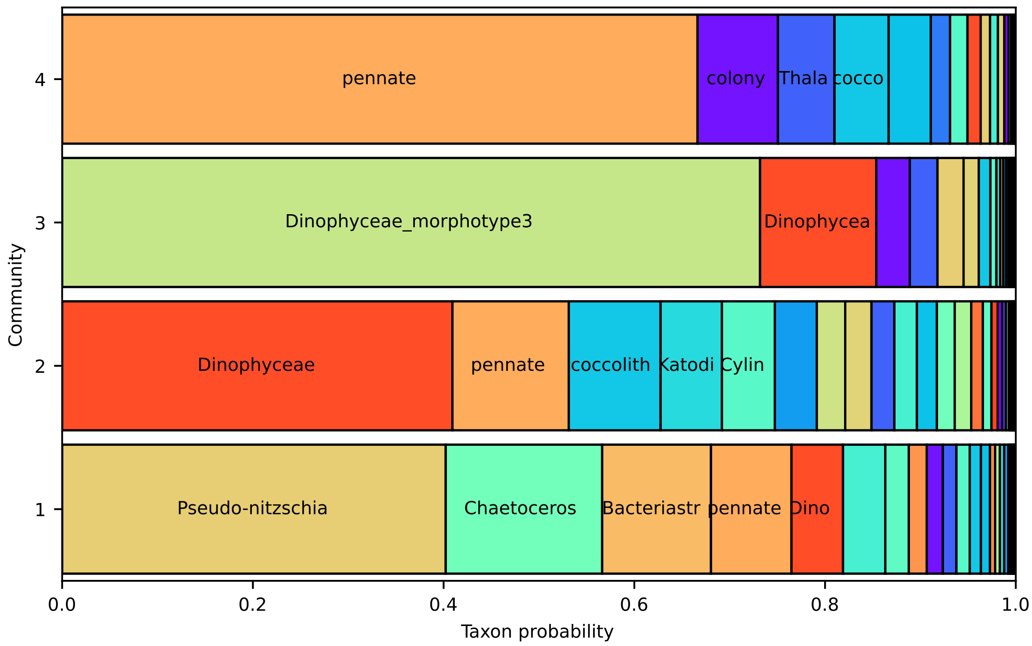

Figure 5. Inferred ROST community model proportions for the different taxa in each community.

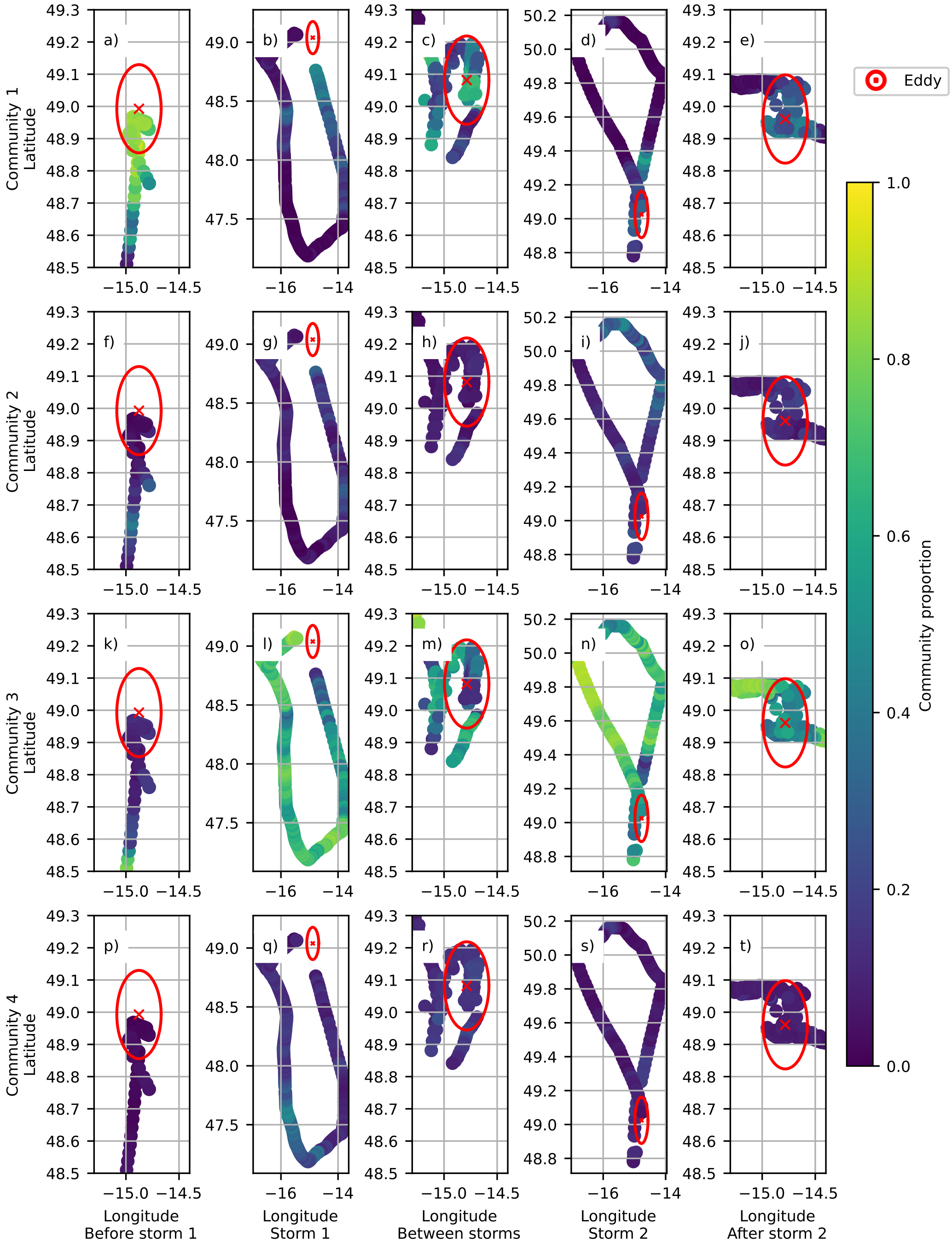

Figure 6. Sptial distribution of proportions for (a-e) Community 1, (f-j): Community 2, (k-o): Community 3 proportions,and (p-t) Community 4 proportions. All panels aggregate data from one of five time periods indicated in Figure 1c and presented left-to-right: Before the first storm, during the first storm, between the two storms, during the second storm, and after the second storm. The mean eddy center and extent (15 km boundary) are marked with a red cross and a circle, respectively. Due to wide deviations in the cruise track during the storms, the second and fourth columns each have their own latitude and longitude bounds. The first, third, and fifth columns share the same latitude and longitude bounds.

Table 1. Mean community proportions by time and location

Cruise period	Location	Com. 1	Com. 2	Com. 3	Com. 4
Before storm 1[a]	Inside[f]	0.840351	0.039149	0.102889	0.017611
	Near[g]	0.651984	0.177954	0.135293	0.034770
	Far[h]	0.141668	0.114440	0.679642	0.064251
Storm 1[b]	Near[g]	0.451712	0.225081	0.216624	0.106584
	Far[h]	0.082952	0.081422	0.658297	0.177329
Between storms[c]	Inside[f]	0.422649	0.081400	0.336334	0.159617
	Near[g]	0.250780	0.085952	0.532940	0.130328
	Far[h]	0.030553	0.217530	0.681103	0.070815
Storm 2[d]	Far[h]	0.034290	0.225775	0.662466	0.077469
After storm 2[e]	Inside[f]	0.320021	0.133385	0.468161	0.078433
	Near[g]	0.116105	0.103529	0.701524	0.078842
	Far[h]	0.021195	0.155310	0.766571	0.056924

[a]May 5-7 [b]May 8-12 [c]May 13-14 [d]May 15-17 [e]May 18-20 [f]<15 km [g]15-45 km [h]>45 km

To better highlight the role mixing plays in altering plankton community structure, we considered an end-member mixing scenario in which the three water masses (`core/eddy`, `warm_salty/filament`, and `cold_fresh`) are mixed in proportions adding to one. Mean plankton concentrations observed before the first storm are treated as ideal (passive) tracers, and the mixed concentrations are normalized to produce an ideally mixed community. For each set of observations taken inside the eddy between the two storms (i.e., after the first storm but before the second storm), the mixture community with the smallest Kullback-Leibler divergence to the observed community at that time was determined (Figure 7d).

The mixing analysis suggests that surface advection drives the warm-salty water mass into the waters above the eddy core. Plankton taxon distributions in the northwest of the eddy seen after the first storm (Figure 7e) closely resemble the mean warm-salty water mass community seen before the storm. East and north-east of the eddy center, post-storm observations resemble none of the pre-storm mean communities. Observations near the eddy center, as well as north and south of it, closely resemble mixtures of pre-storm communities in all water masses.

Analysis of shifts in eddy and filament plankton community composition suggest that water mass mixing may be a significant driver of plankton community variability specifically inside the eddy. Before the first storm, the eddy is dominated by a *Pseudo-nitzschia* bloom, which the topic model represents as a single community dominating over 80% of the eddy plankton community composition. After the first storm, the eddy has a significantly lower proportion of that community, especially near the northwestern edge. There the bloom community is partially succeeded by community 4. Water mass mixing results show that those points with the highest fraction of the warm/salty water mass have the highest proportion of community 3, with the linear fit (Figure 7f) having an r^2 of 0.986.

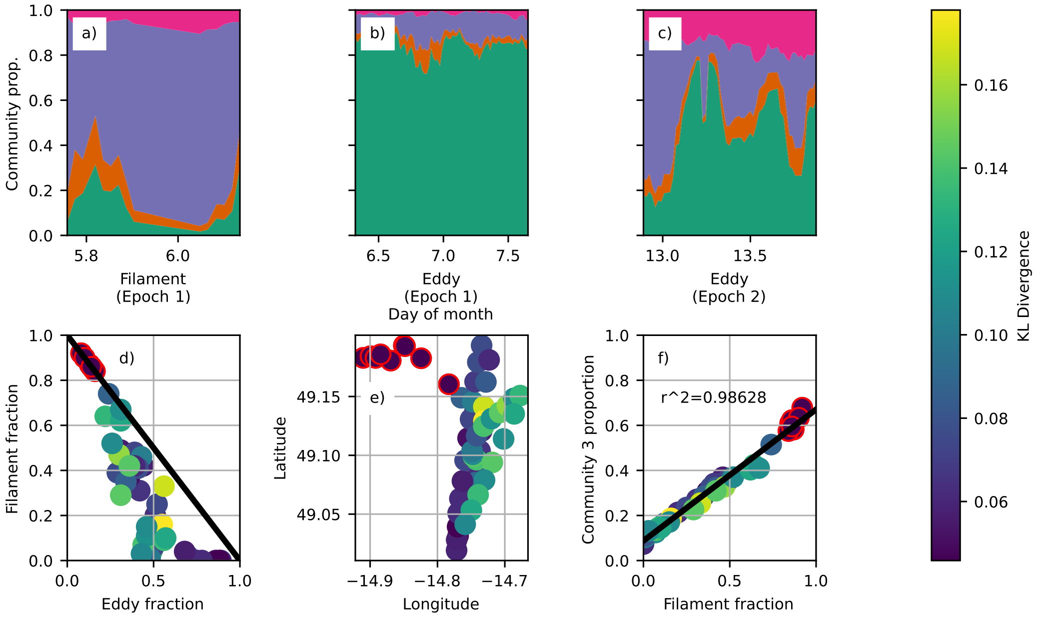

Figure 7. (a-c) The proportions of each community in the filament during epoch 1, the eddy during epoch 1, and the eddy during epoch 2 respectively. Colors indicate the same communities as in Figure 4. (d) A water mass mixing analysis, where the closest water mass mixture to each observation taken in the eddy during epoch 2 is plotted on a 3-component simplex. The three coordinates on the simplex are x (the eddy water mass fraction), y (the filament water mass fraction), and $1 - x - y$ (the cold-fresh water mass fraction). (e) Observations made in the eddy during epoch 2. (f) The relative abundance of community 3 versus the fraction of the warm/salty water mass assigned to each point in the eddy in epoch 2. A linear fit has an r^2 of 0.986. In (d-f), color indicates the KL divergence between the observed plankton distribution in epoch 2 and the lowest KL-divergence distribution of all possible water mass mixtures in epoch 1. Red circles indicate points with a filament fraction above 0.8.

4 Discussion

4.1 Topic models provide a quantitative and interpretable decomposition

The NMDS analysis (Figure 3a) suggests that after the first storm, the eddy surface plankton community became more like the epoch 1 filament community. However, the abstract nature of the NMDS embedding precludes an immediate deeper analysis of the nature of that change. We could, for example, find correlations between the NMDS components and plankton concentrations for various taxa. But NMDS embedding magnitudes and distances do not have any intrinsic meaning. Instead of quantitative analysis, an ordination technique such as NMDS would generally be followed by a qualitative study of correlation with other variables or clustering within the embeddings (Clapham, 2011).

In contrast, topic models directly support quantitative claims about changes in plankton relative abundance. The topic model's communities represent point estimates of relative abundances for each plankton taxon considered in the model. We can therefore inspect spatiotemporal distributions of each community (Figure 6), analyze trends in mean community proportions (Table 1), and model linear relationships between these communities and other hypothetical relative abundance distributions (Figure 7f). The inherent interpretability of topic models also allows for more immediate diagnosing of the nature of major trends seen in data. Consider the temporal distribution of community 1 (Figure 4), along with its associated taxon probabilities (Figure 5). We can immediately spot that community 1 represents a high *Pseudo-nitzschia* abundance, and by looking at its spatial distribution (Figure 6a-e) we conclude that a major source of plankton variability during the cruise was a *Pseudo-nitzschia* bloom in the eddy that dissipated somewhat after the first storm. These kinds of inferences are not possible solely with ordination techniques like NMDS; at a minimum, further processing and analysis of the NMDS output is required.

4.2 Rapid bloom dissipation points to extreme event

Friedland et al. (2018) found that dominant seasonal phytoplankton blooms last on the order of weeks to months across the globe. However, the rather dramatic shift in eddy plankton community composition (from a community dominated by *Pseudo-nitzschia* to a richer community with higher concentrations of other diatoms) occurred over several days of stormy weather. The speed with which the eddy shifted away from a bloom state suggests that the driver of the change may have been an extreme event not well represented by the predominant bloom dissipation mechanisms previously described.

4.3 Upwelling hypothesis and trends in surface chlorophyll

Painter et al. (2016) use a particular North Atlantic storm to highlight how storms structure post-storm plankton communities by enhancing upwelling. This enhanced upwelling brings nutrients to the euphotic zone, setting up conditions for a bloom. Liu and Tang (2018) suggest that this mechanism is responsible for observed post-typhoon chlorophyll fluorescence increases in anti-cyclonic eddies in the South China Sea. In contrast, we found a *decrease* in chlorophyll fluorescence, with high statistical significance (although low r^2) over the course of the cruise (Figure 8a). If the surface was already in the middle of a bloom, we might not expect an increase in productivity. But the observed decrease in chlorophyll fluorescence goes against bloom dynamics being controlled primarily by storm-driven upwelling. Additionally, the mixed layer in the eddy deepened during the storm (Figure 8c). While this points to enhanced vertical mixing, the upper water column has fairly high relative abundance of *Pseudo-*

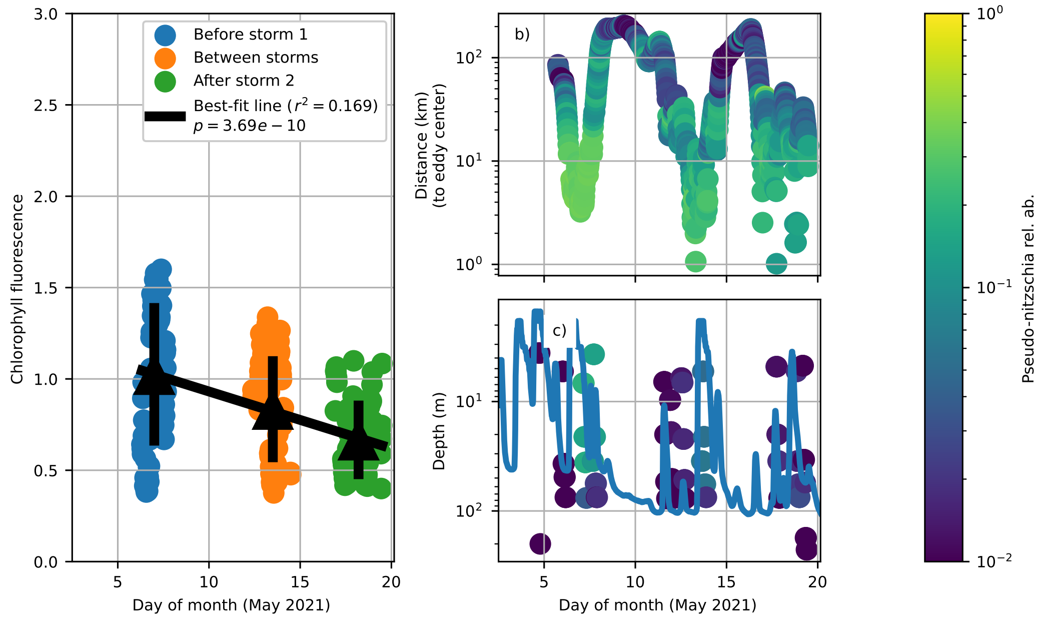

Figure 8. (a) Eddy surface chlorophyll fluorescence versus time during the cruise. Black triangles indicate mean of a color, and the black lines indicate one standard deviation. The line of best fit for all data is indicated in black. (b) Distance to eddy center (km, log scale) versus day of month, with *Pseudo-nitzschia* relative abundance in color. (c) CTD cast Niskin bottle depth (m) with 1D model eddy mixed layer depth (m), with *Pseudo-nitzschia* relative abundance in color.

nitzschia in the eddy before the first storm. Simple dilution through the mixed layer would not account for the observed decrease in *Pseudo-nitzschia* relative abundance.

4.4 Storm-driven advection and stirring control plankton variability

We previously argued that the speed with which the eddy transitioned away from the *Pseudo-nitzschia* bloom community is uncharacteristic of traditional plankton bloom dynamical timescales (section 4.2). We also found evidence against a vertical mixing mechanism for the observed changes in eddy plankton community composition. Instead, our results suggest that horizontal stirring and advection were a major mechanism driving changes in the eddy community. Several observations taken inside the eddy during epoch 2 have plankton communities closely linked to the filament watermass (Figures 7d and 7e). These observations, which have among the lowest kl divergence to the closest water mass mixture of all the observations made during epoch 2, likely represent storm-driven advection of filament water into the northwest corner of the eddy. Some data points in the north, center, and south of the eddy are also fairly well represented as mixtures, with most of the lowest KL-divergence observations found at or near the eddy-filament mixture line (Figure 7d). We can infer that advection likely carried filament plankton communities into the eddy, displacing the bloom community there before the storms. This aligns with Johnson et al. (2023), who found that surface advection and stirring during the storms altered eddy surface temperature and salinity.

4.5 Limitations and future work

This work serves as a demonstration of the successful use of topic modeling for marine plankton ecology, but we do not make any quantitative contrasts between topic models and more traditional dimensionality reduction approaches. The different nature of the outputs of different methods (probability distributions in topic models versus real numbers in NMDS/PCA/etc.) makes direct comparison and evaluation difficult, even though they operate on similar kinds of data. Some of these alternative dimensionality reduction and ordination techniques may offer more quantitative or interpretable outputs.

Our analysis of topic modeling on its own similarly does not quantitatively explore the impacts of the different ROST hyperparameters on the quality or fit of the resulting embeddings. As with other dimensionality reduction techniques, increasing the number of dimensions (topics) in the model improves the fit at the expense of model interpretability and simplicity. The other two hyperparameters control the shape of the prior distribution, and given enough time their impact is washed out in the inferred posterior. The structure of the data likely play a role in determining the importance of all of these hyperparameters, and particularly the sensitivity to the prior distribution. We found that for the plankton data presented here, the prior hyperparameters did not meaningfully impact the visual quality or KL divergence of the resulting community distributions when varied over several orders of magnitude.

Understanding the full scope of spatiotemporal variability requires better resolution of subsurface plankton communities, as well as decoupling surface spatial and temporal observations. IFCBs onboard the other two ships in the field campaign collected surface and CTD cast plankton imagery.

5 Conclusion

In this paper, we demonstrated the power of topic modeling as a tool for uncovering community variability in marine plankton. The 2021 North Atlantic EXPORTS field campaign produced a large quantity of high-resolution phytoplankton image data which allow for the resolution of fine-scale spatiotemporal variability in surface phytoplankton communities. By using topic models to infer latent plankton co-occurrence patterns, we discovered that storm-driven advection was a likely source of surface variability in community structure. Notwithstanding the extreme simplification of treating plankton as pseudo passive tracers, we found strong correlations between a particular co-occurring plankton community and advection of warm, salty water into the eddy. These findings highlight the power of topic modeling as a tool for ecological analysis, particularly in the face of large amounts of spatiotemporally-distributed categorical data. As the resolution and processing power of in-situ imaging systems continues to grow, we foresee an important role for topic models in improving our understanding of marine ecological variability.

Acronyms

IFCB Imaging Flow Cytobot, a high-throughput plankton imaging system that uses flow cytometry and microfluidics to take pictures of phytoplankton precisely when they are in focus of a camera lens

ROST Real-time Online Spatiotemporal Topic model, a Bayesian model for the distribution of categorical information in space-time

CNN Convolutional Neural Network, a neural network architecture which pools data spatially and has been widely applied to image classification tasks

PCA Principal Component Analysis, a statistical technique where a data matrix is decomposed into its eigenvectors to capture major sources of variation

NMDS Non-metric Multi-Dimensional Scaling, a statistical technique for dimensionality reduction which attempts to preserve structural relationships from high dimensions in lower-dimensional embeddings

PAP Porcupine Abyssal Plain, a region of the seafloor in the northeast Atlantic southwest of Ireland

EXPORTS EXport Processes in the Ocean from Remote Sensing, a NASA field campaign to study carbon export in the Earth's oceans

ROI Region of Interest, a portion of an image extracted for further classification

KL Divergence Kullback-Liebler Divergence, a statistical measure of the difference between two probability distributions

Open Research Section

Raw data and products from the NASA EXPORTS program can be found at `https://seabass.gsfc.nasa.gov/`. IFCB images and machine learning labels can be found at `https://ifcb-data.whoi.edu/`. The code for `ROST` can be found at `https://gitlab.com/warplab/rostpy`.

Acknowledgments

This work was part of the Woods Hole Oceanographic Institution's Ocean Twilight Zone Project, funded as part of the Audacious Project housed at TED, with additional support provided by the Simons Foundation (grant 561126 to HMS), NASA Ocean Biology and Biogeochemistry program (grant 80NSSC17K0700 to HMS), NSF-NRI (1734400 to YG), and a National Defense Science and Engineering Graduate Fellowship (to JESS).

References

Allen, S., Henson, S., Hickman, A., Beaulieu, C., Doncaster, P., & Johns, D. (2020, November). Interannual stability of phytoplankton community composition in the North-East Atlantic. *Marine Ecology Progress Series*, *655*, 43–57. Retrieved 2022-11-10, from `https://www.int-res.com/abstracts/meps/v655/p43-57/` doi: 10.3354/meps13515

Banse, K. (1994). Grazing and zooplankton production as key controls of phytoplankton production in the open ocean. *Oceanography*, *7*(1), 13–20. Retrieved 2023-01-27, from `http://www.jstor.org/stable/43925524`

Barcelos e Ramos, J., Schulz, K. G., Voss, M., Narciso, A., Muller, M. N., Reis, F. V., . . . Azevedo, E. B. (2017, 05). Nutrient-specific responses of a phytoplankton community: a case study of the North Atlantic Gyre, Azores. *Journal of Plankton Research*, *39*(4), 744-761. Retrieved from `https://doi.org/10.1093/plankt/fbx025` doi: 10.1093/plankt/fbx025

Brown, A. R., Lilley, M., Shutler, J., Lowe, C., Artioli, Y., Torres, R., . . . Tyler, C. R. (2020). Assessing risks and mitigating impacts of harmful algal blooms on mariculture and marine fisheries. *Reviews in Aquaculture*, *12*(3), 1663-1688. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/raq.12403` doi: https://doi.org/10.1111/raq.12403

Campbell, L., Olson, R. J., Sosik, H. M., Abraham, A., Henrichs, D. W., Hyatt, C. J., & Buskey, E. J. (2010). First harmful dinophysis (dinophyceae, dinophysiales) bloom in the u.s. is revealed by automated imaging flow cytometry1. *Journal of Phycology*, *46*(1), 66-75. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1529-8817.2009.00791.x` doi: https://doi.org/10.1111/j.1529-8817.2009.00791.x

Catlett, D., Peacock, E. E., Crockford, E. T., Futrelle, J., Batchelder, S., Stevens,

B. L. F., . . . Sosik, H. M. (2023). Temperature dependence of parasitoid infection and abundance of a diatom revealed by automated imaging and classification. *Proceedings of the National Academy of Sciences*, *120*(28), e2303356120. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2303356120 doi: 10.1073/pnas.2303356120

Chaudhuri, A. H., Gangopadhyay, A., & Bisagni, J. J. (2011, September). Contrasting Response of the Eastern and Western North Atlantic Circulation to an Episodic Climate Event*. *Journal of Physical Oceanography*, *41*(9), 1630–1638. Retrieved 2022-11-10, from http://journals.ametsoc.org/doi/10.1175/2011JPO4512.1 doi: 10.1175/2011JPO4512.1

Clapham, M. E. (2011). Ordination methods and the evaluation of ediacaran communities. In *Quantifying the evolution of early life: Numerical approaches to the evaluation of fossils and ancient ecosystems* (p. 3-21). Springer Netherlands.

David M. Blei, Ng, Andrew Y, & Jordan, Michael I. (2003, March). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from https://dl.acm.org/doi/10.5555/944919.944937

Eden, C., & Willebrand, J. (2001, May). Mechanism of Interannual to Decadal Variability of the North Atlantic Circulation. *Journal of Climate*, *14*(10), 2266–2280. Retrieved 2022-11-10, from http://journals.ametsoc.org/doi/10.1175/1520-0442(2001)014<2266:MOITDV>2.0.CO;2 doi: 10.1175/1520-0442(2001)014⟨2266:MOITDV⟩2.0.CO;2

Erickson, Z. K., Fields, E., Johnson, L., Thompson, A. F., Dove, L. A., D'Asaro, E., & Siegel, D. A. (2023). Eddy tracking from in situ and satellite observations. *Journal of Geophysical Research: Oceans*, e2023JC019701.

Fiorendino, J. M., Gaonkar, C. C., Henrichs, D. W., & Campbell, L. (2021, 11). Drivers of microplankton community assemblage following tropical cyclones. *Journal of Plankton Research*. Retrieved from https://doi.org/10.1093/plankt/fbab073 (fbab073) doi: 10.1093/plankt/fbab073

Frederiksen, M., Edwards, M., Richardson, A. J., Halliday, N. C., & Wanless, S. (2006). From plankton to top predators: bottom-up control of a marine food web across four trophic levels. *Journal of Animal Ecology*, *75*(6), 1259-1268. Retrieved from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2656.2006.01148.x doi: https://doi.org/10.1111/j.1365-2656.2006.01148.x

Friedland, K. D., Mouw, C. B., Asch, R. G., Ferreira, A. S. A., Henson, S., Hyde, K. J. W., . . . Brady, D. C. (2018). Phenology and time series trends of the dominant seasonal phytoplankton bloom across global scales. *Global Ecology and Biogeography*, *27*(5), 551–569. doi: 10.1111/geb.12717

Girdhar, Y., & Dudek, G. (2015, September). *Gibbs Sampling Strategies for Semantic Perception of Streaming Video Data.* arXiv. Retrieved 2022-11-10, from http://arxiv.org/abs/1509.03242 (arXiv:1509.03242 [cs])

Girdhar, Y., Giguère, P., & Dudek, G. (2014, April). Autonomous adaptive exploration using realtime online spatiotemporal topic modeling. *The International Journal of Robotics Research*, *33*(4), 645–657. Retrieved 2022-11-02, from http://journals.sagepub.com/doi/10.1177/0278364913507325 doi: 10.1177/0278364913507325

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., . . . Gorsky, G. (2016, 02). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*. doi: 10.1038/nature16942

Henson, S., Cael, B., Allen, S., & Dutkiewicz, S. (2021, 09). Future phytoplankton diversity in a changing climate. *Nature Communications*, *12*. doi: 10.1038/s41467-021-25699-w

Henson, S., Lampitt, R., & Johns, D. (2012, November). Variability in phytoplankton community structure in response to the North Atlantic Oscillation and implications for organic carbon flux. *Limnology and Oceanography*, *57*(6), 1591–1601. Retrieved 2022-11-02, from `http://doi.wiley.com/10.4319/lo.2012.57.6.1591` doi: 10.4319/lo.2012.57.6.1591

Ianora, A., Bentley, M., Caldwell, G., Casotti, R., Cembella, A., Engström-Öst, J., ... Vaiciute, D. (2011, 01). The relevance of marine chemical ecology to plankton and ecosystem function: an emerging field. *Marine Drugs*, *9*, 1625-1648.

Jamieson, S., Fathian, K., Khosoussi, K., How, J. P., & Girdhar, Y. (2021, May). Multi-Robot Distributed Semantic Mapping in Unfamiliar Environments through Online Matching of Learned Representations. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 8587–8593). Xi'an, China: IEEE. Retrieved 2022-11-10, from `https://ieeexplore.ieee.org/document/9561934/` doi: 10.1109/ICRA48506.2021.9561934

Johnson, L., Siegel, D., Thompson, A., Fields, E., Erickson, Z., Cetinic, I., ... others (2023). *Assessment of oceanographic conditions during the north atlantic export processes in the ocean from remote sensing (exports) field campaign.*

Kalmbach, A., Girdhar, Y., Sosik, H. M., & Dudek, G. (2017, May). Phytoplankton hotspot prediction with an unsupervised spatial community model. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 4906–4913). Singapore, Singapore: IEEE. Retrieved 2022-11-02, from `http://ieeexplore.ieee.org/document/7989568/` doi: 10.1109/ICRA.2017.7989568

Lévy, M., & Martin, A. P. (2013). The influence of mesoscale and submesoscale heterogeneity on ocean biogeochemical reactions. *Global Biogeochemical Cycles*, *27*(4), 1139–1150.

Liu, F., & Tang, S. (2018). Influence of the Interaction Between Typhoons and Oceanic Mesoscale Eddies on Phytoplankton Blooms. *Journal of Geophysical Research: Oceans*, *123*(4), 2785–2794. doi: 10.1029/2017JC013225

Mahadevan, A. (2016). The impact of submesoscale physics on primary productivity of plankton. *Annual Review of Marine Science*, *8*(1), 161-184. Retrieved from `https://doi.org/10.1146/annurev-marine-010814-015912` (PMID: 26394203) doi: 10.1146/annurev-marine-010814-015912

Mahadevan, A., D'Asaro, E., Lee, C., & Perry, M. J. (2012, July). Eddy-Driven Stratification Initiates North Atlantic Spring Phytoplankton Blooms. *Science*, *337*(6090), 54–58. Retrieved 2022-11-10, from `https://www.science.org/doi/10.1126/science.1218740` doi: 10.1126/science.1218740

Mahadevan, A., Tandon, A., & Ferrari, R. (2010). Rapid changes in mixed layer stratification driven by submesoscale instabilities and winds. *Journal of Geophysical Research: Oceans*, *115*(C3). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008JC005203` doi: https://doi.org/10.1029/2008JC005203

Martin, A. (2003). Phytoplankton patchiness: the role of lateral stirring and mixing. *Progress in oceanography*, *57*(2), 125–174.

McDougall, T. J., Barker, P. M., & Stanley, G. J. (2021). Spice variables and their use in physical oceanography. *Journal of Geophysical Research: Oceans*, *126*(2), e2019JC015936. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015936` (e2019JC015936 2019JC015936) doi: https://doi.org/10.1029/2019JC015936

Olson, R. J., Shalapyonok, A., Kalb, D. J., Graves, S. W., & Sosik, H. M. (2017). Imaging flowcytobot modified for high throughput by in-line acoustic focusing of sample particles. *Limnology and Oceanography: Methods*, *15*(10), 867-874. Retrieved from `https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.1002/lom3.10205` doi: https://doi.org/10.1002/lom3.10205

Olson, R. J., & Sosik, H. M. (2007, June). A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot: In situ imaging of nano- and microplankton. *Limnology and Oceanography: Methods*, *5*(6), 195–203. Retrieved 2022-11-02, from http://doi.wiley.com/10.4319/lom.2007.5.195 doi: 10.4319/lom.2007.5.195

Orenstein, E. C., Beijbom, O., Peacock, E. E., & Sosik, H. M. (2015). *WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification* (Tech. Rep.). Retrieved from https://arxiv.org/abs/1510.00745 doi: 10.48550/ARXIV.1510.00745

Painter, S. C., Finlay, M., Hemsley, V. S., & Martin, A. P. (2016, March). Seasonality, phytoplankton succession and the biogeochemical impacts of an autumn storm in the northeast Atlantic Ocean. *Progress in Oceanography*, *142*, 72–104. doi: 10.1016/j.pocean.2016.02.001

Peacock, E., Sosik, H., & Olson, R. (2014, 04). Parasitic infection of the diatom guinardia delicatula, a recurrent and ecologically important phenomenon on the new england shelf. *Marine Ecology Progress Series*, *503*, 1-10. doi: 10.3354/meps10784

Raitsos, D. E., Pradhan, Y., Lavender, S. J., Hoteit, I., McQuatters-Gollop, A., Reid, P. C., & Richardson, A. J. (2014). From silk to satellite: half a century of ocean colour anomalies in the northeast atlantic. *Global Change Biology*, *20*(7), 2117-2123. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.12457 doi: https://doi.org/10.1111/gcb.12457

Ryther, J. H. (1969). Photosynthesis and fish production in the sea. *Science*, *166*(3901), 72-76. Retrieved from https://www.science.org/doi/abs/10.1126/science.166.3901.72 doi: doi:10.1126/science.166.3901.72

Sverdrup, H. U. (1953, 01). On Conditions for the Vernal Blooming of Phytoplankton. *ICES Journal of Marine Science*, *18*(3), 287-295. Retrieved from https://doi.org/10.1093/icesjms/18.3.287 doi: 10.1093/icesjms/18.3.287

Taylor, J. R., & Ferrari, R. (2011). Shutdown of turbulent convection as a new criterion for the onset of spring phytoplankton blooms. *Limnology and Oceanography*, *56*(6), 2293–2307.

Spatiotemporal Topic Modeling Reveals Storm-Driven Advection and Stirring Control Plankton Community Variability in an Open Ocean Eddy

John E. San Soucie[1,3], Yogesh Girdhar[1], Leah Johnson[4], Emily E. Peacock[2], Alexei Shalapyonok[2], Heidi M. Sosik[2]

[1]Woods Hole Oceanographic Institution, Applied Ocean Physics and Engineering Department
[2]Woods Hole Oceanographic Institution, Biology Department
[3]Massachusetts Institute of Technology, Department of Mechanical Engineering
[4]University of Washington, Applied Physics Laboratory

Key Points:

- Topic models provide a robust alternative to traditional statistical techniques for analysis of sparse, high-dimensional categorical data.
- We perform topic model analyses of machine-classified plankton images taken near a retentive eddy during the 2021 EXPORTS NA field campaign.
- Surface advection and stirring during storms controlled the surface plankton community of the eddy as it transitioned out of a diatom bloom.

Corresponding author: John E. San Soucie, `jsansoucie@whoi.edu`

Abstract

Phytoplankton communities in the open ocean are high-dimensional, sparse, and spatiotemporally heterogeneous. The advent of automated imaging systems has enabled high-resolution observation of these communities, but the amounts of data and their statistical properties make analysis with traditional approaches challenging. Spatiotemporal topic models offer an unsupervised and interpretable approach to dimensionality reduction of sparse, high-dimensional categorical data. Here we use topic modeling to analyze neural-network-classified phytoplankton imagery taken in and around a retentive eddy during the 2021 North Atlantic EXport Processes in the Ocean from Remote Sensing (EXPORTS) field campaign. We investigate the role physical-biological interactions play in altering plankton community composition within the eddy. Analysis of a water mass mixing framework suggests that storm-driven surface advection and stirring were major drivers of the progression of the eddy plankton community away from a diatom bloom over the course of the cruise.

Plain Language Summary

Plankton communities in the ocean can have many different species, with large differences in their abundance and patchy distributions in space. Automated imaging systems allow for high-resolution observation of these plankton communities, but many traditional statistical techniques fail to capture their full complexity. Spatiotemporal topic models, a kind of statistical model designed to work directly with categorical data, can effectively represent this kind of information. In this work, we use topic models to analyze plankton images taken near an eddy in the spring of 2021 and classified into 50 different kinds of plankton with an automated algorithm. We investigate how interactions between ocean physics and biology can change the plankton community inside the eddy. Analysis suggests that storms in the area moved surface water carrying a different plankton community into the eddy.

1 Introduction and background

Marine plankton communities are highly dynamic (Ryther, 1969), with impacts from short- (Mahadevan, 2016) and long-scale (Raitsos et al., 2014) ocean physics, weather (Fiorendino et al., 2021) and climate (Henson et al., 2021), chemical presence (Ianora et al., 2011) and nutrient availability (Barcelos e Ramos et al., 2017), and biological interactions (Banse, 1994). In turn, plankton populations have major impacts on the entire marine food web (Frederiksen et al., 2006), commercial fishing and aquaculture (Brown et al., 2020), and ocean carbon fluxes (Guidi et al., 2016). Understanding how plankton communities respond to external disturbance is therefore crucial for economic and climate forecasting efforts.

In the Northeast Atlantic, which has a strong and energetic eddy field and experiences vigorous wintertime convection, ocean physics plays an important role in mediating phytoplankton community dynamics on a wide range of spatiotemporal scales. Interannually, the North Atlantic Oscillation may impact the community balance between diatoms and dinoflagellates (Henson et al., 2012; Allen et al., 2020). Seasonally, the onset of spring diatom blooms has been linked to a shutdown of wintertime convection (Taylor & Ferrari, 2011) along with solar- (Sverdrup, 1953) and eddy-induced (Mahadevan et al., 2010, 2012) restratification. In addition to temporal changes, the existence of an energetic eddy field creates horizontal heterogeneity and patchiness in phytoplankton productivity and type (Martin, 2003; Lévy & Martin, 2013).

Many approaches for characterizing plankton communities are too low-resolution — either spatiotemporally or in terms of the compositional information acquired —

to fully resolve important internal and external dynamics in marine ecosystems. Genomic data from seawater sampled via bottle casts on a ship is limited by the deployment frequency of the sampling rosette (hours). Conversely, bulk property sensor deployments such as fluorometers on profiling moorings can provide high-frequency measurements but lack fine plankton composition resolution.

In contrast, automated imaging techniques can sample at high temporal resolution, with enough detail to resolve relevant taxonomic distinctions. The Imaging FlowCytobot (IFCB) (Olson & Sosik, 2007) uses flow cytometry integrated with video imaging to detect phytoplankton cells in seawater samples. The IFCB typically samples automatically two to three times per hour, generating thousands of plankton images per sample. Due to the high temporal resolution and information density, full manual review of IFCB datasets is impractical. Instead, classification typically proceeds with machine learning-based classifiers. Ecologically relevant classification of IFCB images with machine learning algorithms such as convolutional neural networks (CNNs) has been well documented (Olson & Sosik, 2007; Olson et al., 2017; Campbell et al., 2010; Catlett et al., 2023; Peacock et al., 2014).

The plankton community composition dynamics which generate image time series are nonlinear, with high-dimensional and spatially heterogeneous (patchy) communities. These properties make data analysis challenging. Statistical tools such as Principal Component Analysis (PCA) and (non-)metric Multidimensional Scaling (NMDS and MDS) greatly reduce the dimensionality of the data while preserving part of the higher-dimensional structures and patterns. But some of these tools make unrealistic assumptions about how data are generated. For example, PCA assumes that observations decompose into real-valued weightings of orthogonal eigenvectors, but actual underlying trends in communities need not be orthogonal. Other tools, like (N)MDS, may not make any generative assumptions at all, and provide a purely descriptive approach to dimensionality reduction.

Topic models offer an approximate but robust and interpretable alternative to classical dimensionality reduction approaches. Topic models are a class of Bayesian graphical model that factor the distribution of categorical observations with latent "topics", which themselves represent distributions over observation categories. A key early topic model, the Latent Dirichlet Allocation model (David M. Blei et al., 2003), was originally used to model text documents. With a Bayesian inference algorithm, the Latent Dirichlet Allocation model converges on topics with semantic meaning, organized by co-occuring clusters of words. The Real-time Online Spatiotemporal Topic (ROST) model extends the Latent Dirichlet Allocation model to operate on data with an associated spatiotemporal context (Girdhar et al., 2014). ROST alters inference so that the topic distribution at a particular point in spacetime incorporates information from nearby points (Girdhar & Dudek, 2015). This allows learned models to generate realistic spatiotemporal distributions for topics. The ROST model has been used to model distributions of corals and seafloor types from robotic surveys of coral reefs (Jamieson et al., 2021), and topics learned from a ROST model have been previously shown to capture meaningful co-occurrence relationships from phytoplankton observation data (Kalmbach et al., 2017).

Compared to standard dimensionality reduction based community modeling approaches such as PCA and NMDS, topic models are more directly interpretable. PCA components are eigenvectors of the covariance matrix, and loadings for a given variable and component represent the correlation between them. But component weights for each observation may be arbitrary positive or negative real numbers. In fact, the location of data in the lower-dimensional space will only be a rotation and flattening of the high-dimensional data. NMDS embeddings are even less directly interpretable than PCA components. NMDS embedding dimensions do not directly correspond to any variables, and the values produced are non-quantitative. Further clustering analysis on NMDS embeddings can identify similar data points, but relationships between observed variables are still not directly encoded and must be inferred. In contrast, topic models produce both a distribution of topics over (space-)time, and a distribution of variables within each topic. The distribution of variables within each topic is a valid categorical probability distribution, and the probabilities can be understood as relative abundances of a particular variable within a given community.

In this paper, we use a Bayesian topic modeling approach to characterize surface plankton community variability, and uncover mechanisms by which disturbance influences that variability. We highlight how topic modeling augments a more traditional NMDS-based approach to link specific co-occurrence patterns to observed similarities in data. With a pseudo passive tracer approach, we show that the learned topic model agrees with a storm-driven surface advection hypothesis for explaining plankton community variability inside a coherent North Atlantic eddy.

2 Methods

2.1 Survey site and timeline

The Porcupine Abyssal Plain (PAP) sits near the transition zone between the North Atlantic subpolar and subtropical gyres (Henson et al., 2012; Chaudhuri et al., 2011; Eden & Willebrand, 2001). The presence of a long-term observatory at PAP, as well as continuous plankton recorder surveys across the region, provide a long history of community level plankton data. This site was chosen for study in the EXport Processes in the Ocean from Remote Sensing (EXPORTS) 2021 spring campaign, which was focused on characterizing the processes controlling carbon flux in the vicinity of a mesoscale eddy (Johnson et al., 2023). An extensive eddy tracking campaign preceded a three ship adaptive sampling effort, allowing for coordinated deployments of instruments and resolution of $\mathcal{O}(100\text{ m})$ spatial variability.

2.2 Data collection

From May 5-21 2021, the R/V *Sarmiento de Gamboa* conducted sampling of a targeted retentive eddy (Figure 1a). With an Imaging FlowCytobot (McLane Reseearch Laboratories, Inc.) plankton imaging system sampling from the Sarmiento's underway seawater pump, images of surface plankton were taken approximately every 20 minutes. These images were classified with a CNN to produce a time series of 50 different plankton taxa concentrations. Niskin bottle samples from CTD casts were also run through an IFCB imaging system, but as the temporal resolution of these data is coarse (about 1 to 2 profiles per day), we do not analyze them in this paper.

The EXPORTS field program targeted sampling within and around a single mesoscale eddy east of the PAP observatory (Johnson et al., 2023). Sophisticated real-time eddy tracking (Erickson et al., 2023) allowed data to be collected in an "eddy center" reference frame, with multiple vessels and assets aimed at characterizing both the eddy center and the variability across the eddy.

Temporal sampling was designed around three epochs of 7 to 10 day duration. These epochs were punctuated by four major storms that passed through the study site. This work will focus on data collected while the R/V *Sarmiento de Gamboa* was on site, which include epoch 1 and 2 and storms on May 7-11 and May 14-16. These two storms limited the ability of the three ships to sample near the target eddy at those times. Major analyses of temporal trends in community composition around the eddy are therefore structured around the impacts of these storms (Figure 1c).

The North Atlantic is characterized by warm salty waters from the south and cold fresh waters from the north. The energetic eddy field stirs these waters, creating sharp variations in temperature and salinity around the eddy edges. Three surface water masses were identified near the survey site, distinguished primarily by spice (e.g. a measure of the temperature and salinity along density surfaces (McDougall et al., 2021)) and density. A surface core water mass is defined as water within 15km of the eddy center (hereafter referred to as core waters). For surface waters outside of the eddy, cold-fresh and warm-salty water masses are distinguished by a spice threshold of 2.1. A particularly relevant source of horizontal variability is a warm/salty (high spice) filament to the south east that is wrapped around the eddy periphery by the geostrophic flow; hereafter referred to as the "filament" (see Figure 1b). Further details about water mass classification are given in Johnson et al. (2023). Johnson et al. showed that storm driven Ekman currents caused exchange between core water and surrounding water classes. In this work we focus on how that exchange impacted phytoplankton communities in the core waters of the eddy.

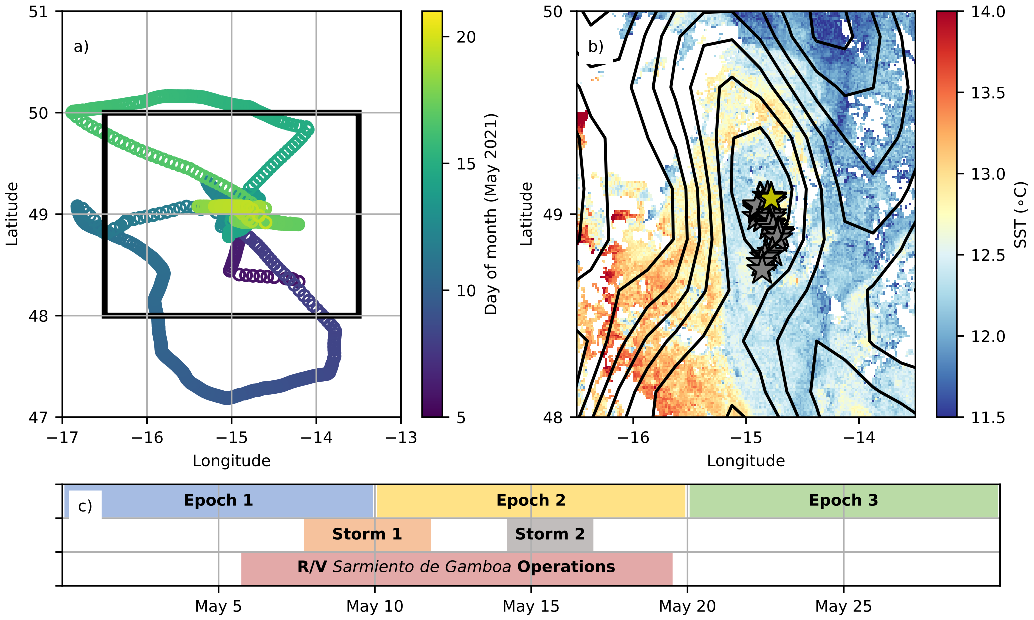

Figure 1. As part of the May 2021 EXPORTS North Atlantic field program, the R/V *Sarmiento de Gamboa* performed extensive oceanographic sampling in and around a retentive eddy in the northeast Atlantic ocean. (a) R/V *Sarmiento de Gamboa* cruise track, with date indicated in color. Two clear deviations from the sampling plan reflect the ship's avoiding of a pair of storms during the cruise. The rectangle indicates the region pictured in Figure 1b. (b) Satellite-derived sea surface temperature and sea surface height in the vicinity of the quasi-retentive eddy, 13 May 2021. Gray stars represent all daily post-processed eddy centers during May 2021, while the yellow star represents the eddy center on May 13. (c) The two main sampling periods planned in advance of the cruise (Epoch 1, May 1-10, and Epoch 2, May 11-20) were interrupted by storm activity.

2.3 Plankton images and classification

Regions of Interest (ROIs) extracted from IFCB images were classified with a CNN-based classifier. The CNN sorted each ROI into one of 50 different classes distinguished morphologically (Orenstein et al., 2015). Of these taxa, two (*bead* and *bubble*) are grouped into the "artifacts" category, and five (*detritus*, *detritus_transparent*, *detritus_theca_fragment*, *fecal_pellet*, and *fiber*) are grouped into "Other not alive". ROIs classified into these categories were removed from the data prior to analysis. An additional taxon, *nanoplankton_mix*, contained ROIs of miscellaneous nanoplankton. Topic models learned with miscellaneous nanoplankton excluded tended to infer more distinguishable topics. In the interest of improving community composition analysis with topic models, these data were also excluded from further analysis.

2.4 NMDS embeddings

NMDS analyses were run with three different dissimilarity matrices. Bray-Curtis dissimilarity was used with direct plankton count data, and KL divergence was used on plankton relative abundance data and ROST model topic proportions. The `scikit-learn python` package's NMDS ordination algorithm was used to calculate lower-dimensional embeddings. Three initialization strategies were compared: random initialization, geographic initialization, PCA initialization, and higher-dimensional NMDS initialization. Initialization of the embeddings with principle coordinates from a PCA analysis resulted in the lowest stress of all strategies, and was used for all further analyses in this paper. NMDS ordinations were used to generate 2D embeddings, to facilitate visualization and further analysis.

2.5 Topic modeling

A ROST model was trained to produce four topics from plankton relative abundance data. Training was done using the `rost-cli` command line program for 1000 epochs, with Dirichlet hyperparameters $\alpha = 0.001$ and $\beta = 0.001$. Most values of Dirichlet hyperparameters below 1.0 produced qualitatively similar results, so no rigorous hyperparameter search was performed. The most important hyperparameter for model quality was the number of topics, K. We chose four topics for the analysis in the rest of the paper, as it effectively captures much of the increase in model accuracy over the bulk of the cruise without including too many negligible communities. Specifically, four topics is the largest number for which each topic has a distinct dominant taxon (plurality relative abundance). With 5 or more topics, the ROST model consistently identified at least 2 topics with a shared dominant taxon. By choosing four topics, the ROST model is forced to identify the primary co-occurrence pattern associated with each of the most common taxa, instead of spreading co-occurrence patterns among multiple topics which causes identifiability issues.

The ROST model (Figure 2) assumes data are produced by a generative process linking each categorical observation of a single plankton to latent (unobserved) assemblages or communities of taxa. Every location in space-time is associated with a particular distribution over communities. To generate an observation, first a community is randomly chosen for that observation. Then, that community's relative abundances are used as probabilities to choose the observed taxon. As both community relative abundances and the spatiotemporal distributions of communities are jointly learned by the model, we infer an effective relative abundance of all plankton taxa at every location containing observations. By comparing this inferred relative abundance to the actual relative abundances, we can quantify the accuracy of a set of learned communities. The Kullbach-Liebler (KL) divergence measures the difference between two probability distributions. Formally, the KL divergence is the expected log likelihood ratio between two distributions P and Q, if an observation is actually drawn from P:

$$D_{KL}(P||Q) = \mathbf{E}_{x\sim P}\left[\log\left(\frac{P(x)}{Q(x)}\right)\right]$$

For absolute and relative abundance data, Bray-Curtis dissimilarity provides a quantifiable measure of the difference between observations. In terms of the relative abun-

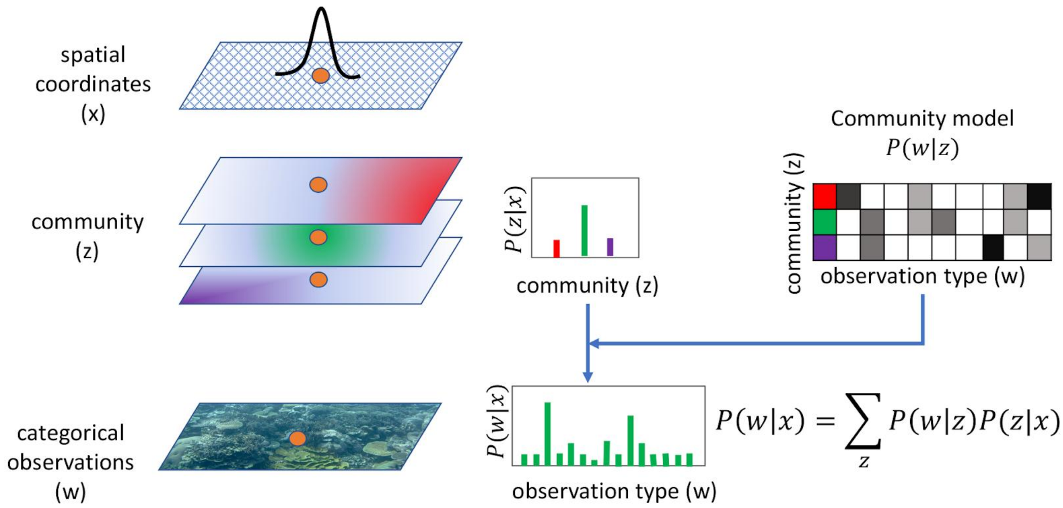

Figure 2. A spatiotemporal topic model factors the distribution of categorical observations in spacetime into a pair of latent distributions. One latent distribution, the "community model", represents a series of communities or topics, each of which is itself a distribution over observation types. The other latent distribution, the spatiotemporal model, represents the probability of finding each topic at any point in spacetime. By multiplying the spatiotemporal model by the matrix community model, we can recover a spatiotemporal distribution over observation types with significantly fewer parameters and desirable structural properties such as sparseness and robustness to rare observation types.

dance, we can calculate it as follows:

$$D_{BC}(P||Q) = 1 - \sum_x \min\left(P(x), Q(x)\right)$$

Bray-Curtis dissimilarity is bounded to be between zero and one, and unlike D_{KL} it is symmetric. We primarily use D_{KL} to compare probability distributions. However, for calculation of dissimilarity matrices as an initial step in other analyses, we also use D_{BC}.

2.6 Water mass mixing

To quantify the extent to which storm events caused surface water mass mixing, we consider the observed plankton concentrations to be ideal (passive) tracers and calculate how close a given sample is to a sample from a mixture of the water masses. First, we take the mean concentration of each taxon in each water mass sampled before the first storm. These are mixed in varying ratios, and normalized to produce a mean mixture relative abundance, representing the hypothetical community of a mixture of the mean of each water mass. At each point between the two storms when the eddy water mass is sampled, we calculate the mean mixture relative abundance with the lowest KL divergence to the observed relative abundance. Zero KL divergence implies that a sampled point's community can be perfectly represented as a mixture of the mean communities seen before the storm. High KL divergence implies that a mixture model is a poor fit for the data, and the observed community variability likely has another mechanism (such as vertical mixing or biological dynamics).

3 Results

3.1 NMDS embedding of observed plankton taxa

NMDS embeddings from a Bray-Curtis dissimilarity matrix calculated with plankton taxon relative abundance data (Figure 3) highlight how the eddy becomes more similar to the filament water mass after the first storm. Separating the observations by epoch and water mass (Figure 3a) identifies a tight cluster of observations for the core water mass in epoch 1. From epoch 1 to epoch 2, the core water mass cluster centroid becomes more negative along the x component, and more positive along the y component. Additionally, the core surface waters get saltier over the same timespan (Figure 3b). This also represents a mean shift of the eddy towards the filament. These results support a mixing/advective source of plankton variability in the core. This is consistent with results from Johnson et al. (2023), which suggests wind driven Ekman transport advected warm salty water from the filament into the 15 km radius around the eddy center.

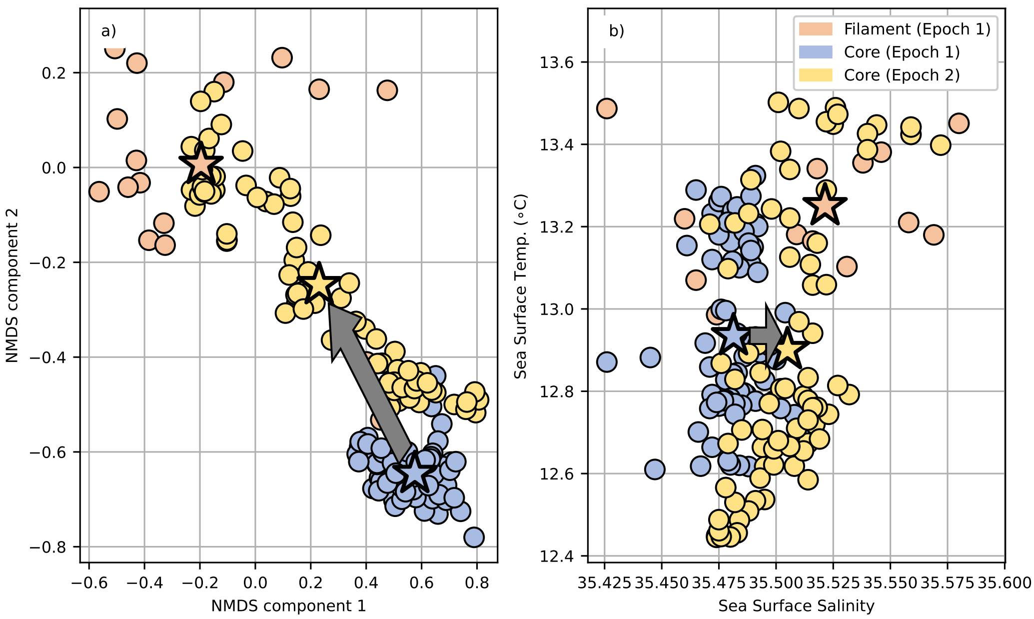

Figure 3. (a) Two-dimensional NMDS embedding of plankton relative abundance data. (b) Temperature-salinity diagram. Stars mark group means for the water mass/epoch combinations listed in the legend. The grey arrow indicates the change in the eddy water mass mean from epoch 1 to epoch 2.

3.2 Community variability inferred by ROST model

Topic models represent co-occurrence patterns as a topic, i.e. a probability distribution over observation categories. These topics are directly interpretable as representing a hypothetical relative abundance matching the co-occurrence pattern, with real observations being drawn from a mixture of these hypothetical abundances. Looking at these patterns temporally (Figure 4) highlights the high variability of community 1. Community 1's relative abundance varies from more than 60% at the peak during epoch 1, to completely absent a few days later during the first storm. This corresponds to a *Pseudo-nitzschia*-dominated community (Figure 5) highly present inside the eddy, especially during epoch 1, but relatively low-proportion far away from it. We proceed by breaking out the community relative abundances in both space and time (Figure 6 and Table 1), in order to characterize broad patterns in community distributions during the cruise.

Community 1 can be identified with initial *Pseudo-nitzschia* bloom conditions inside the eddy. From the initial high proportion in the eddy, community 1 proportions decrease with distance before the first storm and sharply decreases with time after the first storm inside the eddy.

Community 2, dominated by *Dinophyceae_morphotype3*, starts at a low 4% relative abundance inside the eddy, and increases throughout the cruise, ending at a mean eddy relative abundance of 13%. The concentration of this community peaks at 81% far away from the eddy during the first storm excursion.

Community 3 contains a plurality of *Dinophyceae*. Its relative abundance inside the eddy increases throughout the cruise, from about 10% at the start to about 45% at the end. However, these are strictly lower than the abundances seen far from the eddy. The highest abundances of this community are seen far from the eddy at all times.

Community 4 has the highest relative abundance of the *pennate* taxon. Inside the eddy, this community has three distinct relative abundances before, between, and after the two storms, with the mean peaking between the two storms. The distribution is similar just outside the eddy, with lower proportions than inside the eddy. The highest relative abundances of this community are seen during the first storm excursion, as well as inside the eddy between the storms. These peaks are just under 20%, however.

These communities learned by the topic model are highly informative about the water masses sampled, but do not match the water masses exactly. This suggests that water mass variability is linked to (but not the only driver of) plankton community variability in the region surveyed.

Overall, the communities seen in the eddy shift markedly over time, transitioning from a *Pseudo-nitzschia* dominated diatom bloom to a more mixed community. Community 1 (the only community with a significant proportion of *Pseudo-nitzschia*) makes up 84% of the mean community proportions seen in the core in epoch 1, but in epoch 2 it decreases to 42%. All other communities increase in the core from epoch 1 to epoch 2, with communities 2 and 4 reaching a maximum between the storms and community 3 increasing throughout the cruise.

3.3 Topic models decompose compositonal impacts of water mass mixing

The ROST communities inferred in the eddy after the first storm resemble a mixture of the communities in the eddy and filament before the first storm (Figure 3a), further supporting the notion that mixing and/or advection during the storm are primary drivers of plankton variability in the eddy. In epoch 1, the eddy is dominated by the *Pseudo-nitzschia* bloom of community 1 (Figure 7b), while the filament is dominated by community 3, which is primarily *Dinophyceae_morphotype3* (Figure 7a). Later in the cruise, the community distribution in the eddy shifts to be less dominated by community 1 (the bloom community). Instead, the community distribution represents more of a mixture of the community distribution in the eddy and the filament from epoch 1 (Figure 7c). Johnson et al. (2023) showed that Ekman currents during storm 1 flushed approximately 73% of the surface core waters that were replaced with warm/salty waters outside the eddy.

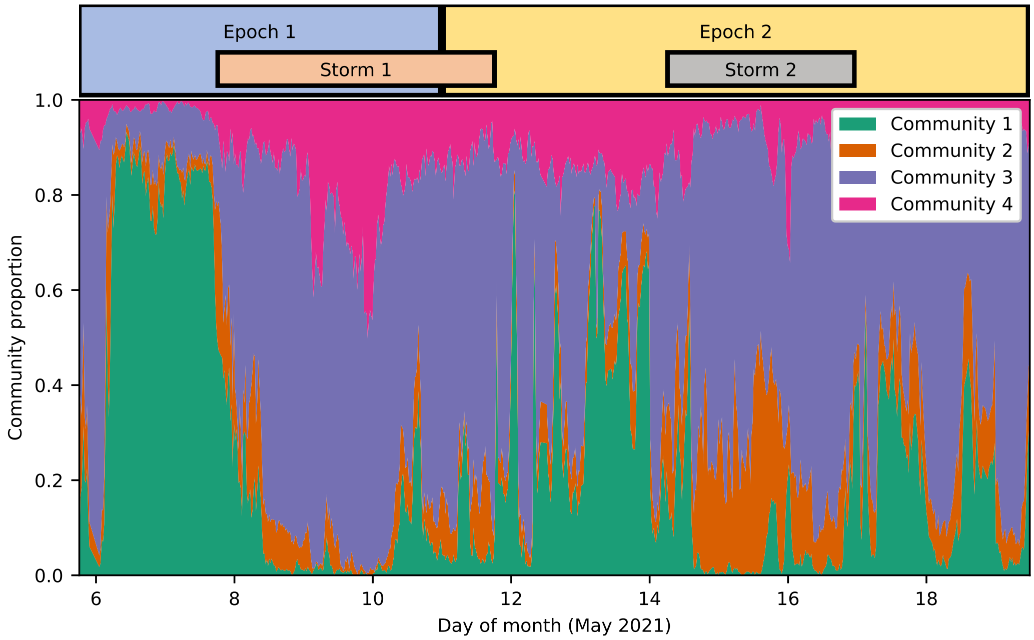

Figure 4. The relative abundance of topics ("communities") inferred by the ROST model versus time over the cruise.

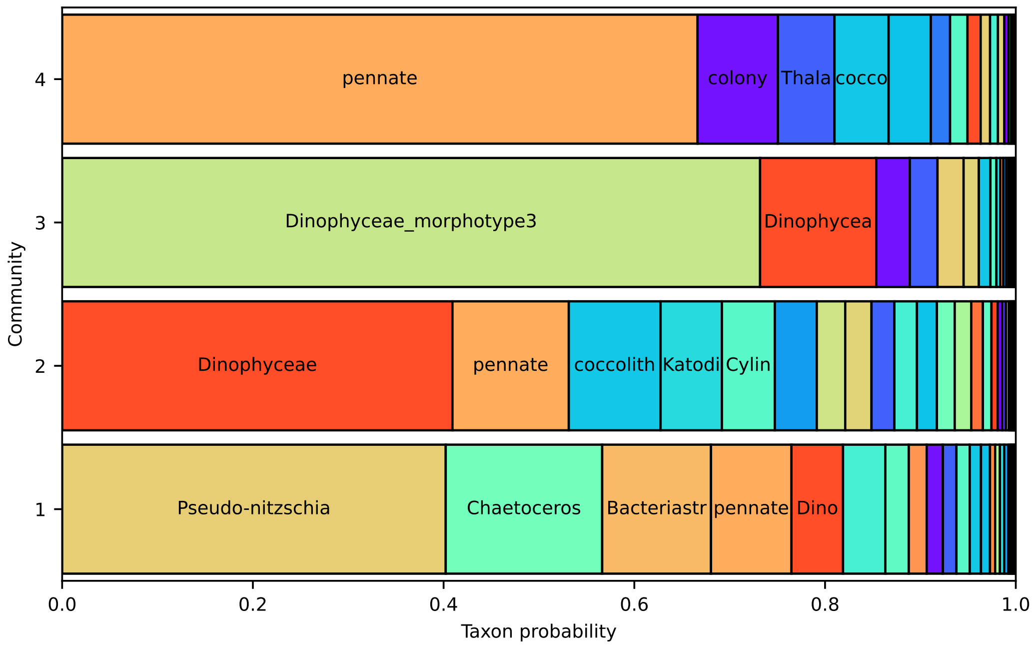

Figure 5. Inferred ROST community model proportions for the different taxa in each community.

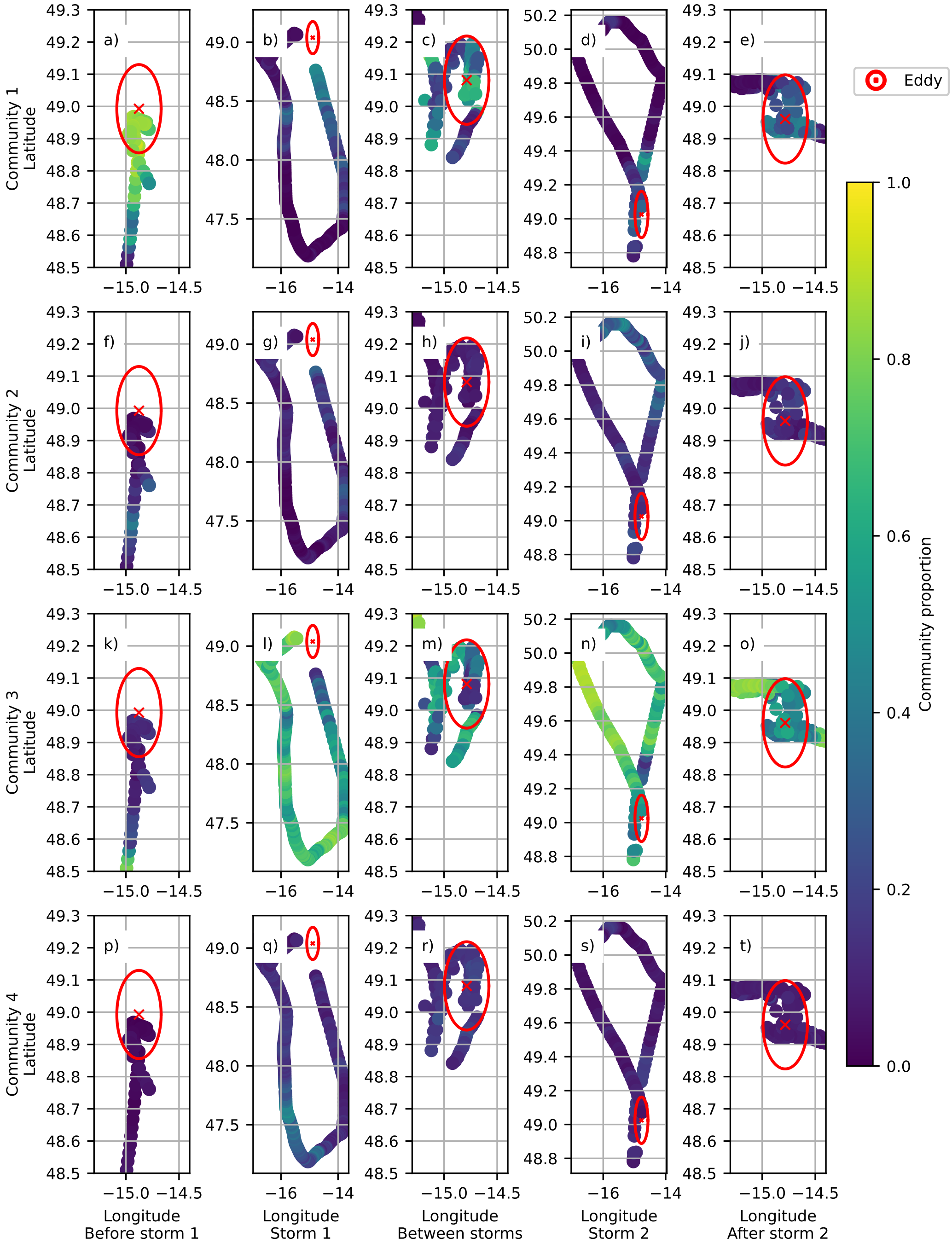

Figure 6. Sptial distribution of proportions for (a-e) Community 1, (f-j): Community 2, (k-o): Community 3 proportions,and (p-t) Community 4 proportions. All panels aggregate data from one of five time periods indicated in Figure 1c and presented left-to-right: Before the first storm, during the first storm, between the two storms, during the second storm, and after the second storm. The mean eddy center and extent (15 km boundary) are marked with a red cross and a circle, respectively. Due to wide deviations in the cruise track during the storms, the second and fourth columns each have their own latitude and longitude bounds. The first, third, and fifth columns share the same latitude and longitude bounds.

Table 1. Mean community proportions by time and location

Cruise period	Location	Com. 1	Com. 2	Com. 3	Com. 4
Before storm 1[a]	Inside[f]	0.840351	0.039149	0.102889	0.017611
	Near[g]	0.651984	0.177954	0.135293	0.034770
	Far[h]	0.141668	0.114440	0.679642	0.064251
Storm 1[b]	Near[g]	0.451712	0.225081	0.216624	0.106584
	Far[h]	0.082952	0.081422	0.658297	0.177329
Between storms[c]	Inside[f]	0.422649	0.081400	0.336334	0.159617
	Near[g]	0.250780	0.085952	0.532940	0.130328
	Far[h]	0.030553	0.217530	0.681103	0.070815
Storm 2[d]	Far[h]	0.034290	0.225775	0.662466	0.077469
After storm 2[e]	Inside[f]	0.320021	0.133385	0.468161	0.078433
	Near[g]	0.116105	0.103529	0.701524	0.078842
	Far[h]	0.021195	0.155310	0.766571	0.056924

[a]May 5-7 [b]May 8-12 [c]May 13-14 [d]May 15-17 [e]May 18-20 [f]<15 km [g]15-45 km [h]>45 km

To better highlight the role mixing plays in altering plankton community structure, we considered an end-member mixing scenario in which the three water masses (`core/eddy`, `warm_salty/filament`, and `cold_fresh`) are mixed in proportions adding to one. Mean plankton concentrations observed before the first storm are treated as ideal (passive) tracers, and the mixed concentrations are normalized to produce an ideally mixed community. For each set of observations taken inside the eddy between the two storms (i.e., after the first storm but before the second storm), the mixture community with the smallest Kullback-Leibler divergence to the observed community at that time was determined (Figure 7d).

The mixing analysis suggests that surface advection drives the warm-salty water mass into the waters above the eddy core. Plankton taxon distributions in the northwest of the eddy seen after the first storm (Figure 7e) closely resemble the mean warm-salty water mass community seen before the storm. East and north-east of the eddy center, post-storm observations resemble none of the pre-storm mean communities. Observations near the eddy center, as well as north and south of it, closely resemble mixtures of pre-storm communities in all water masses.

Analysis of shifts in eddy and filament plankton community composition suggest that water mass mixing may be a significant driver of plankton community variability specifically inside the eddy. Before the first storm, the eddy is dominated by a *Pseudo-nitzschia* bloom, which the topic model represents as a single community dominating over 80% of the eddy plankton community composition. After the first storm, the eddy has a significantly lower proportion of that community, especially near the northwestern edge. There the bloom community is partially succeeded by community 4. Water mass mixing results show that those points with the highest fraction of the warm/salty water mass have the highest proportion of community 3, with the linear fit (Figure 7f) having an r^2 of 0.986.

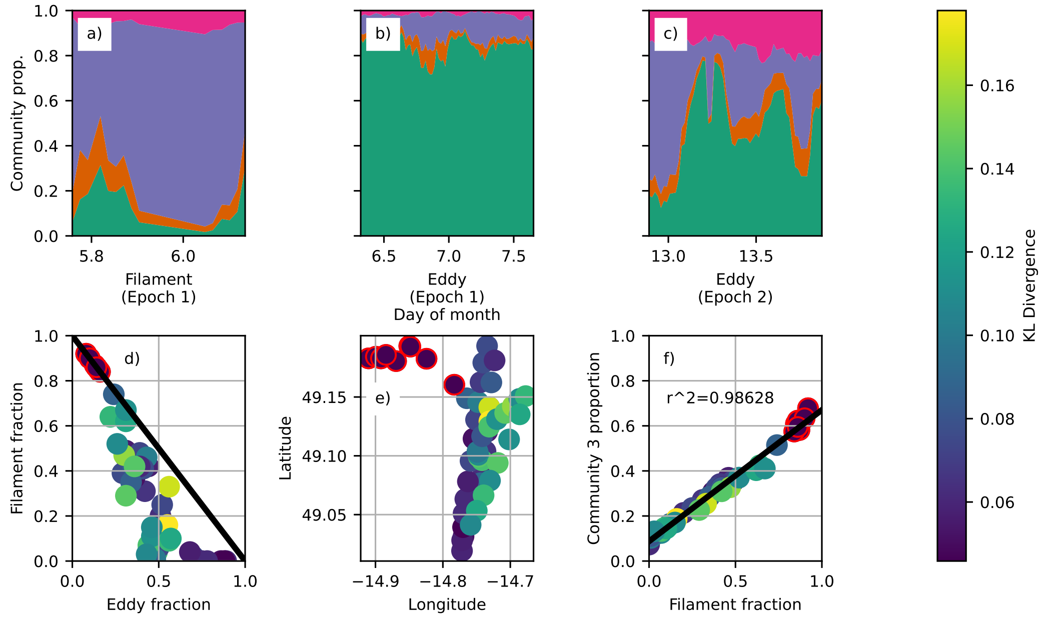

Figure 7. (a-c) The proportions of each community in the filament during epoch 1, the eddy during epoch 1, and the eddy during epoch 2 respectively. Colors indicate the same communities as in Figure 4. (d) A water mass mixing analysis, where the closest water mass mixture to each observation taken in the eddy during epoch 2 is plotted on a 3-component simplex. The three coordinates on the simplex are x (the eddy water mass fraction), y (the filament water mass fraction), and $1 - x - y$ (the cold-fresh water mass fraction). (e) Observations made in the eddy during epoch 2. (f) The relative abundance of community 3 versus the fraction of the warm/salty water mass assigned to each point in the eddy in epoch 2. A linear fit has an r^2 of 0.986. In (d-f), color indicates the KL divergence between the observed plankton distribution in epoch 2 and the lowest KL-divergence distribution of all possible water mass mixtures in epoch 1. Red circles indicate points with a filament fraction above 0.8.

4 Discussion

4.1 Topic models provide a quantitative and interpretable decomposition

The NMDS analysis (Figure 3a) suggests that after the first storm, the eddy surface plankton community became more like the epoch 1 filament community. However, the abstract nature of the NMDS embedding precludes an immediate deeper analysis of the nature of that change. We could, for example, find correlations between the NMDS components and plankton concentrations for various taxa. But NMDS embedding magnitudes and distances do not have any intrinsic meaning. Instead of quantitative analysis, an ordination technique such as NMDS would generally be followed by a qualitative study of correlation with other variables or clustering within the embeddings (Clapham, 2011).

In contrast, topic models directly support quantitative claims about changes in plankton relative abundance. The topic model's communities represent point estimates of relative abundances for each plankton taxon considered in the model. We can therefore inspect spatiotemporal distributions of each community (Figure 6), analyze trends in mean community proportions (Table 1), and model linear relationships between these communities and other hypothetical relative abundance distributions (Figure 7f). The inherent interpretability of topic models also allows for more immediate diagnosing of the nature of major trends seen in data. Consider the temporal distribution of community 1 (Figure 4), along with its associated taxon probabilities (Figure 5). We can immediately spot that community 1 represents a high *Pseudo-nitzschia* abundance, and by looking at its spatial distribution (Figure 6a-e) we conclude that a major source of plankton variability during the cruise was a *Pseudo-nitzschia* bloom in the eddy that dissipated somewhat after the first storm. These kinds of inferences are not possible solely with ordination techniques like NMDS; at a minimum, further processing and analysis of the NMDS output is required.

4.2 Rapid bloom dissipation points to extreme event

Friedland et al. (2018) found that dominant seasonal phytoplankton blooms last on the order of weeks to months across the globe. However, the rather dramatic shift in eddy plankton community composition (from a community dominated by *Pseudo-nitzschia* to a richer community with higher concentrations of other diatoms) occurred over several days of stormy weather. The speed with which the eddy shifted away from a bloom state suggests that the driver of the change may have been an extreme event not well represented by the predominant bloom dissipation mechanisms previously described.

4.3 Upwelling hypothesis and trends in surface chlorophyll

Painter et al. (2016) use a particular North Atlantic storm to highlight how storms structure post-storm plankton communities by enhancing upwelling. This enhanced upwelling brings nutrients to the euphotic zone, setting up conditions for a bloom. Liu and Tang (2018) suggest that this mechanism is responsible for observed post-typhoon chlorophyll fluorescence increases in anti-cyclonic eddies in the South China Sea. In contrast, we found a *decrease* in chlorophyll fluorescence, with high statistical significance (although low r^2) over the course of the cruise (Figure 8a). If the surface was already in the middle of a bloom, we might not expect an increase in productivity. But the observed decrease in chlorophyll fluorescence goes against bloom dynamics being controlled primarily by storm-driven upwelling. Additionally, the mixed layer in the eddy deepened during the storm (Figure 8c). While this points to enhanced vertical mixing, the upper water column has fairly high relative abundance of *Pseudo-*

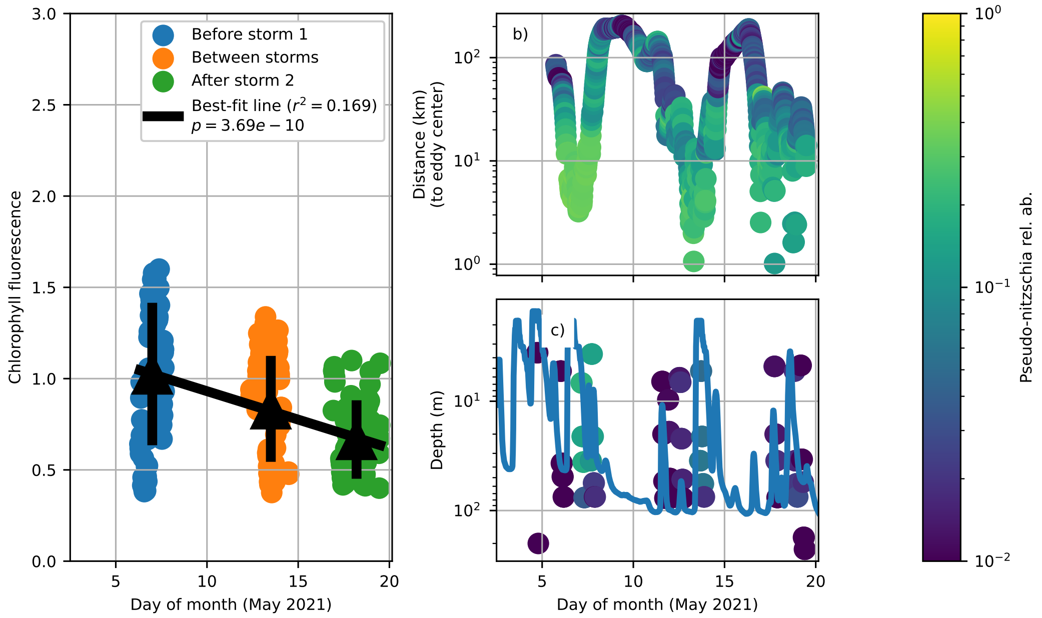

Figure 8. (a) Eddy surface chlorophyll fluorescence versus time during the cruise. Black triangles indicate mean of a color, and the black lines indicate one standard deviation. The line of best fit for all data is indicated in black. (b) Distance to eddy center (km, log scale) versus day of month, with *Pseudo-nitzschia* relative abundance in color. (c) CTD cast Niskin bottle depth (m) with 1D model eddy mixed layer depth (m), with *Pseudo-nitzschia* relative abundance in color.

nitzschia in the eddy before the first storm. Simple dilution through the mixed layer would not account for the observed decrease in *Pseudo-nitzschia* relative abundance.

4.4 Storm-driven advection and stirring control plankton variability

We previously argued that the speed with which the eddy transitioned away from the *Pseudo-nitzschia* bloom community is uncharacteristic of traditional plankton bloom dynamical timescales (section 4.2). We also found evidence against a vertical mixing mechanism for the observed changes in eddy plankton community composition. Instead, our results suggest that horizontal stirring and advection were a major mechanism driving changes in the eddy community. Several observations taken inside the eddy during epoch 2 have plankton communities closely linked to the filament watermass (Figures 7d and 7e). These observations, which have among the lowest kl divergence to the closest water mass mixture of all the observations made during epoch 2, likely represent storm-driven advection of filament water into the northwest corner of the eddy. Some data points in the north, center, and south of the eddy are also fairly well represented as mixtures, with most of the lowest KL-divergence observations found at or near the eddy-filament mixture line (Figure 7d). We can infer that advection likely carried filament plankton communities into the eddy, displacing the bloom community there before the storms. This aligns with Johnson et al. (2023), who found that surface advection and stirring during the storms altered eddy surface temperature and salinity.

4.5 Limitations and future work

This work serves as a demonstration of the successful use of topic modeling for marine plankton ecology, but we do not make any quantitative contrasts between topic models and more traditional dimensionality reduction approaches. The different nature of the outputs of different methods (probability distributions in topic models versus real numbers in NMDS/PCA/etc.) makes direct comparison and evaluation difficult, even though they operate on similar kinds of data. Some of these alternative dimensionality reduction and ordination techniques may offer more quantitative or interpretable outputs.

Our analysis of topic modeling on its own similarly does not quantitatively explore the impacts of the different ROST hyperparameters on the quality or fit of the resulting embeddings. As with other dimensionality reduction techniques, increasing the number of dimensions (topics) in the model improves the fit at the expense of model interpretability and simplicity. The other two hyperparameters control the shape of the prior distribution, and given enough time their impact is washed out in the inferred posterior. The structure of the data likely play a role in determining the importance of all of these hyperparameters, and particularly the sensitivity to the prior distribution. We found that for the plankton data presented here, the prior hyperparameters did not meaningfully impact the visual quality or KL divergence of the resulting community distributions when varied over several orders of magnitude.

Understanding the full scope of spatiotemporal variability requires better resolution of subsurface plankton communities, as well as decoupling surface spatial and temporal observations. IFCBs onboard the other two ships in the field campaign collected surface and CTD cast plankton imagery.

5 Conclusion

In this paper, we demonstrated the power of topic modeling as a tool for uncovering community variability in marine plankton. The 2021 North Atlantic EXPORTS field campaign produced a large quantity of high-resolution phytoplankton image data which allow for the resolution of fine-scale spatiotemporal variability in surface phytoplankton communities. By using topic models to infer latent plankton co-occurrence patterns, we discovered that storm-driven advection was a likely source of surface variability in community structure. Notwithstanding the extreme simplification of treating plankton as pseudo passive tracers, we found strong correlations between a particular co-occurring plankton community and advection of warm, salty water into the eddy. These findings highlight the power of topic modeling as a tool for ecological analysis, particularly in the face of large amounts of spatiotemporally-distributed categorical data. As the resolution and processing power of in-situ imaging systems continues to grow, we foresee an important role for topic models in improving our understanding of marine ecological variability.

Acronyms

IFCB Imaging Flow Cytobot, a high-throughput plankton imaging system that uses flow cytometry and microfluidics to take pictures of phytoplankton precisely when they are in focus of a camera lens

ROST Real-time Online Spatiotemporal Topic model, a Bayesian model for the distribution of categorical information in space-time

CNN Convolutional Neural Network, a neural network architecture which pools data spatially and has been widely applied to image classification tasks

PCA Principal Component Analysis, a statistical technique where a data matrix is decomposed into its eigenvectors to capture major sources of variation

NMDS Non-metric Multi-Dimensional Scaling, a statistical technique for dimensionality reduction which attempts to preserve structural relationships from high dimensions in lower-dimensional embeddings

PAP Porcupine Abyssal Plain, a region of the seafloor in the northeast Atlantic southwest of Ireland

EXPORTS EXport Processes in the Ocean from Remote Sensing, a NASA field campaign to study carbon export in the Earth's oceans

ROI Region of Interest, a portion of an image extracted for further classification

KL Divergence Kullback-Liebler Divergence, a statistical measure of the difference between two probability distributions

Open Research Section

Raw data and products from the NASA EXPORTS program can be found at `https://seabass.gsfc.nasa.gov/`. IFCB images and machine learning labels can be found at `https://ifcb-data.whoi.edu/`. The code for `ROST` can be found at `https://gitlab.com/warplab/rostpy`.

Acknowledgments

This work was part of the Woods Hole Oceanographic Institution's Ocean Twilight Zone Project, funded as part of the Audacious Project housed at TED, with additional support provided by the Simons Foundation (grant 561126 to HMS), NASA Ocean Biology and Biogeochemistry program (grant 80NSSC17K0700 to HMS), NSF-NRI (1734400 to YG), and a National Defense Science and Engineering Graduate Fellowship (to JESS).

References

Allen, S., Henson, S., Hickman, A., Beaulieu, C., Doncaster, P., & Johns, D. (2020, November). Interannual stability of phytoplankton community composition in the North-East Atlantic. *Marine Ecology Progress Series*, *655*, 43–57. Retrieved 2022-11-10, from `https://www.int-res.com/abstracts/meps/v655/p43-57/` doi: 10.3354/meps13515

Banse, K. (1994). Grazing and zooplankton production as key controls of phytoplankton production in the open ocean. *Oceanography*, *7*(1), 13–20. Retrieved 2023-01-27, from `http://www.jstor.org/stable/43925524`

Barcelos e Ramos, J., Schulz, K. G., Voss, M., Narciso, A., Muller, M. N., Reis, F. V., … Azevedo, E. B. (2017, 05). Nutrient-specific responses of a phytoplankton community: a case study of the North Atlantic Gyre, Azores. *Journal of Plankton Research*, *39*(4), 744-761. Retrieved from `https://doi.org/10.1093/plankt/fbx025` doi: 10.1093/plankt/fbx025

Brown, A. R., Lilley, M., Shutler, J., Lowe, C., Artioli, Y., Torres, R., … Tyler, C. R. (2020). Assessing risks and mitigating impacts of harmful algal blooms on mariculture and marine fisheries. *Reviews in Aquaculture*, *12*(3), 1663-1688. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/raq.12403` doi: https://doi.org/10.1111/raq.12403

Campbell, L., Olson, R. J., Sosik, H. M., Abraham, A., Henrichs, D. W., Hyatt, C. J., & Buskey, E. J. (2010). First harmful dinophysis (dinophyceae, dinophysiales) bloom in the u.s. is revealed by automated imaging flow cytometry1. *Journal of Phycology*, *46*(1), 66-75. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1529-8817.2009.00791.x` doi: https://doi.org/10.1111/j.1529-8817.2009.00791.x

Catlett, D., Peacock, E. E., Crockford, E. T., Futrelle, J., Batchelder, S., Stevens,

B. L. F., ... Sosik, H. M. (2023). Temperature dependence of parasitoid infection and abundance of a diatom revealed by automated imaging and classification. *Proceedings of the National Academy of Sciences*, *120*(28), e2303356120. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.2303356120 doi: 10.1073/pnas.2303356120

Chaudhuri, A. H., Gangopadhyay, A., & Bisagni, J. J. (2011, September). Contrasting Response of the Eastern and Western North Atlantic Circulation to an Episodic Climate Event*. *Journal of Physical Oceanography*, *41*(9), 1630–1638. Retrieved 2022-11-10, from http://journals.ametsoc.org/doi/10.1175/2011JPO4512.1 doi: 10.1175/2011JPO4512.1

Clapham, M. E. (2011). Ordination methods and the evaluation of ediacaran communities. In *Quantifying the evolution of early life: Numerical approaches to the evaluation of fossils and ancient ecosystems* (p. 3-21). Springer Netherlands.

David M. Blei, Ng, Andrew Y, & Jordan, Michael I. (2003, March). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from https://dl.acm.org/doi/10.5555/944919.944937

Eden, C., & Willebrand, J. (2001, May). Mechanism of Interannual to Decadal Variability of the North Atlantic Circulation. *Journal of Climate*, *14*(10), 2266–2280. Retrieved 2022-11-10, from http://journals.ametsoc.org/doi/10.1175/1520-0442(2001)014<2266:MOITDV>2.0.CO;2 doi: 10.1175/1520-0442(2001)014⟨2266:MOITDV⟩2.0.CO;2

Erickson, Z. K., Fields, E., Johnson, L., Thompson, A. F., Dove, L. A., D'Asaro, E., & Siegel, D. A. (2023). Eddy tracking from in situ and satellite observations. *Journal of Geophysical Research: Oceans*, e2023JC019701.

Fiorendino, J. M., Gaonkar, C. C., Henrichs, D. W., & Campbell, L. (2021, 11). Drivers of microplankton community assemblage following tropical cyclones. *Journal of Plankton Research*. Retrieved from https://doi.org/10.1093/plankt/fbab073 (fbab073) doi: 10.1093/plankt/fbab073

Frederiksen, M., Edwards, M., Richardson, A. J., Halliday, N. C., & Wanless, S. (2006). From plankton to top predators: bottom-up control of a marine food web across four trophic levels. *Journal of Animal Ecology*, *75*(6), 1259-1268. Retrieved from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2656.2006.01148.x doi: https://doi.org/10.1111/j.1365-2656.2006.01148.x

Friedland, K. D., Mouw, C. B., Asch, R. G., Ferreira, A. S. A., Henson, S., Hyde, K. J. W., ... Brady, D. C. (2018). Phenology and time series trends of the dominant seasonal phytoplankton bloom across global scales. *Global Ecology and Biogeography*, *27*(5), 551–569. doi: 10.1111/geb.12717

Girdhar, Y., & Dudek, G. (2015, September). *Gibbs Sampling Strategies for Semantic Perception of Streaming Video Data.* arXiv. Retrieved 2022-11-10, from http://arxiv.org/abs/1509.03242 (arXiv:1509.03242 [cs])

Girdhar, Y., Giguère, P., & Dudek, G. (2014, April). Autonomous adaptive exploration using realtime online spatiotemporal topic modeling. *The International Journal of Robotics Research*, *33*(4), 645–657. Retrieved 2022-11-02, from http://journals.sagepub.com/doi/10.1177/0278364913507325 doi: 10.1177/0278364913507325

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., ... Gorsky, G. (2016, 02). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*. doi: 10.1038/nature16942

Henson, S., Cael, B., Allen, S., & Dutkiewicz, S. (2021, 09). Future phytoplankton diversity in a changing climate. *Nature Communications*, *12*. doi: 10.1038/s41467-021-25699-w

Henson, S., Lampitt, R., & Johns, D. (2012, November). Variability in phytoplankton community structure in response to the North Atlantic Oscillation and implications for organic carbon flux. *Limnology and Oceanography*, *57*(6), 1591–1601. Retrieved 2022-11-02, from http://doi.wiley.com/10.4319/lo.2012.57.6.1591 doi: 10.4319/lo.2012.57.6.1591

Ianora, A., Bentley, M., Caldwell, G., Casotti, R., Cembella, A., Engström-Öst, J., ... Vaiciute, D. (2011, 01). The relevance of marine chemical ecology to plankton and ecosystem function: an emerging field. *Marine Drugs*, *9*, 1625-1648.

Jamieson, S., Fathian, K., Khosoussi, K., How, J. P., & Girdhar, Y. (2021, May). Multi-Robot Distributed Semantic Mapping in Unfamiliar Environments through Online Matching of Learned Representations. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 8587–8593). Xi'an, China: IEEE. Retrieved 2022-11-10, from https://ieeexplore.ieee.org/document/9561934/ doi: 10.1109/ICRA48506.2021.9561934

Johnson, L., Siegel, D., Thompson, A., Fields, E., Erickson, Z., Cetinic, I., ... others (2023). *Assessment of oceanographic conditions during the north atlantic export processes in the ocean from remote sensing (exports) field campaign.*

Kalmbach, A., Girdhar, Y., Sosik, H. M., & Dudek, G. (2017, May). Phytoplankton hotspot prediction with an unsupervised spatial community model. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 4906–4913). Singapore, Singapore: IEEE. Retrieved 2022-11-02, from http://ieeexplore.ieee.org/document/7989568/ doi: 10.1109/ICRA.2017.7989568

Lévy, M., & Martin, A. P. (2013). The influence of mesoscale and submesoscale heterogeneity on ocean biogeochemical reactions. *Global Biogeochemical Cycles*, *27*(4), 1139–1150.

Liu, F., & Tang, S. (2018). Influence of the Interaction Between Typhoons and Oceanic Mesoscale Eddies on Phytoplankton Blooms. *Journal of Geophysical Research: Oceans*, *123*(4), 2785–2794. doi: 10.1029/2017JC013225

Mahadevan, A. (2016). The impact of submesoscale physics on primary productivity of plankton. *Annual Review of Marine Science*, *8*(1), 161-184. Retrieved from https://doi.org/10.1146/annurev-marine-010814-015912 (PMID: 26394203) doi: 10.1146/annurev-marine-010814-015912

Mahadevan, A., D'Asaro, E., Lee, C., & Perry, M. J. (2012, July). Eddy-Driven Stratification Initiates North Atlantic Spring Phytoplankton Blooms. *Science*, *337*(6090), 54–58. Retrieved 2022-11-10, from https://www.science.org/doi/10.1126/science.1218740 doi: 10.1126/science.1218740

Mahadevan, A., Tandon, A., & Ferrari, R. (2010). Rapid changes in mixed layer stratification driven by submesoscale instabilities and winds. *Journal of Geophysical Research: Oceans*, *115*(C3). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008JC005203 doi: https://doi.org/10.1029/2008JC005203

Martin, A. (2003). Phytoplankton patchiness: the role of lateral stirring and mixing. *Progress in oceanography*, *57*(2), 125–174.

McDougall, T. J., Barker, P. M., & Stanley, G. J. (2021). Spice variables and their use in physical oceanography. *Journal of Geophysical Research: Oceans*, *126*(2), e2019JC015936. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015936 (e2019JC015936 2019JC015936) doi: https://doi.org/10.1029/2019JC015936

Olson, R. J., Shalapyonok, A., Kalb, D. J., Graves, S. W., & Sosik, H. M. (2017). Imaging flowcytobot modified for high throughput by in-line acoustic focusing of sample particles. *Limnology and Oceanography: Methods*, *15*(10), 867-874. Retrieved from https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.1002/lom3.10205 doi: https://doi.org/10.1002/lom3.10205

Olson, R. J., & Sosik, H. M. (2007, June). A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot: In situ imaging of nano- and microplankton. *Limnology and Oceanography: Methods*, *5*(6), 195–203. Retrieved 2022-11-02, from http://doi.wiley.com/10.4319/lom.2007.5.195 doi: 10.4319/lom.2007.5.195

Orenstein, E. C., Beijbom, O., Peacock, E. E., & Sosik, H. M. (2015). *WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification* (Tech. Rep.). Retrieved from https://arxiv.org/abs/1510.00745 doi: 10.48550/ARXIV.1510.00745

Painter, S. C., Finlay, M., Hemsley, V. S., & Martin, A. P. (2016, March). Seasonality, phytoplankton succession and the biogeochemical impacts of an autumn storm in the northeast Atlantic Ocean. *Progress in Oceanography*, *142*, 72–104. doi: 10.1016/j.pocean.2016.02.001

Peacock, E., Sosik, H., & Olson, R. (2014, 04). Parasitic infection of the diatom guinardia delicatula, a recurrent and ecologically important phenomenon on the new england shelf. *Marine Ecology Progress Series*, *503*, 1-10. doi: 10.3354/meps10784

Raitsos, D. E., Pradhan, Y., Lavender, S. J., Hoteit, I., McQuatters-Gollop, A., Reid, P. C., & Richardson, A. J. (2014). From silk to satellite: half a century of ocean colour anomalies in the northeast atlantic. *Global Change Biology*, *20*(7), 2117-2123. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.12457 doi: https://doi.org/10.1111/gcb.12457

Ryther, J. H. (1969). Photosynthesis and fish production in the sea. *Science*, *166*(3901), 72-76. Retrieved from https://www.science.org/doi/abs/10.1126/science.166.3901.72 doi: doi:10.1126/science.166.3901.72

Sverdrup, H. U. (1953, 01). On Conditions for the Vernal Blooming of Phytoplankton. *ICES Journal of Marine Science*, *18*(3), 287-295. Retrieved from https://doi.org/10.1093/icesjms/18.3.287 doi: 10.1093/icesjms/18.3.287

Taylor, J. R., & Ferrari, R. (2011). Shutdown of turbulent convection as a new criterion for the onset of spring phytoplankton blooms. *Limnology and Oceanography*, *56*(6), 2293–2307.