2021

# A Kernel-Based Change Detection Method to Map Shifts in Phytoplankton Communities Measured by Flow Cytometry

Corinne Jones

Sophie Clayton
*Old Dominion University*, sclayton@odu.edu

François Ribalet

E. Virginia Armbrust

Zaid Harchaoui



## Original Publication Citation

Jones, C., Clayton, S., Ribalet, F., Armbrust, E. V., & Harchaoui, Z. (2021). A kernel-based change detection method to map shifts in phytoplankton communities measured by flow cytometry. *Methods in Ecology and Evolution*, *12*(9), 1687-1698. https://doi.org/10.1111/2041-210X.13647

Received: 3 December 2020 | Accepted: 6 May 2021

DOI: 10.1111/2041-210X.13647

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

# A kernel-based change detection method to map shifts in phytoplankton communities measured by flow cytometry

Corinne Jones[1] | Sophie Clayton[2] | François Ribalet[3] | E. Virginia Armbrust[3] | Zaid Harchaoui[4]

[1]Swiss Data Science Center, École polytechnique fédérale de Lausanne, Lausanne, Switzerland

[2]Department of Ocean and Earth Sciences, Old Dominion University, Norfolk, VA, USA

[3]School of Oceanography, University of Washington, Seattle, WA, USA

[4]Department of Statistics, University of Washington, Seattle, WA, USA

**Correspondence**
Corinne Jones
Email: corinne.jones@epfl.ch

**Funding information**
Washington Research Foundation; Alfred P. Sloan Foundation; Gordon and Betty Moore Foundation; Simons Foundation, Grant/Award Number: 329108, 426570SP, 549894 and 574495; Division of Mathematical Sciences, Grant/Award Number: DMS-1810975; Canadian Institute for Advanced Research

**Handling Editor:** Hao Ye

## Abstract

1. Automated, ship-board flow cytometers provide high-resolution maps of phytoplankton composition over large swaths of the world's oceans. They therefore pave the way for understanding how environmental conditions shape community structure. Identification of community changes along a cruise transect commonly segments the data into distinct regions. However, existing segmentation methods are generally not applicable to flow cytometry data, as these data are recorded as 'point cloud' data, with hundreds or thousands of particles measured during each time interval. Moreover, nonparametric segmentation methods that do not rely on prior knowledge of the number of species are desirable to map community shifts.
2. We present CytoSegmenter, a kernel-based change-point estimation method for segmenting point cloud data. Our method allows us to represent and summarize a point cloud of data points by a single element in a Hilbert space. The change-point locations can be found using a fast dynamic programming algorithm.
3. Through an analysis of 12 cruises, we demonstrate that CytoSegmenter allows us to locate abrupt changes in phytoplankton community structure. We show that the changes in community structure generally coincide with changes in the temperature and salinity of the ocean. We also illustrate how the main parameter of CytoSegmenter can be easily calibrated using limited auxiliary annotated data.
4. CytoSegmenter is generally applicable for segmenting series of point cloud data from any domain. Moreover, it readily scales to thousands of point clouds, each containing thousands of points. In the context of flow cytometry data collected during research cruises, it does not require prior clustering of particles to define taxa labels, eliminating a potential source of error. This represents an important advance in automating the analysis of large datasets now emerging in biological oceanography and other fields. It also allows for the approach to be applied during research cruises.

**KEYWORDS**
change-point detection, community structure, flow cytometry, nonparametric statistics, phytoplankton community, point cloud data

## 1 | INTRODUCTION

Determining the number and locations of abrupt changes in distribution of a sequence of observations has played an important role in analyzing ecological data. The study of change points in marine ecology and oceanography dates back several decades. Legendre et al. (1985) performed change-point detection on time series of zooplankton counts near the Mediterranean Sea and in a reservoir in Quebec in the late 1960s and 1970s, respectively. Since then, change points in time have been studied at a wide range of spatial and temporal scales. Spatially, this includes large-scale changes in environmental and/or biological variables measured in the North Pacific and North Atlantic (Friedland et al., 2016; Hare & Mantua, 2000; Mantua et al., 1997) and smaller scale changes in the abundance of species in seas, lakes and estuaries (Alvarez-Fernandez et al., 2012; Gal & Anderson, 2010; Thomson et al., 2010; Weijerman et al., 2005). The time scale between change points depends on the causes of the changes. For example, studies have examined interdecadal changes in the North Pacific related to changes in the climate (Mantua et al., 1997), month-long changes caused by eutrophication (Pace et al., 2017) and daily changes due to vertical migration (Bianchi & Mislan, 2016).

While all of the aforementioned literature focuses on change points in time, it is equally possible to detect change points in space. For example, Nieuwhof et al. (2018) detected changes in water storage capacity as the distance from shellfish reefs increased, and Li et al. (2019) identified the boundary of the Antarctic Intermediate Waterway. In this work, we draw on ideas from the nonparametric statistics and machine learning literature and develop a retrospective change-point estimation method for series of point cloud data that can be applied to segment flow cytometry data. Point cloud data here are a series of observations, where each observation (occurring at a given time point) consists of a large collection of measurements as opposed to a single measurement.

Over the past three decades, flow cytometry has become instrumental in determining the distribution of phytoplankton communities (Sosik et al., 2010). Flow cytometry measures light scattering and fluorescence emissions of individual cells at rates of up to thousands of cells per second. Light scattering is proportional to cell size, and fluorescence is unique to the emission spectra of pigments; together, these parameters can be used to identify populations of phytoplankton with similar optical properties. Automated flow cytometers such as CytoBuoy (Dubelaar et al., 1999), FlowCytobot (Olson et al., 2003) and SeaFlow (Swalwell et al., 2011) have provided unprecedented views of the dynamics of phytoplankton over multiple scales that vary from basin level to mesoscale to microscales. It is now recognized that phytoplankton populations are organized in sub-mesoscale patches (10–100 km) separated by physical fronts induced by horizontal stirring and separating water masses of different origin (d'Ovidio et al., 2010). Consequently, a change in the phytoplankton assemblage is expected to occur when different water masses mix together (Ribalet et al., 2010). The recent release of SeaFlow data collected underway during cruises[1] conducted in the North Pacific Ocean offers unique opportunities to study how phytoplankton communities vary over multiple spatial and temporal scales (Ribalet et al., 2019).

Performing change-point analysis on underway flow cytometry data would allow us to segment a cruise into statistically distinct regions based on measures such as the scatter and fluorescence of individual particles. However, the underway flow cytometry data produced by instruments such as SeaFlow present novel challenges not addressed by existing segmentation methods. Seaflow continually takes measurements of individual phytoplankton, and a batch of thousands of measurements is recorded for each 3-min time period (roughly at a 1 km spatial resolution). Hence, the data corresponding to one 3-min time period can be viewed as a point cloud of phytoplankton measurements. Moreover, as the dataset collected during any given research cruise might contain more than 50 million phytoplankton measurements, the method must scale.

Methods that have been used in ecology and oceanography for detecting change points generally fall into one of the three categories. First, there are methods that repeatedly perform hypothesis tests at every location in the time series and set a testing threshold to yield an appropriate number of change points (Clarke, 1993; Matteson & James, 2014; Page, 1954). Second, there are methods that estimate the locations of a large number of potential change points and then prune them using a penalty term or hypothesis testing (Gordon & Birks, 1972). Finally, there are methods that fit a single model to the time series that allows for an unknown number of changes at unknown times (Fearnhead, 2006; Goldfeld & Quandt, 1973; Hamilton, 1990). However, nearly all of these approaches assume that the data are either real-valued or vectorial, that is, that each observation is a vector of fixed dimension. Applying such methods by reducing each point cloud to its mean as done by Hyrkas et al. (2015) results in the loss of important information regarding the distribution of the data in each point cloud. Of the cited methods, only the dissimilarity-based approaches of Clarke (1993) and Gordon and Birks (1972) could potentially be applied to data consisting of sequences of point clouds. Even if these methods were used, an appropriate dissimilarity measure would still need to be chosen and a principled method for determining the number of changes would need to be proposed.

In this work, we develop a nonparametric statistical method for identifying abrupt changes in large-scale series of point clouds of measurements. The method is nonparametric in that it does not require the modeller to specify a parametric family of probability distributions. It therefore has the advantage that, given enough samples, it can detect

[1]We henceforth refer to such data as 'underway data'.

any change in probability distribution. Moreover, it is capable of scaling to sequences of thousands of point clouds, each with thousands of points. We demonstrate the approach by segmenting SeaFlow data along cruise tracks. We also assess how closely estimates of biological change points coincide with physical shifts to better understand the role of the physical ocean environment in controlling phytoplankton community structure and to estimate the number of biological change points.

## 2 | MATERIALS AND METHODS

### 2.1 | SeaFlow data

To date, the SeaFlow flow cytometer has been used in more than 60 cruises in the North Pacific Ocean. In this work, we focus on 12 of those cruises. These cruises are all of the cruises that are neither coastal cruises nor cruises around the Hawaiian islands and for which data are available. The diverse range of locations of the cruises is depicted in Figure 1 and some characteristics of the cruises may be found in Table 1. Both physical and biological data were collected from the ship's flow-through seawater system (intake between 4 and 7 m depth, depending on the ship and sea state). Note that change points in some cruises (e.g. KOK1606) can occur more than twice as often as in others (e.g. KM1713).

The data we use consist of two components. First, there is physical data, which contain measurements of sea surface temperature and salinity collected at 3-min intervals throughout the cruises from the ship's underway thermosalinograph system. Second, there is biological data, which contain measurements of light scatter and fluorescence emissions of individual particles. The biological data are organized into point clouds resolved to $2^{16}$ discrete values recorded every 3 min. The light intensity measured by a photomultiplier tube (PMT) is converted via a pre-amplifier to a voltage that is then input into a logarithmic amplifier to allow for the visualization of at least four decades of dynamic range on a single graph. It also makes log-normal distributions, common to biological systems, appear more symmetric. The upper bound in light scatter and fluorescence intensity values, which indicates saturation of the photo-multiplier, depends on the degree of signal amplification that is controlled by the PMT voltage set by the operator. It therefore can vary slightly from cruise to cruise. Each post-processed point cloud contains measurements of between 100 and 10,000 particles ranging from 0.5 to 5 microns in diameter. The volume of data in any given file depends on the total abundance of phytoplankton within the sampled region. Each particle is characterized by two measures of fluorescence emission (chlorophyll and phycoerythrin), its light scatter and its label (identified based on a combination of manual gating and a semi-supervised clustering method) (Ribalet et al., 2019). Note that we use the particle labels only for verification of our approach.

The data are cleaned as follows. For the physical data, we first remove the observation times that are not in chronological order. Next, we discard observations for which the temperature and salinity are unavailable or are not between the 1st and the 99th percentiles across all cruises (i.e. between 5°C and 30°C for temperature and 29 PSU and 38 PSU for salinity). For the biological data, we exclude files for which the physical data are not available. We also delete the entries corresponding to added calibration beads rather than phytoplankton. Prior to performing the analysis, we take the log (base 10) of the biological data and standardize the physical and biological data separately for each cruise.

One downside to this dataset is that there are no ground-truth change points in either the biological data or the physical data that could be used to evaluate our method. To ameliorate this problem, we manually annotated change points in the physical data using the following reasoning. Shifts in phytoplankton community structure often occur where different water masses, with distinct temperature and salinity properties, meet (d'Ovidio et al., 2010). As such, large changes in water temperature and salinity measured from a travelling vessel

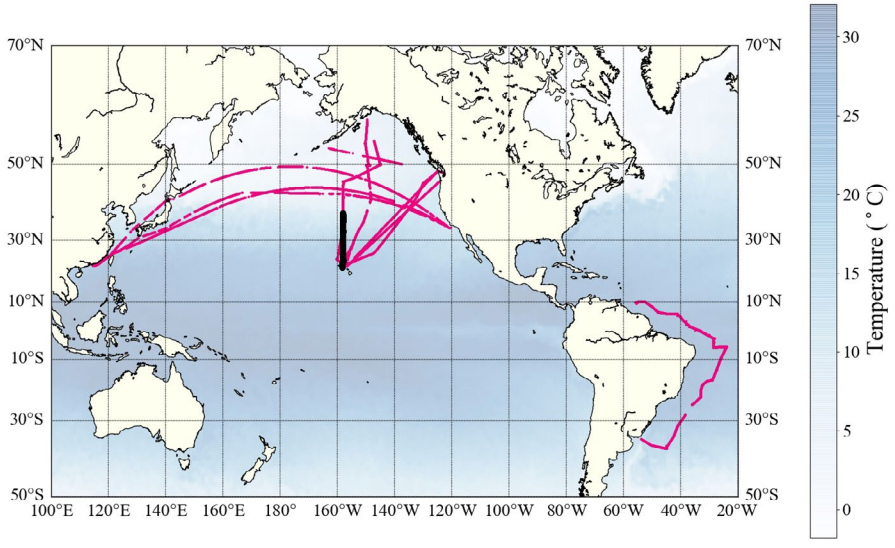


FIGURE 1 Locations of the cruises analyzed in this paper, overlaid on sea surface temperature data from April 26, 2016.[2] In this work, we primarily focus on KOK1606, the cruise in black in the middle of the map, which took place from April 20 to May 4, 2016

[2]NOAA OI SST V2 data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their website at https://psl.noaa.gov/.

| Cruise | Location | Length (km) | # point clouds | # particles | # changes |
|---|---|---|---|---|---|
| KM1502 | Portland–Hawaii | 3,987 | 3,711 | 6.4 million | 43 |
| KM1712 | Hawaii–Alaska | 5,910 | 10,468 | 12.5 million | 48 |
| KM1713 | Alaska–Hawaii | 4,458 | 8,959 | 13.7 million | 27 |
| KN210-04 | South Atlantic | 14,101 | 18,423 | 57.0 million | 92 |
| KOK1606 | Subtropical front | 4,005 | 6,685 | 2.3 million | 59 |
| MGL1704 | Subtropical front | 4,939 | 6,794 | 5.8 million | 32 |
| TN248 | Gulf of Alaska | 1,802 | 943 | 0.1 million | 28 |
| TN271 | Seattle–Hawaii | 4,527 | 3,566 | 1.0 million | 41 |
| TN292 | Seattle–Hawaii | 3,864 | 3,086 | 2.8 million | 29 |
| Tokyo_1 | North Pacific | 11,383 | 4,872 | 0.9 million | 66 |
| Tokyo_2 | North Pacific | 11,481 | 5,145 | 1.5 million | 40 |
| Tokyo_3 | North Pacific | 11,663 | 5,772 | 3.2 million | 118 |

**TABLE 1** Details of the cruises used. The column '# changes' provides the number of change points in the annotated physical data

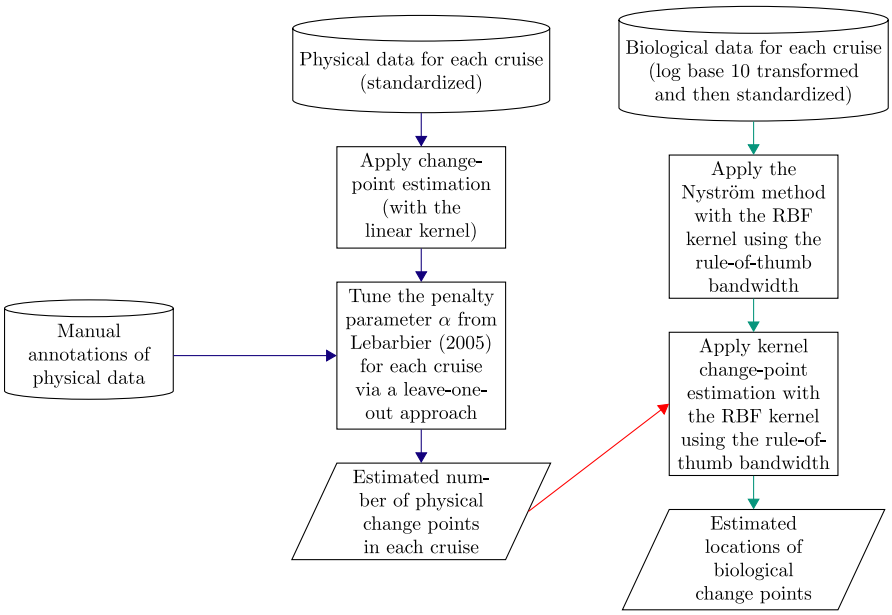


**FIGURE 2** Outline of the proposed approach

are diagnostic of travelling from one water mass to another. In this study, we examined the underway temperature and salinity data measured at around 5-m depth and identified such water mass shifts as occurring where the temperature varied by >2°C and/or the salinity varied by >0.4 PSU over a distance of 20 km or less. For each cruise, these transitions between water masses were manually annotated as the halfway point along lines that connect two clusters of points in a temperature–salinity diagram. This is illustrated in Appendix S1.

## 2.2 | Change-point analysis on point clouds

The goal of change-point estimation here is to locate changes in distribution of an ordered sequence of point clouds $x_1, \ldots, x_T$ from a given cruise. For each time index $t$, the point cloud $x_t = \{x_{t,1}, \ldots, x_{t,nt}\}$ consists of $n_t$ points (particle measurements) $x_{t,i} \in \mathbb{R}^d$. Changes in distribution often occur in the mean of the point clouds (e.g. the average forward scatter and chlorophyll of particles we measure might suddenly increase and the distribution of phycoerythrin might remain the same) but can also occur in higher-order moments of the data such as the variance (e.g. the variance of the forward scatter, chlorophyll and phycoerythrin of particles we measure might all suddenly increase, even though their means might remain the same). Therefore, assume the points within each point cloud $x_t$ are samples from an unknown probability distribution $\mathbb{P}_t$. We seek to estimate the times at which the distributions $\mathbb{P}_t$ change, that is, the times $t$ where $\mathbb{P}_t \neq \mathbb{P}_{t+1}$, to divide the sequence of point clouds into segments that are homogeneous in probability distribution.

Henceforth, we will denote the (unknown) number of change points by $m$ and the change-point times by $t_1 < t_2 < \ldots < t_m$. We furthermore define $t_0 := 1$ and $t_{m+1} := T + 1$. We will now present CytoSegmenter, our approach to change-point analysis on sequences of point clouds. A high-level overview is provided in Figure 2. For a more detailed version, see Figure S2.

### 2.2.1 | Change-point estimation

A natural way of segmenting a sequence of point clouds into a fixed number of parts is to define a notion of similarity between point clouds, and to divide the sequence into contiguous sets of similar point clouds. We will define such a similarity measure using the notion of a positive semi-definite (PSD) kernel. Here a PSD kernel is a function $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for some non-empty set $\mathcal{X}$ such that there exists a Hilbert space $\mathcal{H}$ (i.e. a complete metric space with an inner product) and a map $\phi: \mathcal{X} \to \mathcal{H}$ such that for all $x, x\prime \in \mathcal{X}$, $k(x, x\prime) = \langle \phi(x), \phi(x\prime) \rangle_{\mathcal{H}}$ (Shawe-Taylor & Cristianini, 2004). The notation $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes an inner product in $\mathcal{H}$. There are numerous benefits to the kernel viewpoint. For example, this approach can be applied to generic sets $\mathcal{X}$, so long as a PSD kernel can be defined on $\mathcal{X} \times \mathcal{X}$. Moreover, for kernels like the one we will use, we do not need to make any assumptions about the distribution of the data. Numerous examples of PSD kernels can be found in Shawe-Taylor and Cristianini's (2004) book. In this work, we will use the radial basis function (RBF) kernel $k_p: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $k_p(x_t, x_{t\prime}) = \exp\left(-d(x_t, x_{t\prime})^2 / \left(2\sigma_y^2\right)\right)$, where $d: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a distance measure between two point clouds, $\mathcal{X}$ is the space of point clouds and $\sigma_y > 0$ is the bandwidth of the kernel. When the distance between two point clouds is small, the point clouds are more similar, and when the distance is large, the point clouds are less similar.

The key then lies in defining the distance between a pair of point clouds. One way to look at a point cloud is to see it as an empirical probability density. One could then define the distance between two point clouds as the distance between nonparametric density estimates of the point clouds (Anderson et al., 1994). Gretton et al. (2012) extend this idea, defining a distance called the maximum mean discrepancy (MMD) between two sets of points. The MMD uses another kernel, $k_x: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, defined this time on a pair of points rather than a pair of point clouds. In particular, the squared MMD distance between two point clouds $x_t$ and $x_{t\prime}$ is defined as

$$d(x_t, x_{t\prime})^2 = \frac{1}{n_t^2} \sum_{i,j=1}^{n_t} k_x(x_{t,i}, x_{t,j}) - \frac{2}{n_t n_{t\prime}} \sum_{i=1}^{n_t} \sum_{j=1}^{n_{t\prime}} k_x(x_{t,i}, x_{t\prime,j}) + \frac{1}{n_{t\prime}^2} \sum_{i,j=1}^{n_{t\prime}} k_x(x_{t\prime,i}, x_{t\prime,j}). \quad (1)$$

The first and third terms measure the similarity of points within the two point clouds while the second term assesses the similarity of points across the two point clouds. When the second term is smaller in magnitude relative to the other two terms, that is, the points across point clouds are relatively dissimilar, the distance is larger. In practice, we will take $k_x$ to be another RBF kernel, defined as $k_x(x_{t,i}, x_{t\prime,i\prime}) = \exp(-||x_{t,i} - x_{t\prime,i\prime}||^2/(2\sigma_x^2))$, where $\sigma_x > 0$ is the bandwidth of the kernel. This kernel satisfies the assumptions required for the theory of kernel embeddings of distributions to hold (Gretton et al., 2012). The quantity (1) can be related to the notion of a mean element; see Appendix S2.1.

Having defined a notion of similarity, we now consider estimating the locations of a fixed number, $m$, of change points. To do so, we solve the following kernel change-point problem to locate the change points $t_1, \ldots, t_m$:

$$\min_{t_1,\ldots,t_m} \frac{1}{T} \sum_{j=0}^{m} \left[ \sum_{t=t_j}^{t_{j+1}-1} k_p(x_t, x_t) - \frac{1}{t_{j+1} - t_j} \sum_{t,s=t_j}^{t_{j+1}-1} k_p(x_t, x_s) \right]. \quad (2)$$

This can be viewed as minimizing a least-squares criterion (Harchaoui & Cappé, 2007). Intuitively, the terms within the brackets assess how similar point clouds within a proposed segment are: the first sums how similar each point cloud is to itself and the second subtracts how similar each point cloud is, on average, to the other point clouds in the segment $\left\{x_{t_j}, \ldots, x_{t_{j+1-1}}\right\}$.

The optimization can be performed efficiently using a dynamic programming algorithm similar to the one described by Kay (1993, Ch. 12). A straightforward implementation of the dynamic programming algorithm for a fixed $m$ would make $O(T^2)$ evaluations of the objective (2), which itself can be evaluated with a time complexity $(max_t\, n_t)^2$. To scale the approach to long series of data, we propose to use an approximation scheme, the Nyström method (Williams & Seeger, 2000), to approximate the kernel evaluations in (1). This leads to a complexity in time that is linear in the number of points per point cloud rather than quadratic, allowing us to analyze series of tens or hundreds of millions of data points. For details on this approximation scheme, see Appendix S2.2.

### 2.2.2 | Parameter calibration and domain knowledge

We leverage auxiliary data with change-point annotations to set the number of change points in each sequence of point clouds. We assume here that we have (a) a parallel vectorial time series in which the number of change points is unknown but expected to be the same as in the time series of interest (in this paper, the physical data from the same cruise) and (b) a set of $S$ similar vectorial time series for which the numbers of change points are known (in this paper, physical data from other cruises). The idea is that it can be easier to obtain change-point labels for time series of vectorial data than for time series of point cloud data.

We first use the $S$ vectorial time series in (b) to calibrate the penalty parameter $\alpha$ from the penalty $pen(\alpha, m) = \alpha(m+1)(2\log(T/(m+1)) + 5)/T$ in the penalized change-point problem of Lebarbier (2005). We set the value of $\alpha$ so that, across all $S$ sequences, the difference between the annotated number of change points and the estimated number of change points in these vectorial sequences is minimized. Using this estimated $\alpha$, we then run the method of Lebarbier (2005) on the corresponding parallel series of vectorial data in (a). This provides us with an estimate of the number of change points $m$ in the corresponding sequence of point clouds, which we use when minimizing the objective in (2). For more details, see Appendix S2.3.

To assess the quality of the resulting estimated change points, we can compute the distance from each biological change point to the nearest annotated physical change point. It is reasonable to expect that to first order, shifts in phytoplankton community structure (biological change points) should occur where different water masses meet (physical change points). In reality, due to physical mixing

processes and lags in biological responses to changes in the physical environment, biological change points may not coincide exactly with physical change points. That said, if the majority of biological change points occur within a short distance of the nearest physical change point, this acts as a useful check on the validity of our method. For a given cruise, if $t_1, \ldots, t_m$ and $t'_1, \ldots, t'_{m'}$ are the estimated and annotated change points, respectively, then in terms of distance travelled (by abuse of notation), we could compute for each estimated change point $t_j$ the value $\min_{\ell=1,\ldots,m'} \left|t_j - t'_\ell\right|$. However, change points can be encountered at different rates during different cruises. Therefore, we normalize the aforementioned distances by the average distance between change points in each cruise. That is, for each estimated change point $t_j$, we compute $\min_{\ell=1,\ldots,m'} \left|t_j - t'_\ell\right| / (T/(m'+1))$, where $T$ is the total distance travelled in the corresponding cruise.

The other parameters that need to be calibrated are the parameters of the two kernels we use. For each RBF kernel, we set the bandwidth to the median pairwise distance between inputs, a common rule of thumb. We perform the Nyström approximation with the projection of the kernel onto a subspace of size 128 (chosen as a reasonable balance between the accuracy of the kernel approximation and the runtime). We select the quadrature points by quantizing the data into a codebook of size 128 using 100 iterations of $k$-means. To target phytoplankton community shifts associated with larger-scale oceanographic features, such as mesoscale eddies (~100 km) and gyre boundaries, and to avoid generating a large number of change points associated with high-frequency variability, we set the minimum distance between change points to be five samples, which represents 15 min or roughly 5 km for a ship moving at 10 knots.

### 2.2.3 | Code

The code for this paper, written in Python, builds upon Scikit-learn and Faiss (Johnson et al., 2019; Pedregosa et al., 2011). Code for the Nyström method and the kernel change-point estimation algorithm may be found online at https://github.com/cjones6/chapydette while the code to reproduce the results in this paper is located at http://github.com/cjones6/cytosegmenter. All of the analyses, including the feature generation and change-point estimation, take approximately 22 min in total to run on a machine with an Intel i9-7960X processor, an Nvidia Titan Xp GPU, and 128GB of memory.

## 3 | RESULTS

In applying CytoSegmenter, the kernel change-point method described in Section 2.2, to SeaFlow data, we aim to answer the following questions:

1. Does the method successfully estimate changes in the distribution of phytoplankton communities?
2. Do the estimated change points in the biological data coincide with change points in the physical environment?
3. Can we leverage additional data on the physical environment from other cruises to estimate the number of change points?

For the first two questions, we focus on a single cruise, KOK1606. As the number of change points in a cruise can be subjective, we initially fix the number of change points to 10. For the third question, we expand the analysis to all cruises and estimate the number of change points in each cruise.

### 3.1 | Changes in distribution

We first aim to assess whether CytoSegmenter successfully estimates changes in the distribution of phytoplankton along the course of a cruise. To do so, we run the change-point estimation method on the biological measurements taken on the KOK1606 cruise when fixing the number of change points to 10. The locations of the resulting estimated change points are non-uniformly distributed across the cruise track (Figure 3).

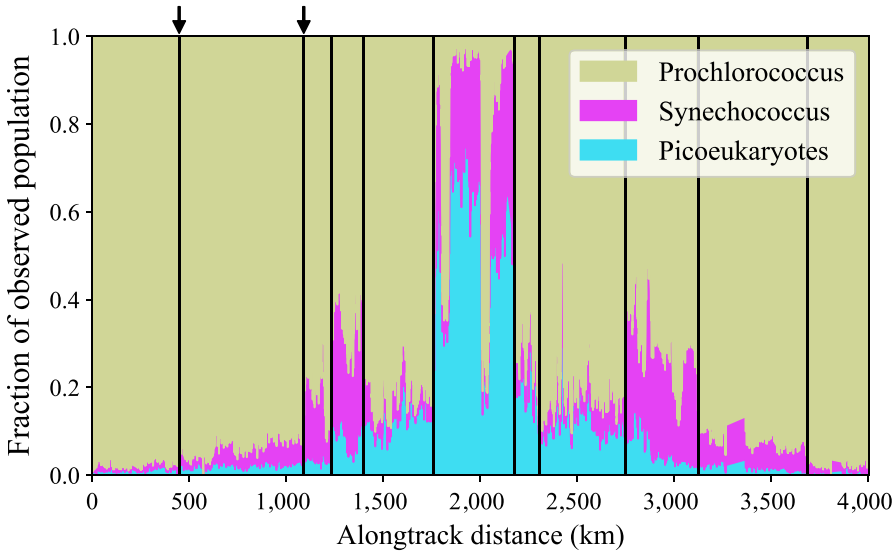


**FIGURE 3** Estimated change points in the biological data from the KOK1606 cruise overlaid on the phytoplankton distribution observed during the cruise when the number of change points is fixed at 10. For reference, with 11 change points the next change point would be at 2078 km. The arrows at the top of the figure indicate the change-point locations examined in Figure 4

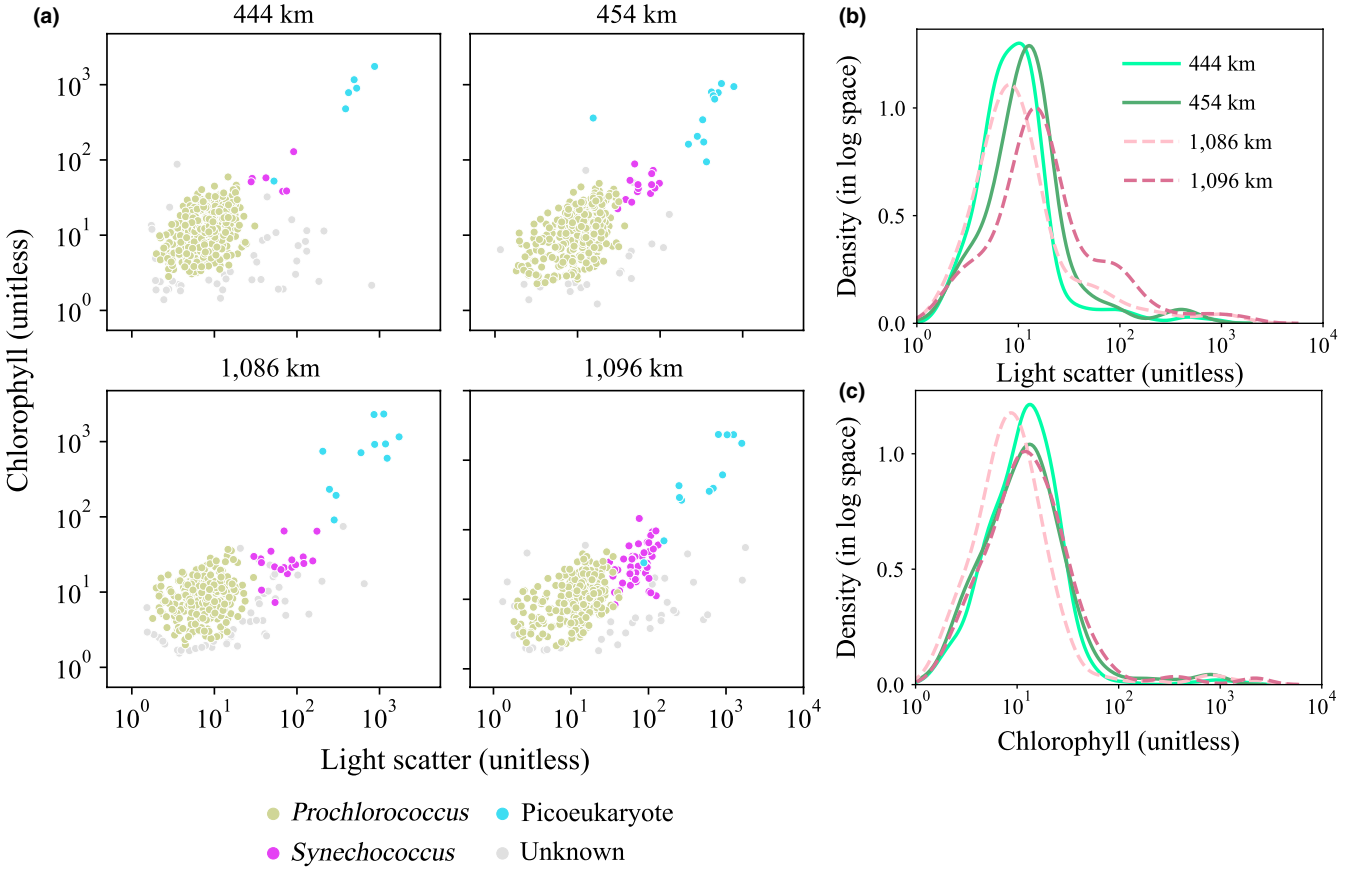


**FIGURE 4** Distribution of the chlorophyll measurements and the light scatter measurements of phytoplankton five kilometers before and after estimated change points at 449 km and 1,091 km along the cruise track (KOK1606 in Table 1)

We assess the quality of the estimated change points in three ways. First, Figure 3 overlays the estimated change points on a plot of the phytoplankton distribution observed at each 3-min time period throughout the course of the cruise. It illustrates that the algorithm locates large, abrupt changes in the phytoplankton community structure. For example, the algorithm detected the large increase from 5% *Synechococcus* to 18% *Synechococcus* at approximately 1,091 km along the cruise track.

Next, for the change points at 449 and 1,091 km, we plot in Figure 4 the distribution of the chlorophyll and the light scatter measurements 5 km before and after the estimated change points. Recall that the cell labels were not available to our method; only the cell measurements were. There is a sudden increase between both 444 and 454 km and between 1,086 and 1,096 km in the fraction of cells that have light scatter between $10^{1.5}$ and $10^{2.5}$ and chlorophyll fluorescence between $10^{1}$ and $10^{2}$. From the labels, we can see that this corresponds to an increase in the fraction of *Synechococcus*. We have included in Figure S3 the analogous plots with phycoethryin and light scatter. The changes in distribution are also visible in the phycoethryin versus light scatter space. For comparison, we also add in Appendix S3.1 the corresponding chlorophyll and light scatter plots for two times at which we did not detect change points. The differences in the univariate densities are much more pronounced around the estimated change points than around these times at which we did not estimate a change point. These changes can also be seen by examining how the average similarity between points in a point cloud differs before and after change points; see Figures S5–S7.

Finally, a distinctive feature of the KOK1606 cruise is that the northward and southward cruise trajectories were nearly identical, occurring over a period of 3 weeks (see Figure 5, which displays the track of the research vessel with the estimated change points overlaid). This provided us with a method to check the efficacy and consistency of our method, as we expect that change points detected on the northbound transect should be similarly detected on the southbound leg. This is generally the case, as four pairs of change points are within 61 km of each other. Given that the ocean is a dynamic environment subject to constant movement driven by ocean currents, it is likely that the detected change points could shift in space. Surface mean current speeds in the region of the Pacific sampled by the KOK1606 cruise are ~0.1 m/s (Lumpkin & Johnson, 2013), which could drive a displacement of ~60 km over the course of a week. Spatial shifts in the ocean temperature during the cruise (cf. Figure 6) support this hypothesis. The close spacing between our outbound and return change points confirms that our method is able to detect change points sampled more than once.

In Appendix S3.2, we include the results of a sensitivity analysis of the parameters of our method. We find that the results are rather insensitive to the parameter values we set: for projections of

dimension 8 through 128, the estimated change points vary by at most 1 km. Moreover, the RBF kernel with the rule-of-thumb bandwidth and the linear kernel yield change-point estimates that disagree by at most 2 km. When varying the bandwidth within a wide range, from 0.058 to 0.63, eight of the ten change points differ by at most 5 km. Finally, the minimum distance between change points had no impact.

We include in Appendix S3.3, a study of the variance of the estimated change point locations. For this, we divided the cruise into 10 disjoint subsamples, obtained by specifying a starting index between 1 and 10 and then taking every 10th point cloud thereafter. Across all subsamples, there are only 11 different intervals of width at most 37 km in which the change points lie. This suggests that the algorithm is quite confident in the locations of 9 of the 10 originally estimated change points.

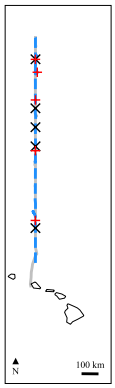


**FIGURE 5** Estimated change points in the biological data from the KOK1606 cruise overlaid on the track of the ship. Beginning in Hawaii, the ship travelled north (silver solid line) and then turned around and went south (blue dashed line). The change points detected on the trip north are marked with black crosses while the change points detected on the trip south are marked with red pluses

## 3.2 | Correspondence between biological and physical change points

To better understand the controls on phytoplankton distributions, we need to gain a better understanding of the balance between physical and biological controls in driving shifts in species' distributions, which are reflected as shifts in overall phytoplankton community structure. By comparing the location of physical change points (based on surface temperature and salinity) and biological change points (based on SeaFlow data), we can begin to understand how important changes in the physical environment control phytoplankton community structure. The left side of Figure 6 plots the

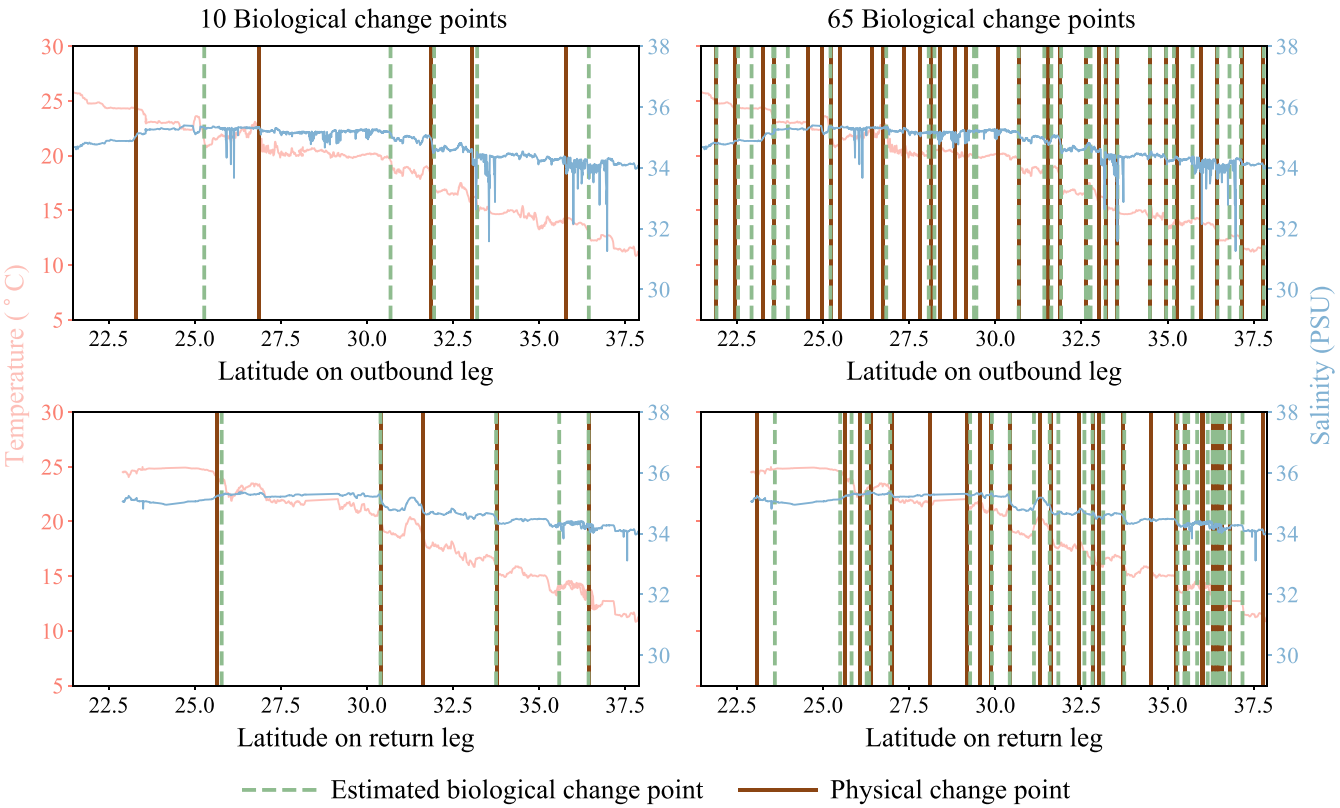


**FIGURE 6** Change points in the biological and physical data from the KOK1606 cruise overlaid on the temperature and salinity data recorded during the cruise. The two plots on the left display the estimated biological and estimated physical change points when fixing the number of change points to 10. The two plots on the right display the estimated biological change points when the number of change points is 65 (the number our method estimated), along with the 59 manually annotated physical change points

estimated change points overlaid on the temperature and salinity measurements throughout the cruise when the number of change points in both the physical and biological data is fixed to 10. The estimated physical and biological change points are within 20 km of each other 60% of the time. However, even when they do not quite coincide, the estimated biological change points are associated with large changes in temperature and salinity. These results suggest that shifts in phytoplankton community structure are largely associated with corresponding shifts in physical ocean properties. The results support previous work that showed that water masses play an important role in structuring phytoplankton communities (e.g. d'Ovidio et al., 2010).

### 3.3 | Estimation of the number of change points

Now, we examine the results from estimating the number of change points. For each of the 12 cruises from Table 1, we estimate the parameter $\alpha$ based on the other 11 cruises, thereby obtaining one value of $\alpha$ per cruise. When considering the grid of $\alpha$'s 0, 0.01, 0.02, ..., 1 in the penalty of Lebarbier (2005), the value chosen is always either 0.12 or 0.13. The corresponding number of estimated physical change points for each cruise is presented in the left panel of Figure 7. In this case, the correlation between the number of estimated change points and the number of annotated change points is 0.62. Moreover, the number of estimated change points for 9 of the 12 cruises is within a factor of two of the number of annotated change points. This suggests that when annotated physical data for a given cruise are unavailable, the number of change points might be reasonably estimated from annotations of other cruises. In Appendix S3.4, we compare this approach to that of Harchaoui and Lévy-Leduc (2007) and demonstrate the superiority of this approach.

Using the estimated number of change points in the physical data from each cruise, we estimate the locations of change points in the biological data from each cruise. The right panel of Figure 7 plots a histogram of the normalized distances from each estimated biological change point to the nearest annotated physical change point. Approximately 52% of the estimated change points are within a normalized distance of 0.25 to the nearest annotated physical change point. In contrast, if we uniformly segmented each cruise (i.e. put the change points at equally spaced intervals), this value would only be 43%. In fact, 21% of the estimated change points are within a normalized distance of 0.05 and 33% are within a normalized distance of 0.1, compared to 9% and 18% based on a uniform segmentation. The right side of Figure 6 displays the estimated biological and annotated physical change points for the KOK1606 cruise. The number of estimated change points is 65, whereas the number of annotated change points is 59. In contrast to the 10 change-point case, most of the locations where there are large changes in temperature and/or salinity are now labelled as change points.

## 4 | DISCUSSION

In this work, we presented a kernel-based change-point detection method for point cloud data. We applied the method to data collected by a shipboard flow cytometer during research cruises to determine where changes in phytoplankton structure occur. The distribution of individual phytoplankton species is a reflection of that species' environmental niche (Hutchinson, 1957), defined as the range of environmental parameters within which a species can

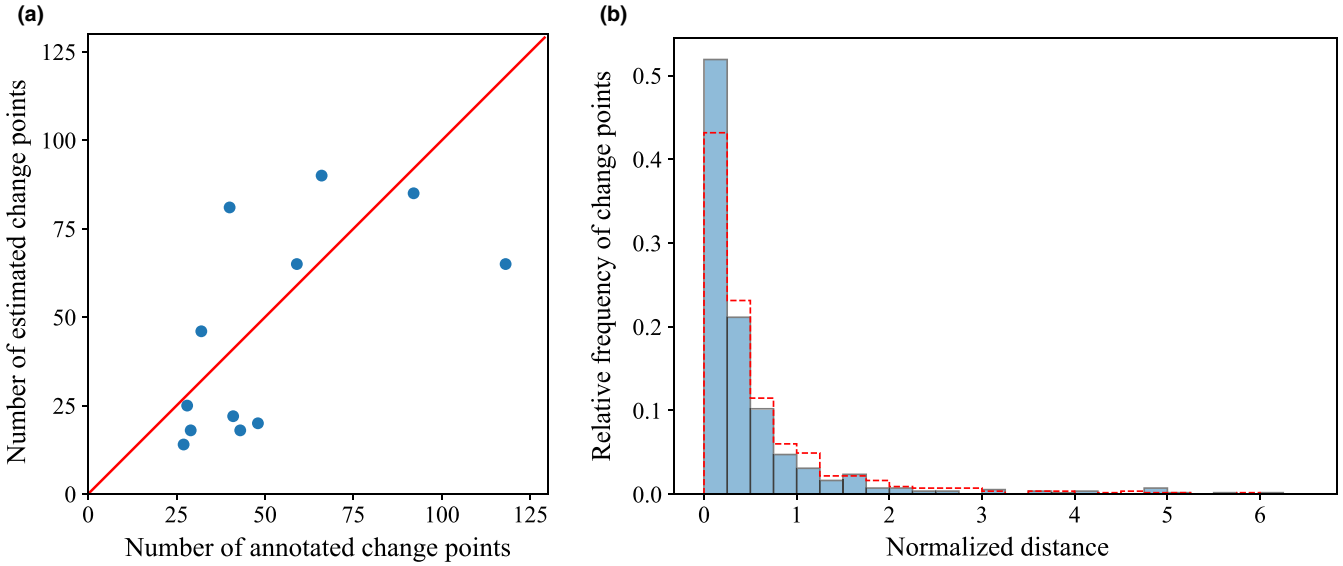


**FIGURE 7** Estimated and annotated number of change points on the physical data (left) and histogram of the distances from each estimated biological change point to the nearest annotated physical change point for the same cruise, normalized by the average distance between annotated change points within the cruise (right). The diagonal red line in the plot on the left denotes the locations where the points would ideally lie. The red dashed line in the plot on the right indicates the histogram one would obtain from uniformly segmenting the cruises

subsist. Niches can be controlled by physicochemical factors such as temperature and nutrient availability, as well as biotic processes such as competition and predation. Previous studies have shown that changes in phytoplankton assemblages are often associated with physical fronts that separate water with different temperature and salinity properties (d'Ovidio et al., 2010) and consequently, an abrupt change in the phytoplankton assemblage is expected to occur across physical fronts (Ribalet et al., 2010). Here we found that changes in the distribution of phytoplankton generally co-occur with changes in temperature and salinity, suggesting that physical processes are driving the observed community shifts. A lack of a lag between physical and biological change points suggests that the physical changes were not associated with changes in nutrient availability or biotic process or that the formation of the physical fronts was recent and the phytoplankton did not have the time to adjust to the changes in environmental conditions. There were instances where biological change points did not coincide with physical change points. This could indicate the presence of a persistent physical front between water masses at a given location, or that water masses have been in contact and mixing for some time, which would allow for a larger spatial lag between physical and biological change points due to the temporal component of the phytoplankton response to that change. These results suggest that the method is able to locate meaningful changes in the distribution of phytoplankton, paving the way for estimating the number of change points in the biological data based on the number of change points in the physical data.

A previously proposed approach to the detection of changes in flow cytometry data is that of Hyrkas et al. (2015). This approach directly averages the points within each point cloud to obtain a single vector, which results in a significant loss of information. In contrast, the approach we propose here accounts for all the moments of a point cloud distribution, hence capturing richer statistical information. In Appendix S3.5, we discuss a case where CytoSegmenter located a noteworthy change point that the change-in-mean method failed to detect. While we focused on identifying changes in *distribution* in a sequence of multi-dimensional point clouds, there may be also times in which one would want to locate both changes in distribution and changes in abundance (i.e. the total number of points per point cloud). This can be done by substituting the kernel on point clouds, $k_x$, with a weighted average of $k_x$ and a kernel $k_c : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ on abundances, for example, $k_c\left(n_t, n_{t'}\right) := min\left(n_t, n_{t'}\right) / max\left(n_t, n_{t'}\right)$.

Three classically difficult facets of change-point analysis are determining the number of change points, assessing the uncertainty of the estimated change points, and handling gradual changes. Our approach estimates the number of change points based on annotations of the physical data in other cruises. One could alternatively use annotations of the physical data from the cruise of interest, if available, or apply methods that would work directly on the biological data. The latter approach can be impeded by an over-reliance on the data from the cruise of interest and a lack of generalizability to a wide range of datasets. To assess the uncertainty of the change-point estimates, we subsampled the cruise. Then we applied the algorithm independently on each subsample, keeping the time ordering within each subsample consistent with the one in the original sample. How to construct confidence intervals around the change points and perform hypothesis testing after the estimation remain active research topics. Finally, our method is designed to detect abrupt changes, and therefore when there is a gradual change it tends to put a single change point in the middle of the change. Simultaneously detecting both abrupt changes and the start and end of gradual changes is left for future work.

Point cloud data also arise in other areas such as computer graphics and computer vision (see, e.g. Suard et al., 2005). Here we have applied this method to identify shifts in phytoplankton community structure from a large dataset of underway flow cytometry measurements. We envision that this statistical framework will be widely applicable to datasets structured in a similar way. A range of broadly similar continuously sampling flow cytometers are currently commercially available (CytoSense, Imaging FlowCytobot and FlowCam) and collect cytometric data (e.g. scatter and fluorescence) as well as images of plankton cells from environmental samples. The images collected by these instruments are used to identify plankton species and to train machine learning models for automated classification (Sosik & Olson, 2007), but full taxonomic analysis of images can be time-consuming and is generally performed post-cruise. However, these images can also be quickly analysed to decompose them into a collection of metrics that describe the shape and size of the cells in addition to their fluorescence and scatter. Applying change-point detection methods such as those described in this paper to metrics easily derived from images can be used to quickly identify shifts in phytoplankton communities during cruises, allowing researchers to target specific features for adaptive sampling while cruises are in progress.

The approach we introduced in this paper is both fast and scalable, taking around 20 min to identify change points based on over 100 million particle measurements from across 12 cruises. Going forward, we intend to apply the method to near-real-time data generated during research cruises to identify potential regions for more intensive sampling.

## ACKNOWLEDGEMENTS

We thank Chris Berthiaume and Dr Annette Hynes for their help in processing and curating SeaFlow data. We also thank the anonymous reviewers and the Associate Editor for their valuable comments and constructive suggestions, which greatly improved the manuscript. This work was supported by grants from the Simons Foundation (Award ID 574495 to F.R., Award IDs 329108, 426570SP and 549894 to E.V.A.). C.J. acknowledges support from NSF grant DMS-1810975. Most of this work was done while C.J. was a PhD student in the Department of Statistics at the University of Washington. S.C. acknowledges support from a Moore/Sloan Data Science and Washington Research Foundation Innovation in Data Science Postdoctoral Fellowship. Z.H. acknowledges support from the CIFAR LMB program and NSF grant DMS-1810975. The authors are grateful to the eScience Institute at the University of Washington for supporting this collaboration. The authors have no conflicts of interest to declare.

## AUTHORS' CONTRIBUTIONS

C.J., Z.H. and S.C. conceived the ideas and designed the methodology; F.R. and E.V.A. collected and curated the data; C.J. analysed the data with the help of F.R. and S.C.; C.J. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

## DATA AVAILABILITY STATEMENT

The physical and biological data may be found at https://doi.org/10.5281/zenodo.4289399 (Ribalet et al., 2021). The kernel change-point estimation code is on Github at https://github.com/cjones6/chapydette, while the code to reproduce the results in the paper may be found at https://github.com/cjones6/cytosegmenter. Both of these repositories have been archived on Zenodo at https://doi.org/10.5281/zenodo.4746627 (Jones & Harchaoui, 2021) and http://doi.org/10.5281/zenodo.4746597 (Jones et al., 2021), respectively.

## ORCID

*Corinne Jones* https://orcid.org/0000-0001-7363-0431
*Sophie Clayton* https://orcid.org/0000-0001-7473-4873
*François Ribalet* https://orcid.org/0000-0002-7431-0234
*E. Virginia Armbrust* https://orcid.org/0000-0001-7865-5101
*Zaid Harchaoui* https://orcid.org/0000-0003-1186-1343

## REFERENCES

Alvarez-Fernandez, S., Lindeboom, H., & Meesters, E. (2012). Temporal changes in plankton of the North Sea: Community shifts and environmental drivers. *Marine Ecology Progress Series*, *462*, 21–38. https://doi.org/10.3354/meps09817

Anderson, N. H., Hall, P., & Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, *500*(1), 41–54. https://doi.org/10.1006/jmva.1994.1033

Bianchi, D., & Mislan, K. A. S. (2016). Global patterns of diel vertical migration times and velocities from acoustic data. *Limnology and Oceanography*, *610*(1), 353–364. https://doi.org/10.1002/lno.10219

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, *180*(1), 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x

d'Ovidio, F., De Monte, S., Alvain, S., Dandonneau, Y., & Lévy, M. (2010). Fluid dynamical niches of phytoplankton types. *Proceedings of the National Academy of Sciences of the United States of America*, *1070*(43), 18366–18370. https://doi.org/10.1073/pnas.1004620107

Dubelaar, G. B., Gerritzen, P. L., Beeker, A. E., Jonker, R. R., & Tangen, K. (1999). Design and first results of CytoBuoy: A wireless flow cytometer for in situ analysis of marine and fresh waters. *Cytometry*, *370*(4), 247–254. https://doi.org/10.1002/(SICI)1097-0320(19991201)37:4<247:AID-CYTO1>3.0.CO;2-9

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, *160*(2), 203–213. https://doi.org/10.1007/s11222-006-8450-8

Friedland, K. D., Record, N. R., Asch, R. G., Kristiansen, T., Saba, V. S., Drinkwater, K. F., Henson, S., Leaf, R. T., Morse, R. E., Johns, D. G., Large, S. I., Hjøllo, S. S., Nye, J. A., Alexander, M. A., & Ji, R. (2016). Seasonal phytoplankton blooms in the North Atlantic linked to the overwintering strategies of copepods. *Elementa: Science of the Anthropocene*, *4*. https://doi.org/10.12952/journal.elementa.000099

Gal, G., & Anderson, W. (2010). A novel approach to detecting a regime shift in a lake ecosystem. *Methods in Ecology and Evolution*, *10*(1), 45–52. https://doi.org/10.1111/j.2041-210X.2009.00006.x

Goldfeld, S. M., & Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, *1*, 3–16. https://doi.org/10.1016/0304-4076(73)90002-X

Gordon, A. D., & Birks, H. J. B. (1972). Numerical methods in quaternary palaeoecology I. Zonation of pollen diagrams. *New Phytologist*, *710*(5), 961–979. https://doi.org/10.1111/j.1469-8137.1972.tb01976.x

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *130*, 723–773.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, *450*(1–2), 39–70. https://doi.org/10.1016/0304-4076(90)90093-9

Harchaoui, Z., & Cappé, O. (2007). Retrospective multiple change-point estimation with kernels. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *IEEE workshop on statistical signal processing* (pp. 768–772). Curran Associates, Inc. https://doi.org/10.1109/SSP.2007.4301363

Harchaoui, Z., & Lévy-Leduc, C. (2007). Catching change-points with lasso. *Advances in neural information processing systems* (pp. 617–624).

Hare, S. R., & Mantua, N. J. (2000). Empirical evidence for North Pacific regime shifts in 1977 and 1989. *Progress in Oceanography*, *470*(2–4), 103–145. https://doi.org/10.1016/S0079-6611(00)00033-1

Hutchinson, G. E. (1957). Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, *22*, 415–427. https://doi.org/10.1101/SQB.1957.022.01.039

Hyrkas, J., Clayton, S., Ribalet, F., Halperin, D., Virginia Armbrust, E., & Howe, B. (2015). Scalable clustering algorithms for continuous environmental flow cytometry. *Bioinformatics*, *320*(3), 417–423. https://doi.org/10.1093/bioinformatics/btv594

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big*. https://doi.org/10.1109/TBDATA.2019.2921572

Jones, C., Clayton, S., Ribalet, F., Armbrust, E. V., & Harchaoui, Z. (2021). cjones6/cytosegmenter: First Release. https://doi.org/10.5281/zenodo.4746597

Jones, C., & Harchaoui, Z. (2021). cjones6/chapydette: First release. https://doi.org/10.5281/zenodo.4746627

Kay, S. M. (1993). *Fundamentals of statistical signal processing: Detection theory*. Prentice Hall PTR.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, *850*(4), 717–736. https://doi.org/10.1016/j.sigpro.2004.11.012

Legendre, P., Dallot, S., & Legendre, L. (1985). Succession of species within a community: Chronological clustering, with applications to marine and freshwater zooplankton. *The American Naturalist*, *1250*(2), 257–288. https://doi.org/10.1086/284340

Li, S., Xie, Y., Dai, H., & Song, L. (2019). Scan *B*-statistic for kernel change-point detection. *Sequential Analysis*, *380*(4), 503–544. https://doi.org/10.1080/07474946.2019.1686886

Lumpkin, R., & Johnson, G. C. (2013). Global ocean surface velocities from drifters: Mean, variance, El Niño-Southern Oscillation response, and seasonal cycle. *Journal of Geophysical Research: Oceans*, *1180*(6), 2992–3006. https://doi.org/10.1002/jgrc.20210

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., & Francis, R. C. (1997). A Pacific interdecadal climate oscillation with impacts on salmon production. *Bulletin of the American Meteorological Society*, *780*(6), 1069–1080. https://doi.org/10.1175/1520-0477(1997)078<1069:APICOW>2.0.CO;2

Matteson, D. S., & James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, *1090*(505), 334–345. https://doi.org/10.1080/01621459.2013.849605

Nieuwhof, S., Van Belzen, J., Oteman, B., Van De Koppel, J., Herman, P. M., & Van Der Wal, D. (2018). Shellfish reefs increase water storage capacity on intertidal flats over extensive spatial scales. *Ecosystems*, *210*(2), 360–372. https://doi.org/10.1007/s10021-017-0153-9

Olson, R. J., Shalapyonok, A., & Sosik, H. M. (2003). An automated submersible flow cytometer for analyzing pico- and nanophytoplankton: FlowCytobot. *Deep Sea Research Part I: Oceanographic Research Papers*, *500*(2), 301–315. https://doi.org/10.1016/S0967-0637(03)00003-7

Pace, M. L., Batt, R. D., Buelo, C. D., Carpenter, S. R., Cole, J. J., Kurtzweil, J. T., & Wilkinson, G. M. (2017). Reversal of a cyanobacterial bloom in response to early warnings. *Proceedings of the National Academy of Sciences of the United States of America*, *1140*(2), 352–357. https://doi.org/10.1073/pnas.1612424114

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *410*(1–2), 100–115. https://doi.org/10.2307/2333009

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ribalet, F., Berthiaume, C., Hynes, A., & Jones, C. (2021). SeaFlow data: Optical properties of single celled phytoplankton + auxillary physical data. https://doi.org/10.5281/zenodo.4682238

Ribalet, F., Berthiaume, C., Hynes, A., Swalwell, J., Carlson, M., Clayton, S., Hennon, G., Poirier, C., Shimabukuro, E., White, A., & Armbrust, E. V. (2019). Seaflow data v1, high-resolution abundance, size and biomass of small phytoplankton in the North Pacific. *Scientific Data*, *60*(1), 277. https://doi.org/10.1038/s41597-019-0292-2

Ribalet, F., Marchetti, A., Hubbard, K. A., Brown, K., Durkin, C. A., Morales, R., Armbrust, E. V. (2010). Unveiling a phytoplankton hotspot at a narrow boundary between coastal and offshore waters. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(38), 16571–16576. https://doi.org/10.1073/pnas.1005638107

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809682

Sosik, H. M., & Olson, R. J. (2007). Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods*, *50*(6), 204–216. https://doi.org/10.4319/lom.2007.5.204

Sosik, H. M., Olson, R. J., & Armbrust, E. V. (2010). Flow Cytometry in Phytoplankton Research. *Chlorophyll a fluorescence in aquatic sciences: Methods and applications* (pp. 171–185). Springer Netherlands. https://doi.org/10.1007/978-90-481-9268-7

Suard, F., Guigue, V., Rakotomamonjy, A., & Benshrair, A. (2005). Pedestrian detection using stereo-vision and graph kernels. *Proceedings of the IEEE Intelligent Vehicles Symposium* (pp. 267–272). https://doi.org/10.1109/IVS.2005.1505113

Swalwell, J. E., Ribalet, F., & Armbrust, E. V. (2011). Seaflow: A novel underway flow-cytometer for continuous observations of phytoplankton in the ocean. *Limnology and Oceanography: Methods*, *90*(10), 466–477. https://doi.org/10.4319/lom.2011.9.466

Thomson, J. R., Kimmerer, W. J., Brown, L. R., Newman, K. B., Nally, R. M., Bennett, W. A., Feyrer, F., & Fleishman, E. (2010). Bayesian change point analysis of abundance trends for pelagic fishes in the upper San Francisco Estuary. *Ecological Applications*, *200*(5), 1431–1448. https://doi.org/10.1890/09-0998.1

Weijerman, M., Lindeboom, H., & Zuur, A. F. (2005). Regime shifts in marine ecosystems of the North Sea and Wadden Sea. *Marine Ecology Progress Series*, *298*, 21–39. https://doi.org/10.3354/meps298021

Williams, C. K., & Seeger, M. (2000). Using the Nyström method to speed up kernel machines. In *Advances in neural information processing systems* (pp. 661–667). MIT Press.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Jones, C., Clayton, S., Ribalet, F., Armbrust, E. V., & Harchaoui, Z. (2021). A kernel-based change detection method to map shifts in phytoplankton communities measured by flow cytometry. *Methods in Ecology and Evolution*, 12, 1687–1698. https://doi.org/10.1111/2041-210X.13647