El Niño detection via unsupervised clustering of Argo temperature profiles

Isabel A. Houghton [1], James D. Wilson[1,2]

[1]The Data Institute, University of San Francisco, San Francisco, CA 94105
[2]Department of Mathematics and Statistics, University of San Francisco, San Francisco, CA 94105

Key Points:

- Unsupervised clustering based solely on temperature profiles effectively partitions water masses in the Pacific Ocean.
- The temporal evolution of the clusters reveals spatial oscillations associated with El Niño events.
- Unsupervised machine learning serves as a flexible and robust approach to anomaly detection in oceanographic data.

Corresponding author: Isabel A. Houghton, `izhoughton@gmail.com`

Abstract

Variability in the El Niño-Southern Oscillation (ENSO) has global impacts on seasonal temperatures and rainfall. Current detection methods for extreme phases, which occur with irregular periodicity, rely upon sea surface temperature anomalies within a strictly defined geographic region of the Pacific Ocean. However, under changing climate conditions and ocean warming, these historically motivated indicators may not be reliable into the future. In this work, we demonstrate the power of data clustering as a robust, automatic way to detect anomalies in climate patterns. Ocean temperature profiles from Argo floats are partitioned into similar groups utilizing unsupervised machine learning methods. The automatically identified groups of measurements represent spatially coherent, large-scale water masses in the Pacific, despite no inclusion of geospatial information in the clustering task. Further, spatiotemporal dynamics of the clusters are strongly indicative of El Niño events, the east Pacific warming phase of ENSO. The fitting of a cluster model on a collection of ocean profiles identifies changes in the vertical structure of the temperature profiles through reassignment to a different group, concisely capturing physical changes to the water column during an El Niño event, such as thermocline tilting. Clustering proves to be an effective tool for analysis of the irregularly sampled (in space and time) data from Argo floats and may serve as a novel approach for detecting anomalies given the freedom from thresholding decisions. Unsupervised machine learning could be particularly valuable due to its ability to identify patterns in datasets without user-imposed expectations, facilitating further discovery of anomaly indicators.

Plain Language Summary

The climate phenomenon known as El Niño leads to variable temperatures and rainfall amounts around the world and occurs at unpredictable intervals. The most commonly used measurement to determine an El Niño is occurring relies on the difference between the three-month average temperature and the thirty-year average at the surface of the ocean in a rectangular region near the equator. However, as climate changes, these historically defined ways of measuring an El Niño may no longer be helpful. In order to develop a more flexible way to observe an El Niño, we use tools from the field of machine learning. Specifically, temperature measurements in the Pacific Ocean from the surface down to a depth of 1,000 m are grouped automatically (i.e. without pre-defined rules) using machine learning methods. Without using information about the location of the

measurements, this process groups measurements that are also close together in space. Changes over time of group assignments closely matches an El Niño happening, and also point to physical changes to that region in the ocean. The automatic grouping of ocean profiles works very well to signal an El Niño and could potentially be a useful tool for future study of data from the ocean.

1 Introduction

The oceans are critical in governing global climate through heat transport and absorption of carbon from the atmosphere (Marshall & Plumb, 2008). Extensive effort is put toward monitoring and predicting the state of the ocean, providing valuable data for daily weather prediction as well as long term understanding of climate variability. The Pacific Ocean, the world's largest ocean basin, has recurring patterns of variability, most notably as part of the El Niño-Southern Oscillation (ENSO). Due to complex coupling between the ocean and atmosphere, sea surface temperatures and atmospheric winds in the Pacific region interact in a positive feedback loop to produce major oscillations in climate with repercussions at a global scale. An El Niño event, characterized by anomalous warming of eastern equatorial Pacific waters, occurs approximately every 3-8 years and, due to global teleconnections, results in varying temperatures and precipitation levels around the globe (Wyrtki, 1975; Rasmusson & Carpenter, 1982). The ensuing shift in seasonal temperatures and rainfall leads to droughts and flooding in Africa, Latin America, North America, and Southeast Asia. These extreme events have major consequences for human health and economic costs in the billions (Buizer et al., 2000; Iizumi et al., 2014). Despite the importance of forecasting such events, El Niño prediction remains challenging, particularly beyond a six-month horizon, due to the high non-linearity of the system and the relatively unique development of each El Niño event (Timmermann et al., 2018; Dijkstra et al., 2019).

The dynamics of the El Niño-Southern Oscillation are associated with a high pressure system over the eastern Pacific Ocean and a low pressure system over the western Pacific and Indonesia. This pressure gradient across the Pacific leads to persistent easterly winds near the equator that drive upwelling along the eastern Pacific coasts, leading to cooler surface temperatures and a tilted thermocline. During an El Niño event, the pressure gradient driven atmospheric circulation decreases, reducing upwelling along the eastern Pacific, enhancing sea surface temperatures and deepening of the thermocline in that region (Wang et al., 2000; Meinen & McPhaden, 2000).

Current El Niño detection relies on sea surface temperature anomalies within a specifically designated region (Niño 3.4, defined from 5°S to 5°N and 170°W to 120°W) in the equatorial Pacific. Extensive study of historical patterns have identified this region as the dominant location of the coupled ocean-atmosphere interactions (Bamston et al., 1997). The exclusive consideration of surface measurements in a small geographic location potentially disregards indicators in other regions of the Pacific Ocean basin and in subsurface variation of the vertical structure. As a consequence of a significant warming trend in the Niño 3.4 region since 1950, the thirty-year period against which temperature values are compared is updated on a five-year basis (NOAA National Centers for Environmental Prediction, 2020). In the context of global climate change and ocean warming, these updates will likely continue to be necessary (Ashok et al., 2007; Yeh et al., 2009). Therefore, methods for El Niño detection incorporating large horizontal and vertical scales and utilizing in situ data without empirical thresholds are of particular value (Yang & Wang, 2009).

In situ measurements of the ocean are valuable sources for subsurface observations as well as for model validation and improvement, particularly in a changing climate. In situ instruments have begun collecting increasing amounts of data, thus methods for effective analysis are critical for data utilization and could provide new approaches to ocean observation and prediction. To date, the Argo program (Riser et al., 2016) has massively increased the extent of in situ measurements of the ocean with profiling floats. Those direct measurements have provided insight into ENSO dynamics, specifically allowing detection of changes in the distribution of ocean heat content due to tilting of the thermocline (Roemmich & Gilson, 2011; Johnson & Birnbaum, 2017) and adjustment of the Equatorial Pacific Thermostad, a layer of low vertical stratification below the pycnocline (Johnson & Birnbaum, 2016). In situ measurements do come with additional challenges over uniform model data, particularly in terms of spatial and temporal sparsity (relative to model grid cells) and nonuniform sampling. As a result, development of novel methods for data utilization may prove particularly useful with the growing deluge of data becoming available.

Unsupervised machine learning methods for clustering data provide an effective and robust approach for partitioning complex data, particularly adaptable to the spatial and temporal irregularity of many in situ ocean observations. Additionally, clustering can reveal patterns or similarities in a dataset while avoiding biased expectations of what patterns should exist (e.g., thresholds derived from prior assumptions of the system). Previous work developed a profile classification model using clustering and considered unsupervised clustering of temperature profile measurements in the Atlantic and Southern Oceans (Maze, Mercier, & Cabanes, 2017; Maze, Mercier, Fablet, et al., 2017; Jones et al., 2019) and found groupings consistent with known oceanic water masses. In this work, we analyze measurements in the Pacific Ocean basin and consider the temporal evolution of the clustered data. The openly-available dataset of ocean temperature profiles from the Argo program is analyzed with unsupervised machine learning methods to reveal novel El Niño indicators free from user-imposed decisions. We find that temporal dynamics in the spatial location of cluster assignments are strongly correlated with the current leading metric for El Niño occurrence. The unsupervised methods successfully partition the temperature profiles into physically meaningful groups and the variation over time identifies changes in both thermocline depth and sea surface temperatures, key physics associated with ENSO. The data and analysis methods are described in the following section. Section 3 describes the patterns identified by the clustering algorithm and section 4 discusses their relationship to current oceanographic understanding. Finally, section 5 summarizes the utility of unsupervised methods for analyzing oceanographic data as illustrated by effective ENSO detection and highlights future directions.

2 Data and Methods

Temperature profiles in the Pacific Ocean acquired by the Argo project (Riser et al., 2016) were reduced to a lower-dimensional embedding using principal component analysis (PCA) and then grouped via k-means clustering, an unsupervised clustering method. The evolution of the spatial patterns of measurements assigned to each cluster were then considered over a thirteen-year time period (2006–2019). Measurements within each three-month period during this time period were aggregated for analysis. Oscillations in the spatial extent of clusters were compared to an indicator of El Niño. A description of the Argo temperature dataset, dimensionality reduction and clustering methods, and methods used to evaluate the spatial evolution of our clusters relative to an existing El Niño-Southern Oscillation indicator over time are included below.

2.1 Argo Float Dataset

The Argo program was initiated in the early 2000's and consists of a global array of free-drifting profiling floats that have served to massively expand our global ocean observing network. Each profiling float in the array measures the vertical structure of temperature and salinity in the ocean, with newer profiling floats also measuring bio-optical traits and biogeochemical properties. Currently, nearly 4,000 individual profiling floats are deployed, each acquiring vertical profile measurements to a depth of approximately 2,000 m every ten days. Collected data is then made publicly available in near real-time. The free-floating nature of the instruments leads to a global array of sensors distributed at roughly every three degrees ($\sim$300 km), with dynamically changing positions over time. Argo is the leading source of global subsurface data, particularly for use in ocean data assimilation and model reanalysis (Riser et al., 2016).

Argo profiling float measurements of temperature were acquired in October 2019 in the Pacific Ocean basin between 30°S and 50°N from January 2006 to September 2019 (Argo, 2019). The start date was chosen to achieve a sufficiently long window to observe several El Niño events while utilizing a similar number of measurements at a given time (see supplementary Figure S1). Prior to 2006, the number of profile measurements available becomes sparser. The latitude range was chosen to initially focus on the low- to mid-latitudes of the Pacific basin Each measurement had an associated latitude, longitude, and acquisition timestamp. Only delayed-mode data were used. All temperature profiles containing missing data, insufficient data points, or nonphysical values were removed as defined below. This corresponded to profiles with fewer than 50 data points, the initial data point more than 25 dbar from the surface, the final data point less than 1,000 dbar, or temperature values less than -5°C. Temperature values in the remaining profiles were linearly interpolated onto a uniform grid with 5 dbar spacing from 5 dbar down to 1,000 dbar, the same as Jones et al. (Jones et al., 2019). Data was only stored down to 1,000 dbar despite measurements extending down to approximately 2,000 dbar due to the majority of temperature variability of interest associated with the thermocline and occurring in the upper 1,000 dbar (Yang & Wang, 2009). This yielded a set of approximately 560,000 temperature profiles consisting of 200 data points each for the thirteen year time span that were subsequently assigned to clusters. The count of profiles increased from approximately 35,000 per year in 2006 to approximately 70,000 in 2019 (see supplementary Figure S1).

2.2 Dimensionality Reduction and Clustering

A critical first step toward effective clustering for a high-dimensional variable is dimensionality reduction (Beyer et al., 1998). Effective dimensionality reduction casts a given sample with many features into a lower-dimensional space where a distance metric between two samples reasonably captures differences within the dataset. For the temperature profiles consisting of hundreds of data points over a uniform depth grid, calculating a point-wise difference between each profile would not fully capture critical differences between profiles, such as the shape of the temperature profile with depth (e.g. thermocline location).

In this work, principal component analysis (PCA) was applied utilizing the *scikit-learn* machine learning library for Python (Pedregosa et al., 2011). This algorithm implements linear dimensionality reduction using singular value decomposition of the data to project each sample into a lower dimensional space of linearly uncorrelated (orthogonal) basis functions, termed principal components (Shlens, 2003). The first principal component accounts for the largest possible variance in the data, and each subsequent component attempts to further maximally account for variance under the constraint of orthogonality to preceding components. Thus, one can specify the desired variance to account for in the data and the number of components to describe that variance between samples will be retained. PCA was applied to the 200-data-point profiles to capture 99.9% of the variance with 17 principal components. Across all of the profiles, each depth level was mean-centered but not scaled, therefore the lower depth levels with less temperature variability contributed less to the final representation. Seventeen components was notably higher than previous work by Jones et al. (2019), which only required six components for 99.9% of the variance, likely due to the strong vertical coherence of the Southern Ocean (Karsten & Marshall, 2002) as well as Maze et al. (2017b), which prescribed retaining 11 principal components for 99.88% of the variance. Beyond the first few, the principal components had very small magnitudes which is perhaps why an additional six principal components were necessary to account for 99.9% of the data variance here compared to the 99.88% used in Maze et al. (2017b).

With dimensionality reduction applied, properties such as Euclidean distance between each new feature (i.e., principal component) become notably more effective at describing sample differences (Beyer et al., 1998). Clustering methods were next applied with the goal of grouping the profiles solely based on differences in temperature structure without any geospatial information or external constraints applied. A wide variety of clustering methods exist with different advantages and levels of complexity (Xu & Tian, 2015). While exploration of the different clustering outcomes from the variety of methods (e.g. spectral clustering, hierarchical models) would potentially reveal interesting insights, the primary goal of this study was to find a straightforward approach to assign temperature profiles to groups. Previous work utilized Gaussian mixture modeling (GMM), which aims to fit the data as a linear combination of multidimensional Gaussian distributions. In this work, k-means clustering, a widely utilized and efficient approach in a variety of applications (Jain, 2010), was chosen, primarily due to its computational efficiency and straightforward implementation. Results from k-means were compared with GMM (see supplement).

Given a set of samples $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$, where each sample is represented by a d-dimensional vector, the k-means clustering algorithm aims to partition the n samples into k clusters, $\boldsymbol{C}$=$\{C_1, C_2, ..., C_k\}$, with the objective of minimizing the within-cluster sum of squares (WCSS). In particular, let μ_ℓ be the mean of the data within the ℓth cluster, C_ℓ. The k-means algorithm seeks to identify the partition, $\boldsymbol{C}$, that minimizes

$$WCSS = \sum_{\ell=1}^{k} \sum_{\mathbf{x} \in C_\ell} \|\mathbf{x} - \boldsymbol{\mu}_\ell\|^2 . \tag{1}$$

The embeddings of the temperature profiles produced by PCA were clustered following the *scikit-learn* implementation of the k-means clustering task to assign each profile measurement to a cluster.

One limitation of k-means clustering lies in the required choice of number of clusters, k, to create. However, due to the efficiency of implementation of the algorithm, a range of cluster counts can be tested and cluster characteristics can be analyzed to assess optimal cluster count. A common strategy to assess the cohesion of clusters in a partition, i.e. how similar every object is to its cluster, is to measure the average silhouette score of the cluster assignment (Rousseeuw, 1987).

To obtain a silhouette score, for each data point $i \in C_\ell$, the mean distance between i and all other data points in the same cluster is given by:

$$a(i) = \frac{1}{|C_\ell| - 1} \sum_{j \in C_\ell, i \neq j} d(i,j) \tag{2}$$

where $d(i,j)$ is the distance between cluster points i and j in the cluster C_ℓ, and $|C_\ell|$ denotes the number of data points in cluster ℓ. The dissimilarity of point $i \in C_\ell$ to other clusters is then defined by:

$$b(i) = \min_{k \neq \ell} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \tag{3}$$

where the cluster to which sample i is closest, but not assigned, is used (indicated by the *min* operator). Combining the similarity of a sample to its assigned cluster ($a(i)$) and dissimilarity to the nearest cluster to which it is not assigned ($b(i)$), yields a silhouette score, s, defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4}$$

which can then be aggregated for all partitioned points. To assess the cohesion of a partition, $\boldsymbol{C}$, we measure the average silhouette score across all data points. An optimal silhouette score of 1.0 indicates the sample is a large distance to non-assigned clusters and small distance to other samples in the assigned cluster. A score of -1.0 indicates the sample is closer to another cluster than its own and a score of 0 indicates the object is on the border of two natural clusters. The global silhouette score can be calculated for varying cluster counts, ideally encountering a cluster count, k, that maximizes the global silhouette score. The silhouette score was taken into account with physical intuition regarding the Pacific Ocean in order to find an optimal cluster count that maximizes uniqueness of data in the clusters with sufficient clusters to describe variability in the Pacific. Specifically, inspection of the unique water masses in the Pacific Ocean (Emery, 2008) indicated likely more than three clusters (the value of k with the global maximum silhouette score, see below and Figure 1) would be useful to capture the variability given the presence of several upper water masses in the Pacific (e.g. Equatorial, Eastern and Western Central waters, Northern and Southern Eastern Transition waters) overlapping with intermediate waters (e.g. North Pacific Intermediate Water, California Intermediate Water, and Antarctic Intermediate Water).

K-means clustering was found to be effective at partitioning, reproducible, and highly computationally efficient. The silhouette score for cluster counts ranging from 3 to 10 exhibited a stable point at $k = 7$ (Figure 1), indicating partitioning at that granularity aligned with separations in the data. Seven clusters were chosen in order to balance obtaining a reasonable number of clusters with improvement seen in the silhouette score. While choice of k did involve decision making in an otherwise unsupervised process, variation of cluster count did not fundamentally alter the partitioning, but rather led to a coarsening (for fewer clusters) or refining (for more clusters) of the divisions along similar lines (see supplementary Figure S3). Following selection of an appropriate k, data across all time (2006-2019) were simultaneously clustered and the assigned cluster identity was used for subsequent analysis. Alternatively, temperature profiles could be divided into shorter time periods and then subsequently clustered (not shown). However, simultaneous clustering across all time yielded similar partitions and provided a more consistent approach, particularly given the free-floating, intermittent nature of the measurements in contrast to a fixed set of sampling locations.

Because the k-means clustering algorithm is randomly initialized and can converge on a local (rather than global) minimum, repeatability of the clustering assignment was quantified with an adjusted Rand index measuring the similarity between two different groupings, adjusted for random chance of assignment (Rand, 1971). An index of 1.0 indicates exactly identical clustering, regardless of specific label changes (i.e. a cluster labelled #1 in one partitioning can be labelled cluster #4 in a subsequent partitioning but have the same members). The adjusted Rand index was calculated before analysis was carried out to confirm that repeated clustering would yield similar results. Repetition of the clustering produced very similar results such that the same profiles were consistently grouped together. Ten repeated clusterings produced an average adjusted Rand index of 0.997, indicating high repeatability of the analysis.

2.3 El Niño-Southern Oscillation Indicator

The current leading diagnostic metric of El Niño-Southern Oscillation state utilized by the National Oceanic and Atmospheric Administration (NOAA) relies on the sea surface temperature anomaly within the rectangular Niño 3.4 region of the Pacific defined from 5°S to 5°N and 170°W to 120°W (NOAA National Centers for Environmental Prediction, 2020). The three-month running mean of the anomaly from the most recent 30-year historical period (updated every five years) in this region is termed the Oceanic Niño Index (ONI). This index must exceed ± 0.5°C for at least five consecutive overlapping three-month periods to classify the period as a full-fledged El Niño (+0.5°C) or La Niña (-0.5°C) (NOAA National Centers for Environmental Prediction, 2020). ONI values were obtained from NOAA (NOAA National Centers for Environmental Prediction, 2020) and used directly for comparison.

2.4 Spatio-temporal Cluster Analysis

Following the clustering of temperature measurements without any associated temporal or geospatial information, the locations of measurements assigned to each cluster were analyzed over time and compared to historic El Niño events, utilizing the ONI as a ground truth on the historic presence or absence of an event. All profile measurements occurring in a 90 day window were aggregated into a single timestep with the window shifting by 30 days for each subsequent timestep, providing statistics representing a three-month running mean for comparison with the ONI values. For each cluster (ℓ) and for each timestep (t), the anomaly from the average longitude of the cluster over the full 13-year period ($\lambda_{\ell,2006-2019}$) was then considered. However, there could be many measurements in a given area (because the profiling floats are not enforced to be uniformly distributed) skewing the mean toward that region. To effectively capture how far a given cluster extends east or west beyond its $\lambda_{\ell,2006-2019}$ position at a given timestep, all unique longitudes of measurements within the cluster were aggregated, essentially indicating the spatial coverage of that cluster. In practice, all measurements in the same 0.5 degree longitude bin were considered as a single unique longitude. The average of this unique set of longitudes was calculated ($\lambda_{\ell,t}$) and then differenced from the $\lambda_{\ell,2006-2019}$ position for this cluster to determine the anomaly in the average of the unique longitudes for that cluster at that timestep. This method minimized the importance of several measurements at the same longitude (but potentially different latitude) and highlighted oscillations in the zonal extent of a cluster.

3 Results

3.1 Clustering

Each group produced by the clustering algorithm contained profiles with relatively similar vertical structure and surface temperature values indicated by the uniqueness of the average temperature profile of each cluster and the standard deviation within the group relative to variation between groups (Figure 2). The unsupervised clustering method was able to detect differences and partition profiles with similar surface temperatures but unique vertical structures (e.g. clusters 0 and 5), as well as similar vertical structures but shifted temperatures (e.g. clusters 2 and 5), a complex task to achieve with hard-coded selection rules. Each measurement assigned to a cluster also had an associated latitude and longitude allowing visualization of clusters in geographic space. A map of all measurement locations, each colored by its corresponding cluster assignment, illustrates the spatial coherency of each cluster, with few outliers and minimal spatial overlap between clusters (Figure 3a). This spatial coherency was similar to previous analyses by Maze, Mercier, Fablet, et al. (2017) and Jones et al. (2019), despite utilization of a different clustering method (k-means versus Gaussian mixture model). Notably, when only sea surface temperature (i.e. the uppermost measurement by the profiling float) was used for clustering (Figure 3b), the clusters were significantly less spatially well-defined with a scattered overlap of measurements belonging to different groups, indicating the subsurface structure of the temperature profile was critical in partitioning.

3.2 Spatio-Temporal Dynamics

Assignment of measurements to clusters from three-month time periods exhibited clear spatial oscillations correlated with the Oceanic Niño Index. Oscillations were primarily observed in clusters with measurements at lower latitudes (see Figure 4 and supplementary video). Figure 4 revealed a noticeable change in clustering assignments which closely matched El Niño events.

3.2.1 Niño 3.4 Region

For direct comparison with the current region considered for diagnosis of El Niño conditions, clustering of measurements in the constrained geographic region of Niño 3.4 (N3.4) was considered first. The cluster assignments, rather than the traditional surface temperature values, were analyzed. Two groups primarily populated the N3.4 region over the thirteen years, a low latitude western group (cluster 5, teal) and a low latitude eastern group (cluster 2, orange). The two groups occupied unique spatial regions with an east-west division. Qualitatively, the division oscillated east and west irregularly, in synchrony with the ONI (inner boxed regions, Figure 4). During neutral ENSO periods, the N3.4 region was approximately evenly divided between one group in the western half and one group in the eastern half (Figure 4a,c). During a positive ONI anomaly (El Niño event), the western cluster distinctly shifted eastward to occupy the majority of the N3.4 region (Figure 4b,d). Following an event, as the ONI rapidly returned to neutral levels, the western cluster shifted back to its original balance partially occupying the N3.4 region along with eastern cluster measurements. The shifting of the spatial locations of measurements assigned to a group is quantified by the anomaly in the average unique longitudes of measurements in the eastern cluster (Figure 5a). By the anomaly of the $\lambda_{2,t}$ positions for the eastern cluster at each time step from that cluster's $\lambda_{2,2006-2019}$ position as defined above, the average unique longitudinal position of measurements in cluster 2 was consistently farther east (positive longitudinal anomaly) during periods above the El Niño threshold, and near average or farther west during other periods.

3.2.2 Tropical Pacific Region

Temporal dynamics of cluster assignments in the entire tropical Pacific region spanning $\pm$ 23.4° latitude indicated additional larger-scale patterns. The tropics were primarily populated by three groups: one group (cluster 2, orange) in the eastern Pacific spanning the tropical latitudes, a second group in the western Pacific confined to lower latitudes (cluster 5, teal), and a third group (cluster 0, maroon) also in the western Pacific to the north and south of the second group (Figure 3a). During an elevated ONI period, the eastern cluster that had shifted farther east at very low latitudes (N3.4 region), simultaneously significantly expanded its extent westward at slightly northern latitudes, leading to the presence of measurements assigned to this cluster all the way across the Pacific in a narrow band around 10°N (Figure 4b,d). This phenomenon exhibited itself during every El Niño event during the time period assessed (2006-2019). This oscillation was quantified with the anomalous average unique longitude ($\lambda_{2,t}$-$\lambda_{2,2006-2019}$) of the eastern cluster (Figure 5b), and had a Pearson correlation coefficient with the ONI of -0.75 and a peak cross-correlation with zero time lag.

4 Discussion

The ocean is composed of a distribution of water masses with unique temperature and salinity characteristics that can be related to the region of water mass formation (Emery, 2008). These water masses typically have both a horizontal and vertical extent. Therefore, a profile measurement down to 1,000 dbar would likely sample multiple water masses, indicated by temperature and salinity variability over depth in the profile. This layering of unique water masses with variable horizontal extents results in the high variability seen in temperature profiles. However, temperature profiles obtained physically proximate are likely sampling the same set of water masses and therefore likely to exhibit similar structure. The effective clustering of similarly structured temperature profiles in turn led measurements within a given cluster to be spatially proximate, as seen in Figure 3. Further, the clusters identified align well with traditional water masses and the pattern of overlap of different depth water masses. Specifically, cluster 3 (yellow) aligns well with the boundaries of Pacific Subarctic Water (PSUW) (Figure 3). Cluster 4 (light green) appears to represent the regions occupied by the North and South Pacific Subtropical Mode Waters (NPSMW, SPSMW). The path of the Kuroshio Current aligns well with the boundary between clusters 4 (light green) and 1 (dark orange) while the path of the Oyashio Current appears to align well with the boundary between clusters 3 (yellow) and 1 (dark orange). The boundaries on the meridional extent of the Pacific Equatorial Water relative to the Western North Pacific Central Water (WNPCW) and the Western South Pacific Central Water (WSPCW) in the western equatorial Pacific can also be seen as the boundary between the clusters 5 (teal) and 0 (red) (Figure 3). Intermediate depth waters are also formed off the coast of California in the northern hemisphere and off the coast of South America in the southern hemisphere, termed California Intermediate Water (CIW) and East Southern Pacific Intermediate Water (ESPIW) (Emery, 2008), respectively, as a consequence of coastal upwelling. Both of these water masses appear clustered together in cluster 1 (dark orange) near the sites of coastal upwelling. Meanwhile, cluster 2 (light orange) appears to represent water masses related to equatorial dynamics and Eastern South Pacific Central Water/Eastern North Pacific Central Water. The Pacific Equatorial Water (PEW) forms a notable band of water at low latitudes (Emery, 2008). This region was also partitioned by the clustering task, and was divided into an eastern and western cluster at low latitudes. This east-west division of the PEW was due to the variable thermocline depth and surface temperature across the Pacific, with sufficiently high variability relative to the rest of the profiles for the algorithm to identify two unique clusters (Figure 2). Interestingly, even very few partitions (e.g. $k = 3$) still divided the PEW, indicating the east-west variability was relatively dominant (see supplement, Figure S3). Despite thermocline depth and surface temperature varying continuously across the Pacific, the partitioning divided the PEW at a consistent threshold, with the location of that division found to be particularly relevant in terms of temporal variability.

The switching of cluster assignment in the regions of interest signaled a physical change to the water column indicated by the differences in temperature profiles in the two dominant oscillating clusters (i.e. orange and teal profiles in Figure 2). The eastern cluster is characterized by cooler surface temperatures and a shallower thermocline, therefore a shift of that cluster out of the N3.4 region aligns with the positive ONI temperature anomaly. All anomalies in the average of the unique longitudes within the eastern cluster ($\lambda_{2,t}$-$\lambda_{2,2006-2019}$) beyond one standard deviation occur simultaneously with an El Niño event, and only the major event in 2015-2016 exceeds two standard deviations (Figure 5). At the surface, the profiles in the western cluster (5) have warmer temperatures than profiles in the eastern cluster (2). In terms of vertical structure, the thermocline is deeper in the western cluster and shallower in the eastern cluster. Thus, during neutral conditions, the east-west division in the two clusters corresponds to a tilted thermocline and colder surface temperatures in the east. During an El Niño, the western cluster extends farther eastward at the equator, indicating warmer surface temperatures and a deeper thermocline than under neutral conditions, consistent with physical understanding of ENSO dynamics (Meinen & McPhaden, 2000). Additionally, the eastern cluster extends far westward in a band north of the western cluster, leading to a north-south gradient in cluster identity and accompanying north-south surface temperature gradient and thermocline tilt that is unique to periods with an elevated Oceanic Niño Index. The spatial extent of the clusters thus provided a concise method for observation of oscillations characteristic of Kelvin and Rossby wave-driven ENSO dynamics (Battisti, 1989; Kim & Kim, 2002). The ability to compare the general characteristics of profiles in each group produced by the clustering provided a concise way to identify complex shifts in water column structure over time and clearly identify anomalous periods.

Unsupervised clustering provided a robust way to delineate regions with distinct water masses without imposing thresholds or arbitrary latitude or longitude limits. Additionally, the spatial locations of measurements within a cluster evolved over time, and relating back to the original temperature profiles in a given cluster indicated the physical dynamics at work, such as a shift in thermocline depth.

5 Conclusions

Approximately 560,000 temperature profiles in the Pacific Ocean taken from 2006-2019 were partitioned into seven groups via the k-means clustering method. Analysis of all measurement assignments illustrate spatially coherent patterns associated with known water masses of the Pacific despite no inclusion of geospatial information in the clustering decision. Cluster assignments over time oscillate in spatial extent, particularly at lower latitudes. These oscillations are strongly correlated with the Oceanic Niño Index, the broadly utilized indicator of an El Niño event. The representative profiles of each cluster correspond to the current understanding of oceanic dynamics, particularly the shift in sea surface temperature and thermocline depth as a result of reduced eastern Pacific upwelling during El Niño events. Despite the difficult task of uniformly sampling a massive extent of the worlds oceans with free-drifting devices, Argo sensors are gathering sufficient data to observe oscillations in oceanic dynamics over relatively short time periods (i.e. three months) at relatively high resolution (3-5 degrees), indicating the unparalleled value of the ever-increasing observing network and the real-time data distribution.

While clustering methods have been applied across a variety of fields, utilization within ocean and climate sciences remains limited (Karpatne et al., 2019). However, as climate change continues and potentially accelerates (IPCC, 2019), identifying robust methods to identify patterns and anomalies within climate and environmental data could prove invaluable as historic means continuously shift. In the context of climate models in a changing climate, this objective approach could further serve to account for biases in ENSO representation. Unsupervised methods such as clustering and other complex network theory approaches (e.g. anomaly detection on a graph) provide an automated approach to segmentation and analysis driven by statistics of the dataset rather than potentially imposing biases toward expected, but not necessarily fully representative, patterns.

Altogether, unsupervised machine learning techniques prove to be a highly effective approach for analyzing Argo data and gaining physical insights into the system.

Acknowledgments

These data were collected and made freely available by the International Argo Program and the national programs that contribute to it; http://www.argo.ucsd.edu, http://argo.jcommops.org. The Argo Program is part of the Global Ocean Observing System. The ONI was provided by the National Oceanic and Atmospheric Administration National Centers for Environmental Prediction; https://origin.cpc.ncep.noaa.gov/products/analysis_ monitoring/ensostuff/ONI_v5.php.

JDW was partially funded by the National Science Foundation grant NSF DMS-1830547. IAH was supported by The Data Institute at University of San Francisco.

References

Argo. (2019). Argo float data and metadata from Global Data Assembly Center (Argo GDAC) - Snapshot of Argo GDAC of October 8st 2019. *SEANOE*. Retrieved from `https://www.seanoe.org/data/00311/42182/#67548` doi: 10.17882/42182

Ashok, K., Behera, S. K., Rao, S. A., Weng, H., & Yamagata, T. (2007). El Niño Modoki and its possible teleconnection. *Journal of Geophysical Research: Oceans*. doi: 10.1029/2006JC003798

Bamston, A. G., Chelliah, M., & Goldenberg, S. B. (1997). Documentation of a highly enso-related sst region in the equatorial pacific: Research note. *Atmosphere - Ocean*. doi: 10.1080/07055900.1997.9649597

Battisti, D. S. (1989). On the Role of Off-Equatorial Oceanic Rossby Waves during ENSO. *Journal of Physical Oceanography*. doi: 10.1175/1520-0485(1989)019⟨0551:otrooe⟩2.0.co;2

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1998). When is "nearest neighbor" meaningful? In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. doi: 10.1007/3-540-49257-7{_}15

Buizer, J. L., Foster, J., & Lund, D. (2000). Global impacts and regional actions: Preparing for the 1997-98 El Niño. *Bulletin of the American Meteorological Society*. doi: 10.1175/1520-0477(2000)081⟨2121:GIARAP⟩2.3.CO;2

Dijkstra, H. A., Petersik, P., Hernández-García, E., & López, C. (2019, 10). The Application of Machine Learning Techniques to Improve El Niño Prediction Skill.

Frontiers in Physics, *7*. Retrieved from `https://www.frontiersin.org/article/10.3389/fphy.2019.00153/full` doi: 10.3389/fphy.2019.00153

Emery, W. J. (2008). Water Types and Water Masses. In *Encyclopedia of ocean sciences: Second edition.* doi: 10.1016/B978-012374473-9.00108-9

Iizumi, T., Luo, J. J., Challinor, A. J., Sakurai, G., Yokozawa, M., Sakuma, H., . . . Yamagata, T. (2014). Impacts of El Niño Southern Oscillation on the global yields of major crops. *Nature Communications*. doi: 10.1038/ncomms4712

IPCC. (2019). IPCC Special Report on the Ocean and Cryosphere in a Changing Climate. In *Ipcc summary for policymakers.* doi: https://www.ipcc.ch/report/srocc/

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. doi: 10.1016/j.patrec.2009.09.011

Johnson, G. C., & Birnbaum, A. N. (2016). Equatorial Pacific Thermostad response to El Niño. *Journal of Geophysical Research: Oceans*. doi: 10.1002/2016JC012304

Johnson, G. C., & Birnbaum, A. N. (2017). As El Niño builds, Pacific Warm Pool expands, ocean gains more heat. *Geophysical Research Letters*. doi: 10.1002/2016GL071767

Jones, D. C., Holt, H. J., Meijers, A. J., & Shuckburgh, E. (2019). Unsupervised Clustering of Southern Ocean Argo Float Temperature Profiles. *Journal of Geophysical Research: Oceans*. doi: 10.1029/2018JC014629

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2019). Machine Learning for the Geosciences: Challenges and Opportunities. *IEEE Transactions on Knowledge and Data Engineering*. doi: 10.1109/TKDE.2018.2861006

Karsten, R. H., & Marshall, J. (2002). Constructing the residual circulation of the ACC from observations. *Journal of Physical Oceanography*. doi: 10.1175/1520-0485(2002)032⟨3315:CTRCOT⟩2.0.CO;2

Kim, K. Y., & Kim, Y. Y. (2002). Mechanism of Kelvin and Rossby waves during ENSO events. *Meteorology and Atmospheric Physics*. doi: 10.1007/s00703-002-0547-9

Marshall, J., & Plumb, R. A. (2008). *Atmosphere, Ocean, and Climate Dynamics.* doi: 10.1017/CBO9781107415324.004

Maze, G., Mercier, H., & Cabanes, C. (2017). Profile Classification Models. *Mercator Ocean Journal*.

Maze, G., Mercier, H., Fablet, R., Tandeo, P., Lopez Radcenco, M., Lenca, P., ... Le Goff, C. (2017). Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean. *Progress in Oceanography*. doi: 10.1016/j.pocean.2016.12.008

Meinen, C. S., & McPhaden, M. J. (2000). Observations of warm water volume changes in the equatorial Pacific and their relationship to El Nino and La Nina. *Journal of Climate*. doi: 10.1175/1520-0442(2000)013⟨3551:OOWWVC⟩2.0.CO;2

NOAA National Centers for Environmental Prediction. (2020). *Cold & Warm Episodes by Season.* Retrieved from `https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php`

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., ... Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python* (Vol. 12; Tech. Rep.). Retrieved from `http://scikit-learn.sourceforge.net.`

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*. doi: 10.1080/01621459.1971.10482356

Rasmusson, E. M., & Carpenter, T. H. (1982). Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation/ El Nino (Pacific) . *Monthly Weather Review*. doi: 10.1175/1520-0493(1982)110⟨0354:VITSST⟩2.0.CO;2

Riser, S. C., Freeland, H. J., Roemmich, D., Wijffels, S., Troisi, A., Belbéoch, M., ... Jayne, S. R. (2016). *Fifteen years of ocean observations with the global Argo array.* doi: 10.1038/nclimate2872

Roemmich, D., & Gilson, J. (2011). The global ocean imprint of ENSO. *Geophysical Research Letters*. doi: 10.1029/2011GL047992

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. doi: 10.1016/0377-0427(87)90125-7

Shlens, J. (2003). A tutorial on principal component analysis: derivation, discussion and singular value decomposition. *Online Note httpwww snl salk edushlenspub-*

notespca pdf. doi: 10.1.1.115.3503

Timmermann, A., An, S. I., Kug, J. S., Jin, F. F., Cai, W., Capotondi, A., … Zhang, X. (2018). *El Niño–Southern Oscillation complexity.* doi: 10.1038/s41586-018-0252-6

Wang, B., Wu, R., & Lukas, R. (2000). Annual adjustment of the thermocline in the tropical Pacific Ocean. *Journal of Climate*. doi: 10.1175/1520-0442(2000)013⟨0596:AAOTTI⟩2.0.CO;2

Wyrtki, K. (1975). El Niño—The Dynamic Response of the Equatorial Pacific Oceanto Atmospheric Forcing. *Journal of Physical Oceanography*. doi: 10.1175/1520-0485(1975)005⟨0572:entdro⟩2.0.co;2

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*. doi: 10.1007/s40745-015-0040-1

Yang, H., & Wang, F. (2009). Revisiting the thermocline depth in the equatorial Pacific. *Journal of Climate*. doi: 10.1175/2009JCLI2836.1

Yeh, S. W., Kug, J. S., Dewitte, B., Kwon, M. H., Kirtman, B. P., & Jin, F. F. (2009). El Nino in a changing climate. *Nature*. doi: 10.1038/nature08316

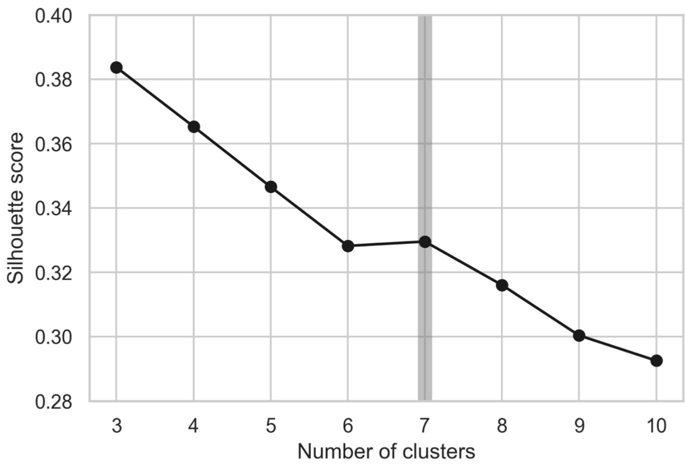

Figure 1. Silhouette score as a function of number of clusters, k, from 3 to 10 calculated following equation 4. A local maximum (highlighted in gray) is observed at $k = 7$.

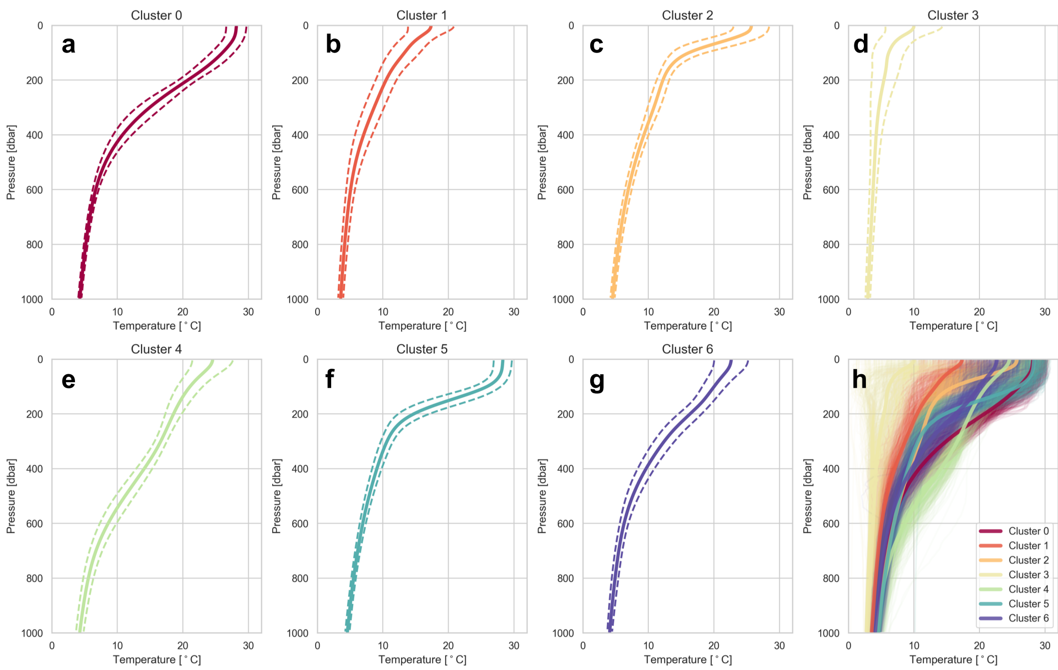

Figure 2. (a-g) For each cluster, the mean temperature profile (solid line) and ± one standard deviation of temperature (dashed line) is plotted. (h) Overlay of a random subset of profiles from each cluster, with thicker lines indicating the mean temperature profile in each cluster, colored by cluster assignment.

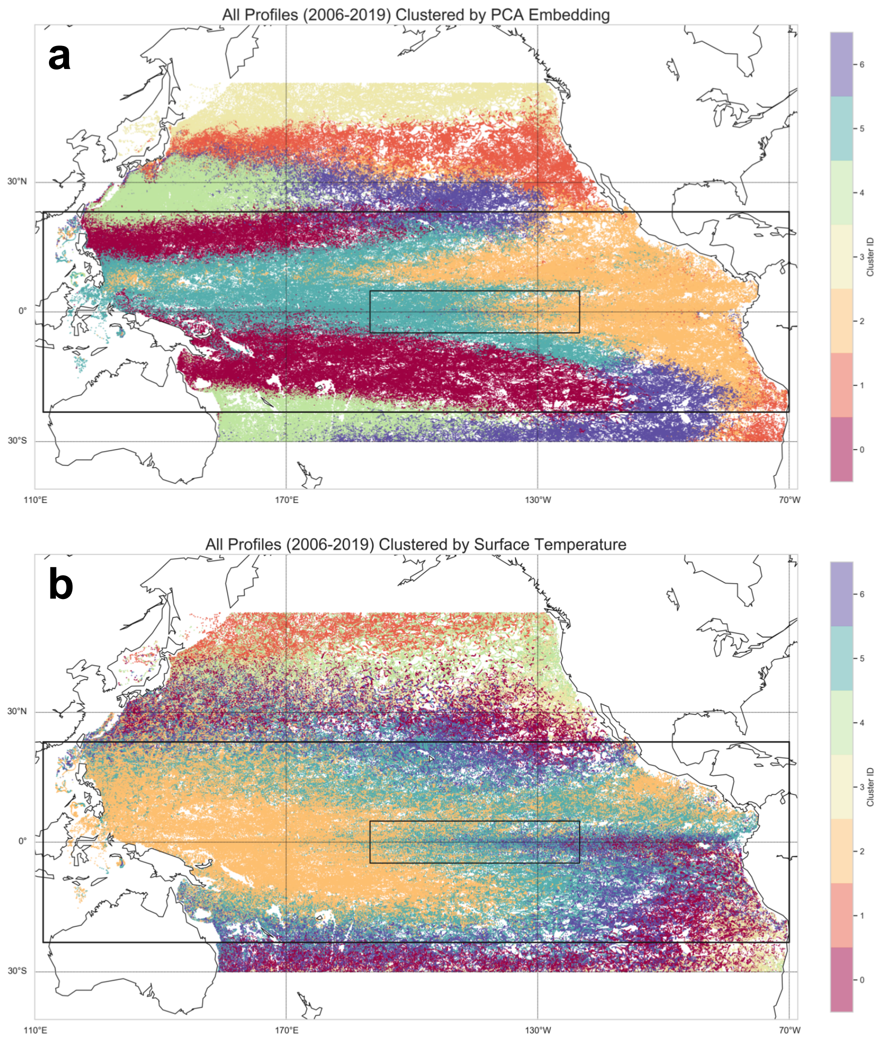

Figure 3. The spatial distribution of Argo measurements in the Pacific, colored by cluster assignment. Cluster IDs are randomly set by the clustering algorithm initialization, therefore ID magnitudes are arbitrary. The large black box corresponds to the tropical zone ($\pm$ 23.4° latitude), and the smaller inner box corresponds to the Niño 3.4 region. (a) Measurements grouped by PCA embedding of full temperature profile, used for subsequent analysis. (b) Measurements grouped by sea surface temperature (uppermost profile measurement only).

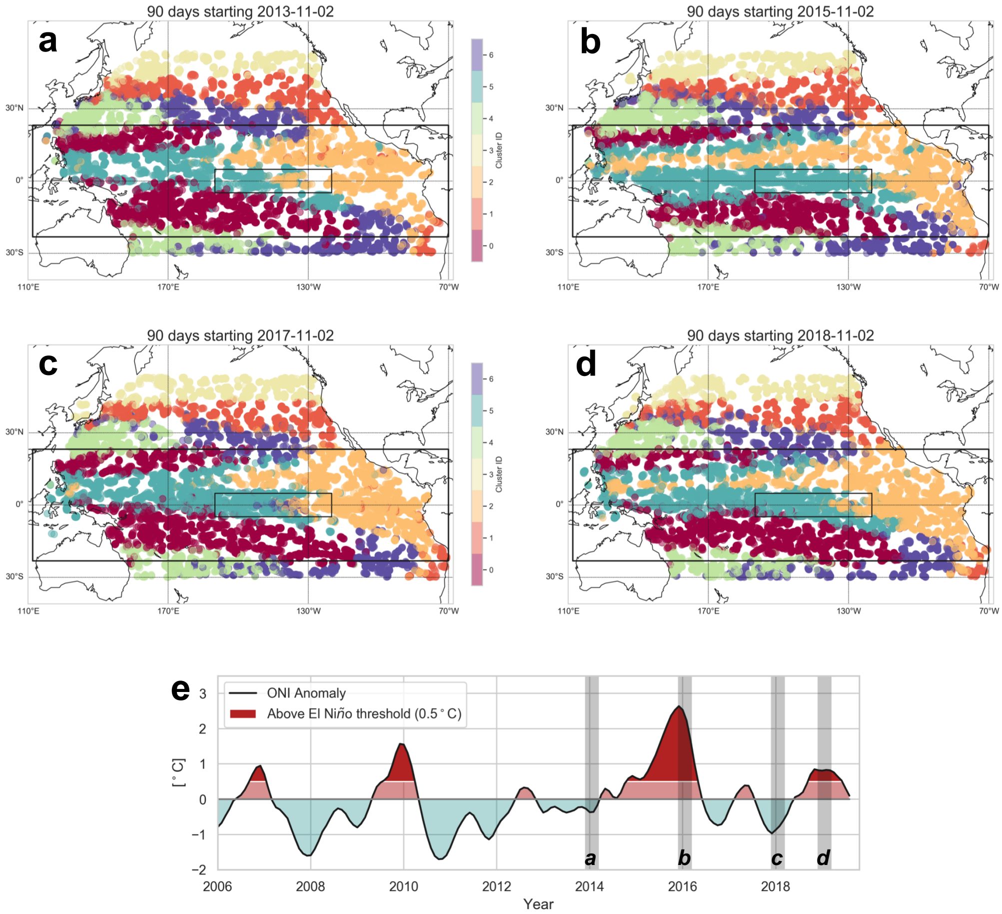

Figure 4. (a-d) Measurements within each three-month period indicated colored by cluster assignment (see supplementary video for cluster assignments over all time). (e) The ONI anomaly from 2006 to 2019 indicating several El Niño events. Vertical gray shaded bars correspond to time periods visualized in upper plots.

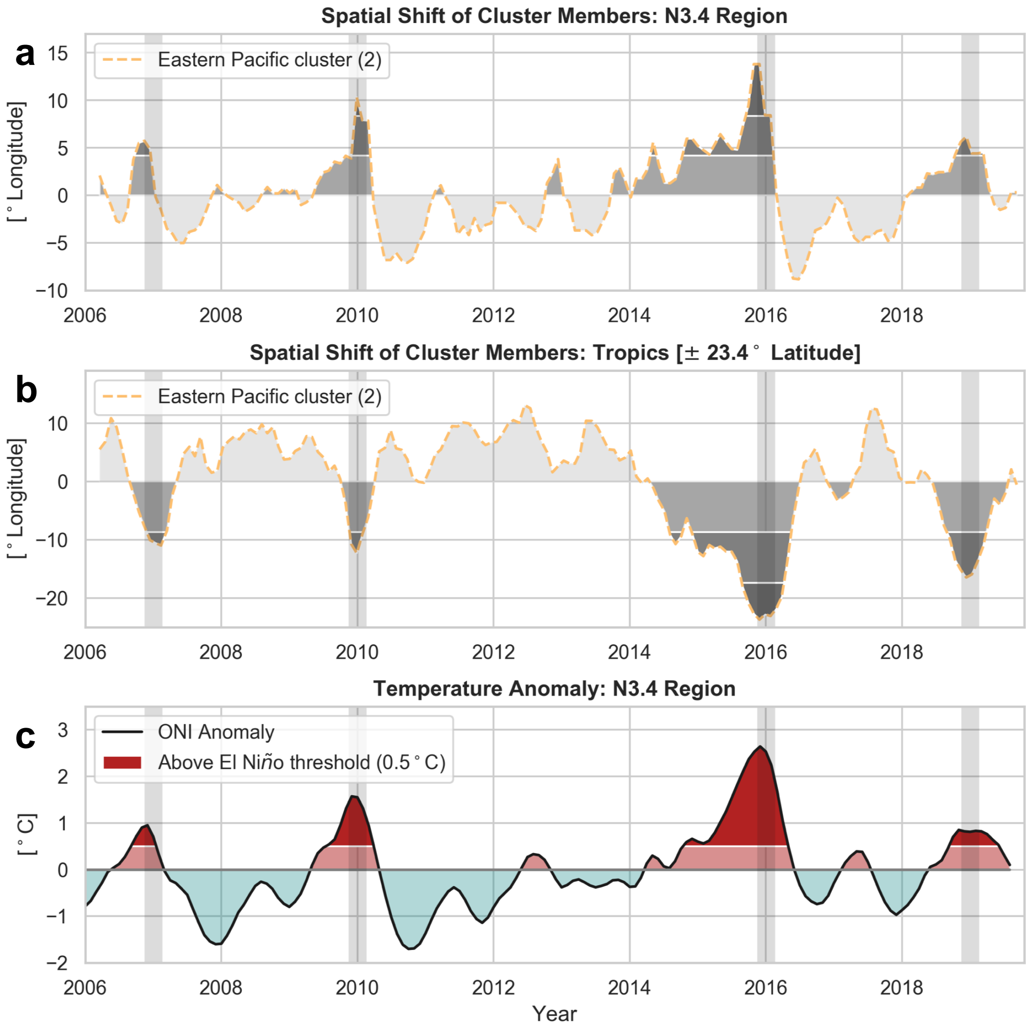

Figure 5. (a) The longitudinal anomaly of the eastern cluster members within the Niño 3.4 region. (b) The longitudinal anomaly of the Eastern cluster members over the entire tropics. White lines and gray shading correspond to standard deviations from the mean. Vertical gray bars on all plots correspond to an El Niño event occurring. (c) ONI during the same period, dark red region corresponds to events above 0.5°C threshold.