El Nino detection via unsupervised clustering of Argo temperature profiles

Isabel A Houghton[1] and James Wilson[1]

[1]University of San Francisco

November 24, 2022

Abstract

Variability in the El Nino-Southern Oscillation 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, temporal dynamics of the clusters are strongly indicative of El Nino events, the Pacific warming phase of the El Nino-Southern Oscillation. The unsupervised clustering task successfully 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 Nino event, such as tilting of the thermocline. Clustering proves to be an effective tool for analysis of the irregularly sampled (in space and time) data from ocean floats and may serve as a novel approach for detecting future anomalies given the freedom from thresholding decisions. Unsupervised machine learning approaches could be particularly valuable due to their ability to identify patterns in datasets without user-imposed expectations, facilitating further discovery of anomaly indicators.

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 Math 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, ihoughton@usfca.edu

Abstract

Variability in the El Niño-Southern Oscillation 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, temporal dynamics of the clusters are strongly indicative of El Niño events, the Pacific warming phase of the El Niño-Southern Oscillation. The unsupervised clustering task successfully 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 tilting of the thermocline. Clustering proves to be an effective tool for analysis of the irregularly sampled (in space and time) data from ocean floats and may serve as a novel approach for detecting future anomalies given the freedom from thresholding decisions. Unsupervised machine learning approaches could be particularly valuable due to their ability to identify patterns in datasets without user-imposed expectations, facilitating further discovery of anomaly indicators.

Plain Language Summary

The climate phenomenon know 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 differences in the average temperature 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 are very tightly matched with an El Niño happening, and also point to physical changes to that region in the ocean. Altogether, automatic grouping by machine learning 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 many associated oscillations, 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 period, 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 (Rasmusson & Carpenter, 1982; Wyrtki, 1975). 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 (Dijkstra et al., 2019).

Current El Niño detection relies on sea surface temperature anomalies within a specifically designated region (Niño 3.4) in the equatorial Pacific. Extensive study of historical patterns have identified this region as the dominant location of the coupled ocean-atmosphere interactions (Trenberth, 2019). However, a strictly defined rectangular geographic region and empirical thresholds are likely not robust to change, even minor shifts in oceanic and atmospheric circulation. 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. Similarly, an anomaly threshold assumes the historic running average will remain stationary into the future, an unlikely scenario in the context of global climate change and ocean warming (Yeh et al., 2009; Ashok & Yamagata, 2009). Therefore, methods for El Niño detection incorporating large horizontal and vertical scales and utilizing directly measured data without empirical thresholds are of particular value.

Direct measurements of the state of the ocean are relatively limited and substantial analysis and prediction relies on remotely sensed (e.g. sea surface temperature) or model calculated data. In situ measurements are valuable sources for subsurface measurements as well as for model validation and improvement, particularly in a changing climate. In situ measurements come with additional challenges, particularly in terms of spatial and temporal sparsity and nonuniform sampling for free-floating measurement profilers. 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.

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 (i.e. thresholds derived from prior assumptions of the system). Previous work has considered unsupervised clustering of temperature profile measurements in the Atlantic and Southern Oceans (Jones et al., 2019; Maze et al., 2017) 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 for the first time. The openly-available dataset of ocean temperature profiles from the Argo program is analyzed with unsupervised machine learning methods to reveal El Niño indicators without thresholding decisions. We find that temporal dynamics in the spatial location of cluster assignments are strongly correlated with current metrics 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 (ARGO, 2000) were reduced to a lower-dimensional embedding using principal component analysis (PCA) and then grouped via k-means clustering, an unsupervised clustering method. The spatial locations of measurements assigned to each cluster were then considered over a thirteen year time period as well as over season-length (three month) time periods. Oscillations in the spatial extent of clusters were compared to indicators of climate phenomena (El Niño) originating in the Pacific Ocean. A description of the Argo temperature dataset, dimensionality reduction and clustering methods, and comparison to El Niño-Southern Oscillation indicators are included below.

2.1 Argo Float Dataset

The Argo program was initiated in the 1990's and consists of a global array of free-drifting profiling floats that have served to substantially expand our global ocean observing network. Each profiler in the array measures the vertical structure of temperature and salinity in the ocean, with newer profilers taking into account currents and bio-optical traits. Currently, nearly 4,000 individual profilers 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 (ARGO, 2000).

Argo profiler measurements of temperature were acquired in the Pacific Ocean basin between 30°S and 50°N from January 2006 to September 2019 via the Argovis API (argovis.colorado.edu). Each measurement had an associated latitude, longitude, and acquisition timestamp. All temperature profiles containing missing data, insufficient data points, or nonphysical values were removed. This corresponded to profiles with fewer than 50 data points, the initial data point more than 25 mbar from the surface, the final data point less than 1,000 mbar, or temperature values less than -5°C. Temperature values in the remaining profiles were linearly interpolated onto a uniform grid with 5 mbar spacing from 5 mbar down to 1,000 mbar. Data was only stored down to 1,000 mbar despite measurements down to approximately 2,000 mbar due to the majority of temperature variability of interest occurring in the upper 1,000 mbar. This yielded a set of approximately 560,000 temperature profiles consisting of 199 data points each for the thirteen year time span that were subsequently assigned to clusters.

2.2 Dimensionality Reduction and Clustering

A critical first step toward effective clustering for a high-dimensional variable is dimensionality reduction (Aggarwal et al., 2001). 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) values, 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 additional components will be calculated to more completely describe variance between samples. PCA was applied to cast the 199-data-point profiles into 17 principal components to capture 99.9% of the variance.

With dimensionality reduction applied, properties such as Euclidean distance between each representation become notably more effective at describing sample differences (Aggarwal et al., 2001). Clustering methods were then applied with the goal of grouping the profiles solely based on differences in temperature and 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 (i.e. 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. In comparison to GMM, which works best when the data are multivariate Gaussian, k-means is non-parametric, is computationally efficient, and provides hard assignments to each sample. 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_i be the mean of the data within the ith cluster, C_i. The k-means algorithm seeks to identify the partition, $\boldsymbol{C}$, that minimizes

$$WCSS = \arg\min_{\mathbf{C}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_\ell} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 . \tag{1}$$

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 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 clusters 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 a large distance to non-assigned clusters and small distance to other samples in the assigned cluster. The global silhouette score can be calculated for varying cluster counts, ideally encountering a cluster count, k, that maximizes s_{global}. 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 would be useful to capture variability.

Following selection of 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. 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.

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).

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 (Trenberth, 2019). The three-month running mean of the anomaly in this region is termed the Ocean Niño Index (ONI). This index must exceed ±0.5°C for at least five consecutive months to classify the period as a full-fledged El Niño (+5°C) or La Niña (-5°C) (Trenberth, 2019). ONI values were obtained from NOAA (noaa.gov) and used directly for comparison.

2.4 Spatio-temporal Cluster Analysis

Following 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 ENSO events, utilizing the ONI as a ground truth on the historic presence or absence of an El Niño 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 NOAA reported values. The zonal (east-west) extent of measurements within a cluster was then considered. To effectively capture the changes in zonal extent of a cluster, all unique longitudes of measurements within a cluster were aggregated. The unique set of longitudes represented within a cluster were then averaged and zero-meaned. 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

K-means clustering was found to be effective at partitioning, reproducible, and highly computationally efficient. The silhouette score for cluster counts ranging 3 to 10 exhibited no global maximum, but 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 uniqueness of clusters from sufficient partitions with improvement in 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 occurring, but rather led to a coarsening (for fewer clusters) or refining (for more clusters) of the divisions along similar lines (see supplementary figure 1).

Repetition of the PCA embedding process and clustering produced very similar results such that the same profiles were consistently grouped together. Ten repeated embeddings and clusterings produced an average adjusted Rand index of 0.997, indicating high repeatability of the analysis.

Each group produced by the clustering algorithm contained profiles with relatively similar vertical structure and temperature values (figure 2) indicated by the uniqueness of the average temperature profile of each cluster and the standard deviation within the group relative to variation between groups. 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. Each measurement displayed on a map and colored by its corresponding cluster assignment (figure 3b) illustrated the spatial coherency of measurements in each cluster, with few outliers and minimal spatial overlap of cluster members. This spatial coherency was similar to previous analyses by Maze 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 profiler) was used for clustering (figure 3a), the clusters were significantly less spatially well-defined with a scattered overlap of measurements belonging to different groups, indicating the full vertical structure of the temperature profile was critical in partitioning.

3.2 Temporal Dynamics

Measurements from three-month time periods exhibited clear spatial oscillations in cluster assignments correlated with the Ocean 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, measurements in the constrained geographic region of Niño 3.4 (N3.4) were 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. During a positive ONI anomaly (El Niño event), the western cluster distinctly shifted eastward to occupy the majority of the N3.4 region. 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 longitudinal extent of measurements in the eastern cluster (figure 6a). The average longitudinal position of measurements in cluster 2 was consistently further east (positive longitudinal anomaly) during periods above the El Niño threshold, and near average or further 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 (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 3). During an elevated ONI period, the eastern cluster that had shifted further 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 the majority eastern group (cluster 2) all the way in the western Pacific in a narrow band around 10°N (figure 4). This phenomenon exhibited itself during every El Niño event during the time period assessed (2006-2019). This oscillation was quantified with the anomalous longitudinal extent of the eastern cluster (figure 6b). Opposite to the N3.4 region, on a large scale, the eastern cluster exhibited strong location anomalies to the west during El Niño events, once again in synchrony with ONI oscillations.

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 (e.g. upper, intermediate or deep) extent. Therefore, a profile measurement down to 1,000 mbar 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. The Pacific is known to have strong east-west variations in upper water masses (Emery, 2008) and contains east and west central waters in both the northern and southern hemispheres, which was seen in the partitioning of profiles in both the meridional and zonal direction. Intermediate waters are formed off the coast of California in the northern hemisphere and off the coast of South

America in the southern hemisphere as a consequence of coastal upwelling, and were also partitioned. Additionally, the Pacific is unique for its Pacific Equatorial Water, a large band spanning the low latitudes. This region was also partitioned by the clustering task, and was divided into an eastern and western cluster at low latitudes which were found to be particularly relevant in terms of temporal variability.

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 westerly 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 leveling the depth of the thermocline in that region (Wang et al., 2000; Meinen & McPhaden, 2000).

The switching of cluster assignment in a region signals a physical change to the water column indicated by the differences in temperature profiles in the two dominant oscillating clusters (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 temperatures in the east. During an El Niño, the western cluster extends further 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 Ocean 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 (Kim & Kim, 2002; Battisti, 1989). 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 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.

By analyzing the sparse (relative to grid cells of a model) but directly measured set of profiles, unsupervised clustering methods are shown to be highly effective at revealing anomalies. 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 unsupervised 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 metrics like temperature anomalies from historic means become obsolete. 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. 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

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Lecture notes in computer science.* doi: 10.1007/3-540-44503-x{_}27

ARGO. (2000). Argo floats data and metadata from Global Data Assembly Centre (Argo GDAC). *Ifremer*. doi: 10.12770/1282383d-9b35-4eaa-a9d6-4b0c24c0cfc9

Ashok, K., & Yamagata, T. (2009). The El Niño with a difference. *Nature*. doi: 10.1038/461481a

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

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 policymalers.* 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

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

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., 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

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

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 edushlenspubnotespca pdf*. doi: 10.1.1.115.3503

Trenberth, K. (2019). *The Climate Data Guide: Nino SST Indices (Nino 1+2, 3, 3.4, 4; ONI and TNI).* Retrieved from `https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni`

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

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

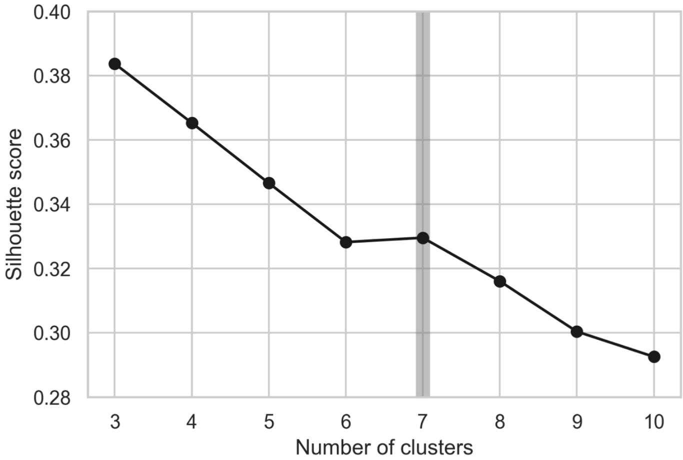

Figure 1. Silhouette score as a function of number of clusters, k, from 3 to 10 calculated following equation 4. A local maximum (highlighted in gray) is observed at $k = 7$.

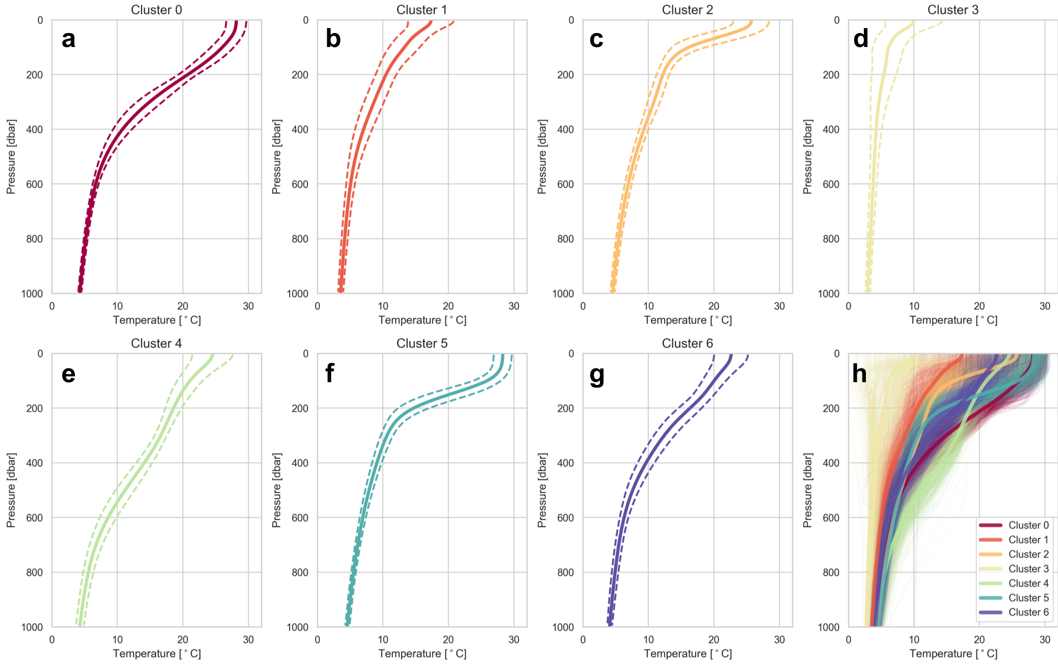

Figure 2. Temperature profiles collected by the Argo project, colored corresponding to cluster assignment. (a-g) For each cluster, the mean temperature profile (solid line) and ± one standard deviation of temperature (dashed line) is plotted. (h) Overlay of a random subset of profiles from each cluster,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

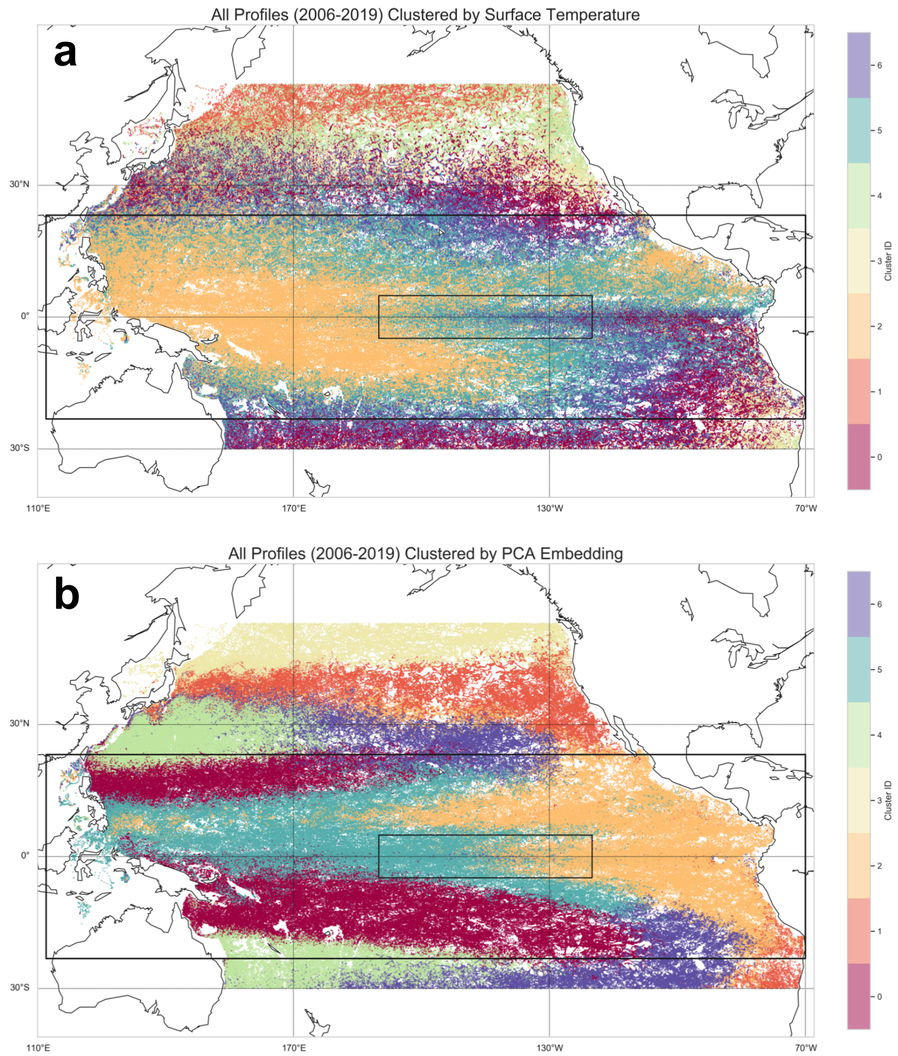

Figure 3. The spatial distribution of Argo measurements in the Pacific, colored by cluster assignment. Cluster IDs are randomly set by the clustering algorithm initialization, therefore ID magnitudes are arbitrary. The large black box corresponds to the tropical zone (± 23.4° latitude), and the smaller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

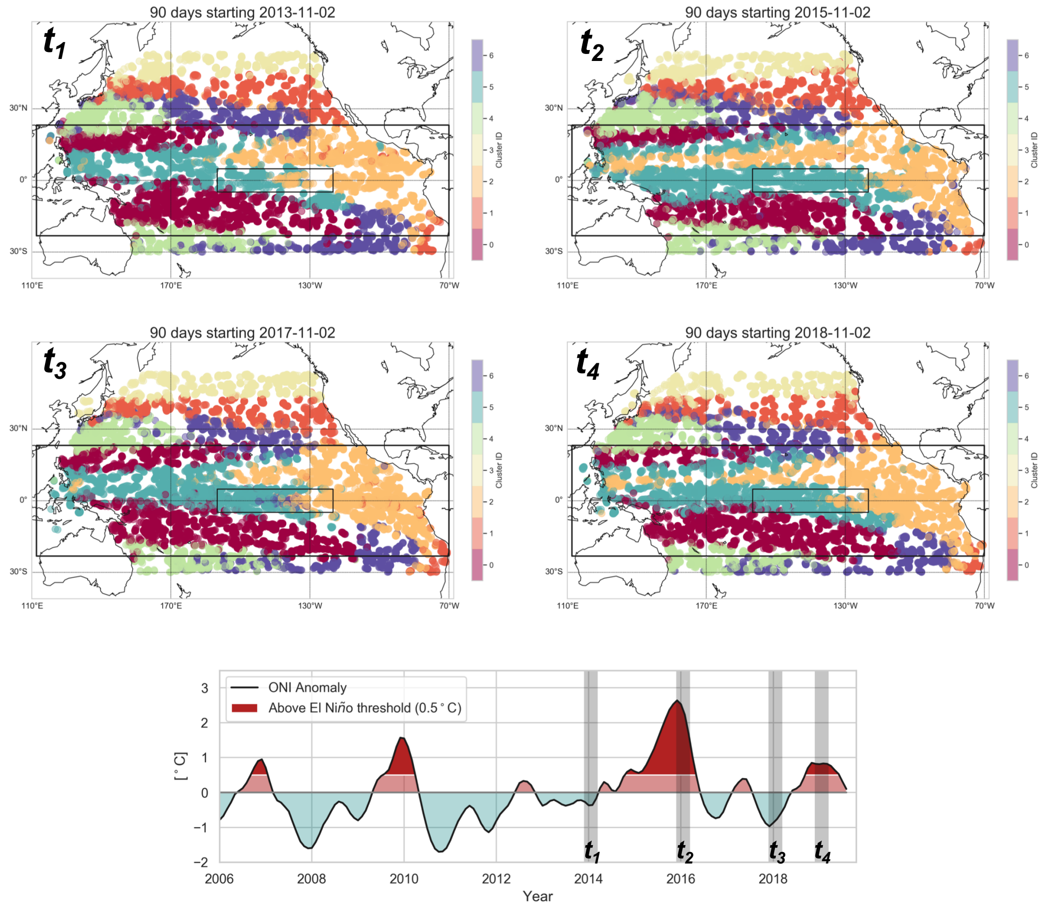

Figure 4. Upper: Three-month periods of measurements colored by cluster assignment. Two periods (t_1, t_3) correspond to a neutral ENSO phase and two periods (t_2, t_4) correspond to El Niño events during northern winter. During elevated ONI periods, the eastern cluster (2, orange) extends in a narrow band across the Pacific at approximately 10°N while simultaneously shifting westward out of the Niño 3.4 designated region. During neutral periods, the eastern cluster shifts back eastward overall, but extends slightly westward in the Niño 3.4 region (see supplementary video for cluster assignments over all time). Lower: The ONI anomaly from 2006 to 2019 indicating several El Niño events. Vertical gray shaded bars correspond to time periods visualized in upper plots.

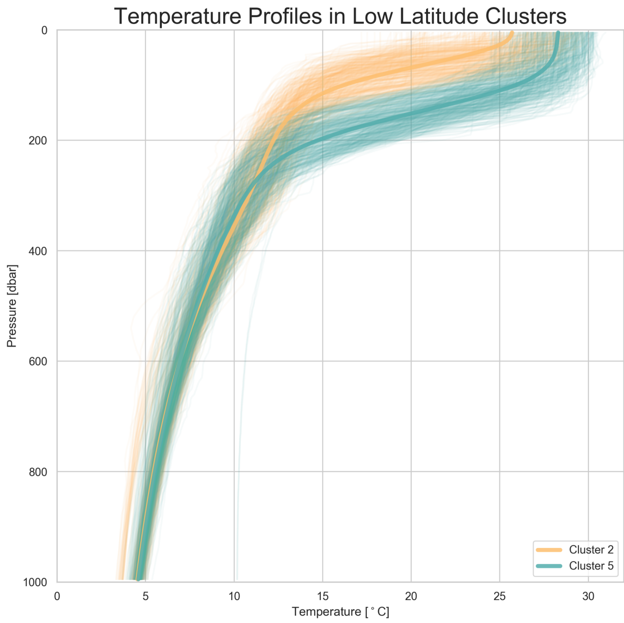

Figure 5. Relative to the eastern cluster (2), the western cluster (5) contains profiles with a warmer surface temperature and deeper thermocline. A shift in cluster assignment from 5 to 2 in a spatial region indicates a decrease in the thermocline depth and a decrease of surface temperatures.

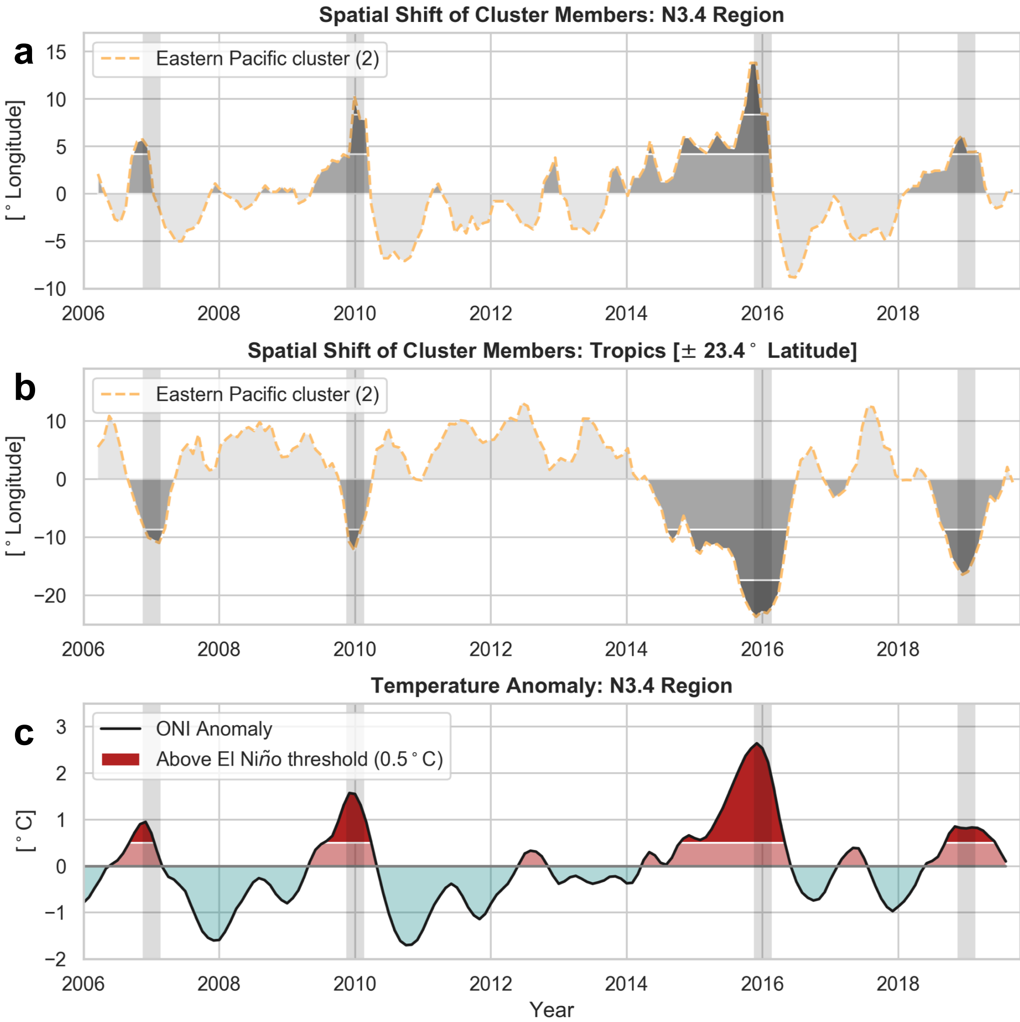

Figure 6. Spatial oscillations in the eastern low latitude cluster (2) are indicative of ENSO events. (a) During an El Niño, a shift eastward of measurements assigned to the eastern cluster is seen in the Niño 3.4 region. (b) Over the entire tropics, the eastern cluster measurements shift westward. White lines and gray shading correspond to standard deviations from the mean. All anomalies in spatial location beyond one standard deviation occur simultaneously with an El Niño event, and only the major event in 2015-2016 exceeds two standard deviations. The eastern cluster is characterized by cooler surface temperatures and a shallower thermocline (figure 5), therefore a shift of that cluster out of the N3.4 Region aligns with the positive ONI temperature anomaly. Vertical gray bars on all plots correspond to a full El Niño event occurring.