

# Discovery of Climate Indices using Clustering

Michael Steinbach  
Pang-Ning Tan  
Vipin Kumar  
Dept. of Comp. Sci. & Eng.  
University of Minnesota

steinbac,ptan,kumar@cs.umn.edu

Steven Klooster  
California State University,  
Monterey Bay  
klooster@gaia.arc.nasa.gov

Christopher Potter  
NASA Ames Research Center  
cpotter@mail.  
arc.nasa.gov

## ABSTRACT

To analyze the effect of the oceans and atmosphere on land climate, Earth Scientists have developed climate indices, which are time series that summarize the behavior of selected regions of the Earth's oceans and atmosphere. In the past, Earth scientists have used observation and, more recently, eigenvalue analysis techniques, such as principal components analysis (PCA) and singular value decomposition (SVD), to discover climate indices. However, eigenvalue techniques are only useful for finding a few of the strongest signals. Furthermore, they impose a condition that all discovered signals must be orthogonal to each other, making it difficult to attach a physical interpretation to them. This paper presents an alternative clustering-based methodology for the discovery of climate indices that overcomes these limitations and is based on clusters that represent regions with relatively homogeneous behavior. The centroids of these clusters are time series that summarize the behavior of the ocean or atmosphere in those regions. Some of these centroids correspond to known climate indices and provide a validation of our methodology; other centroids are variants of known indices that may provide better predictive power for some land areas; and still other indices may represent potentially new Earth science phenomena. Finally, we show that cluster based indices generally outperform SVD derived indices, both in terms of area weighted correlation and direct correlation with the known indices.

## Categories and Subject Descriptors

I.5.3 [Pattern Recognition]: Clustering; I.5.4 [Pattern Recognition]: Applications—*Climate*

## Keywords

clustering, singular value decomposition, time series, Earth science data, mining scientific data

## 1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD 2003 Washington, DC, USA

Copyright 2003 ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

It is well known that ocean, atmosphere and land processes are highly coupled, i.e., climate phenomena occurring in one location can affect the climate at a far away location. Indeed, understanding these climate teleconnections is critical for finding the answers to questions such as how the Earth's climate is changing and how ecosystems respond to global environmental change. A common way to study

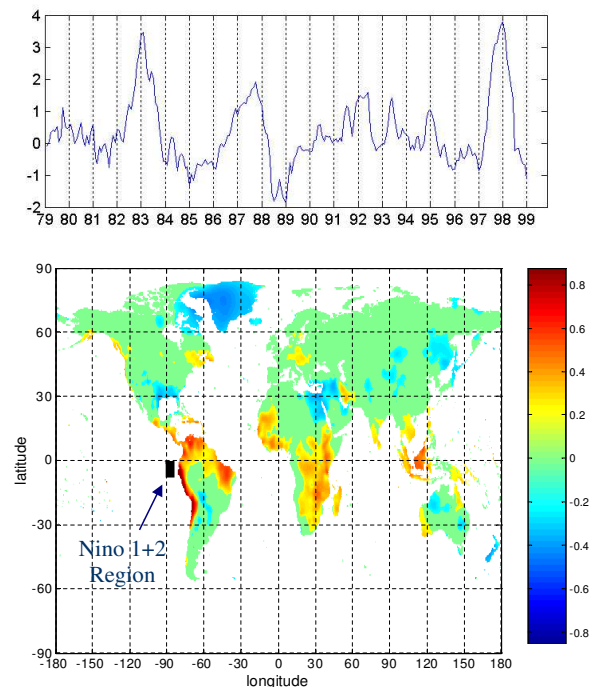


Figure 1: The NINO 1+2 index and its correlation to land temperature anomalies. (Best viewed in color.)

such teleconnections is by using climate indices [6, 7], which distill climate variability at a regional or global scale into a single time series. For example, the NINO 1+2 index, which is defined as the average sea surface temperature anomaly in a region off the coast of Peru, is a climate index that is associated with the El Nino phenomenon, which is the anomalous warming of the eastern tropical region of the Pacific. El Nino has been linked to climate anomalies in many parts of the world such as droughts in Australia and heavy

rainfall along the eastern coast of South America [18]. Figure 1 shows the correlation between the NINO 1+2 index and land temperature anomalies, which are deviations from the mean. Observe that this index is highly correlated to the land temperature anomalies on the western coast of South America, which is not surprising given the proximity of this region to the ocean region defining the index. However, few outside the field of Earth Science would expect that NINO 1+2 is also highly correlated to land regions that are far away from the eastern coast of South America, e.g., Africa and South-East Asia.

Most commonly used climate indices are based on sea level pressure (SLP) and sea surface temperature (SST) in ocean regions. These indices can ease the discovery of relationships of SST and SLP to land temperature and precipitation. These variables in turn, impact plant growth, and are therefore important for understanding the global carbon cycle and the ecological dynamics of the Earth.

Because of this, Earth Scientists have devoted a considerable amount of time to developing/discovering climate indices, such as NINO 1+2 and the other indices described in Table 1. One of the approaches used to discover climate indices has been the direct observation of climate phenomenon. For instance, the El Nino phenomenon was first noticed by Peruvian fishermen centuries ago. The fishermen observed that in some years the warm southward current, which appeared around Christmas, would persist for an unusually long time, with a disastrous impact on fishing. In the early 20th century, while studying the trade winds and Indian monsoon, scientists noticed large scale changes in pressure in the equatorial Pacific region which they referred to as the ‘Southern Oscillation.’ Scientists developed a climate index called the Southern Oscillation Index (SOI) to capture this pressure phenomenon. In the mid and late 60’s, the Southern Oscillation was conclusively tied to El Nino, and their impact on global climate was recognized. Needless to say, finding climate indices in this fashion is a very slow and tedious process.

More recently, motivated by the massive amounts of new data being produced by satellite observations, Earth Scientists have been using eigenvalue analysis techniques, such as principal components analysis (PCA) and singular value decomposition (SVD), to discover climate indices [16]. While eigenvalue techniques do provide a way to quickly and automatically detect patterns in large amounts of data, they also have the following limitations: (i) all discovered signals must be orthogonal to each other, making it difficult to attach a physical interpretation to them, and (ii) weaker signals may be masked by stronger signals. These points are discussed in more detail in Section 3.

This paper presents an alternative clustering-based methodology for the discovery of climate indices that overcomes these limitations. The use of clustering is driven by the intuition that a climate phenomenon is expected to involve a significant region of the ocean or atmosphere, and that we expect that such a phenomenon will be ‘stronger’ if it involves a region where the behavior is relatively uniform over the entire area. SNN clustering [2, 3, 4] has been shown to find such homogeneous clusters. Each of these clusters can be characterized by a centroid, i.e., the mean of all the time series describing the ocean points in the cluster, and thus, these centroids represent potential climate indices. This approach offers a number of benefits: (i) discovered signals

do not need to be orthogonal to each other, (ii) signals are more easily interpreted, (iii) weaker signals are more readily detected, and (iv) it provides an efficient way to determine the influence of a large set of points, e.g. all ocean points, on another large set of points, e.g., all land points.

The results of applying our methodology to discover cluster indices are encouraging. Some of the cluster centroids, i.e., candidate indices, that we found are very highly correlated to known indices. This represents a rediscovery of well-known indices and serves to validate our approach. In fact, we are able to rediscover most of the known major climate indices using our approach. In addition, some of the cluster centroids that have a high correlation to well-known indices may represent variants to well-known indices in that, while they may represent the same phenomena, they may be potentially better predictors of land behavior for some regions of the land. Finally, cluster centroids that have medium or low correlation with known indices may represent potentially new Earth science phenomena.

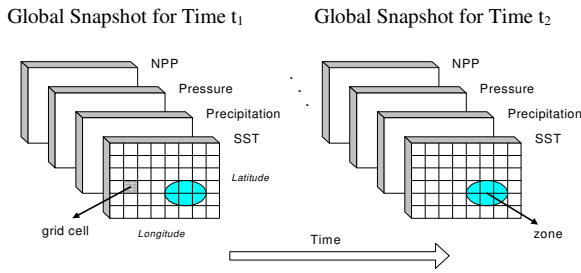
This paper is organized as follows: Section 2 provides a quick introduction to Earth Science data and climate indices, while Section 3 provides a more detailed look at how eigenvalue techniques are used to discover climate indices and the limitations of this approach. Section 4 describes our methodology and Sections 5 presents the results of applying this methodology to find climate indices that have a strong connection to land temperature. Section 6 summarizes our work and indicates future directions. Our preliminary results on this topic have appeared in several workshop papers [17, 15, 14, 13].

## 2. EARTH SCIENCE DATA AND CLIMATE INDICES

The Earth science data for our analysis consists of global snapshots of measurement values for a number of variables (e.g., temperature, pressure and precipitation) collected for all land and sea surfaces (see Figure 2). These variable values are either observations from different sensors, e.g., precipitation, Sea Level Pressure (SLP), sea surface temperature (SST), or the result of model predictions, e.g., NPP (Net Primary Production or plant growth) from the CASA model [10], and are typically available at monthly intervals that span a range of 10 to 50 years. For the analysis presented here, we focus on attributes measured at points (grid cells) on latitude-longitude spherical grids of different resolutions, e.g., land temperature, which is available at a resolution of  $0.5^\circ \times 0.5^\circ$ , and SST, which is available for a  $1^\circ \times 1^\circ$  grid, and SLP, which is available for a  $2.5^\circ \times 2.5^\circ$  grid.

Most of the well-known climate indices [6, 7] based upon SST and SLP are shown in Table 1. Many of the indices represent the El Nino phenomenon and are highly correlated, as shown in Table 2. Figure 1 shows the time series for the NINO 1+2 index. The peaks correspond to El Nino events.

The spatial and temporal nature of Earth Science poses a number of challenges. For instance, Earth Science time series data is noisy, has cycles of varying lengths and regularity, and can contain long term trends. In addition, such data displays spatial and temporal autocorrelation, i.e., measured values that are close in time and space tend to be highly correlated, or similar. To handle the issues of seasonality (yearly cycles) and temporal autocorrelation, we preprocess the data to remove seasonality. In particular, we use



**Figure 2: A simplified view of the problem domain.**

the ‘monthly Z score’ transformation, which takes the set of values for a given month, calculates the mean and standard deviation of that set of values, and then ‘standardizes’ the data by calculating the Z-score of each value, i.e., by subtracting off the corresponding monthly mean and dividing by the monthly standard deviation. The Z-score transformation also reduces temporal autocorrelation significantly. Since we used time series spanning 41 years (12 monthly measurements each year), we also detrended the data before our analysis. For further details on these issues, we refer the reader to [15] and [17].

**Table 1: Description of well-known climate indices.**

Index	Description
SOI	(Southern Oscillation Index) Measures the SLP anomalies between Darwin and Tahiti
NAO	(North Atlantic Oscillation) Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	(Arctic Oscillation) Defined as the first principal component of SLP poleward of 20° N
PDO	(Pacific Decadal Oscillation) Derived as the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20°N
QBO	(Quasi-Biennial Oscillation Index) Measures the regular variation of zonal (i.e. east-west) stratospheric winds above the equator
CTI	(Cold Tongue Index) Captures SST variations in the cold tongue region of the equatorial Pacific Ocean (6°N-6°S, 180°-90°W)
WP	(Western Pacific) Represents a low-frequency temporal function of the ‘zonal dipole’ SLP spatial pattern involving the Kamchatka Peninsula, southeastern Asia and far western tropical and subtropical North Pacific
NINO1+2	Sea surface temperature anomalies in the region bounded by 80°W-90°W and 0°-10°S
NINO3	Sea surface temperature anomalies in the region bounded by 90°W-150°W and 5°S-5°N
NINO3.4	Sea surface temperature anomalies in the region bounded by 120°W-170°W and 5°S-5°N
NINO4	Sea surface temperature anomalies in the region bounded by 150°W-160°W and 5°S-5°N

**Table 2: Correlation between known indices**

Index	NINO1+2	NINO3	NINO3.4	NINO4	CTI
NINO1+2	1	0.8336	0.6711	0.4851	0.7826
NINO3	0.8336	1	0.9440	0.7368	0.9632
NINO3.4	0.6711	0.9440	1	0.8758	0.9369
NINO4	0.4851	0.7368	0.8758	1	0.7629
CTI	0.7826	0.9632	0.9369	0.7629	1

### 3. FINDING STRONG SPATIAL OR TEMPORAL PATTERNS USING SVD ANALYSIS

Given a data matrix, whose rows consist of time series from various points on the globe, the objective is to discover the strong temporal or spatial patterns in the data. Earth scientists routinely use Empirical Orthogonal Functions (EOF), to find spatial patterns, and temporal patterns [16]. EOF is just another name for a statistical technique known as Principal Components Analysis (PCA), which, in turn, is similar to a technique from linear algebra, which is known as singular value decomposition (SVD). (For true equivalence, it is necessary to remove the mean from the data before applying SVD.) At a high level, SVD decomposes a matrix into two sets of patterns, which, for Earth science data, correspond to a set of spatial patterns and a set of temporal patterns. These patterns come in pairs, i.e., for every temporal pattern there is a corresponding spatial pattern. See [16] for a more technical description.

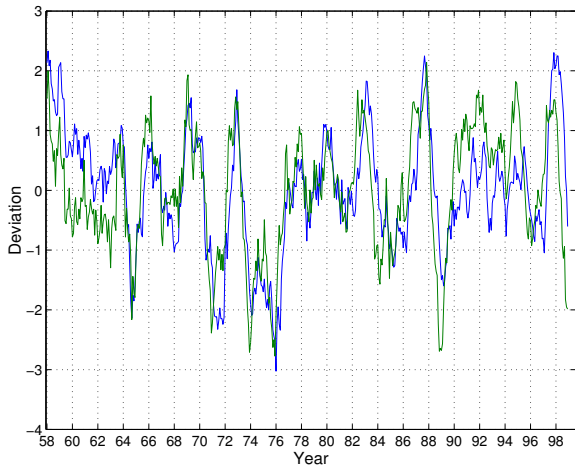
Also, for each pair of patterns, there is an associated value (called a singular value), which is greater than or equal to 0. The strongest patterns (or the patterns that capture the largest amount of variation in the data) are associated with the largest singular values, and often, by looking at only the first few singular values and their associated pairs of spatial and temporal patterns, it is possible to account for most of the variation in the data. Looked at another way, the original data can be approximated as a linear combination of these strongest patterns.

Finally, for Earth science data, we can plot the temporal patterns (right singular vectors, known as t-EOFs) in a regular line plot and the spatial patterns (left singular vectors, plain EOFs) on a spatial grid, and thus, visualize these patterns.

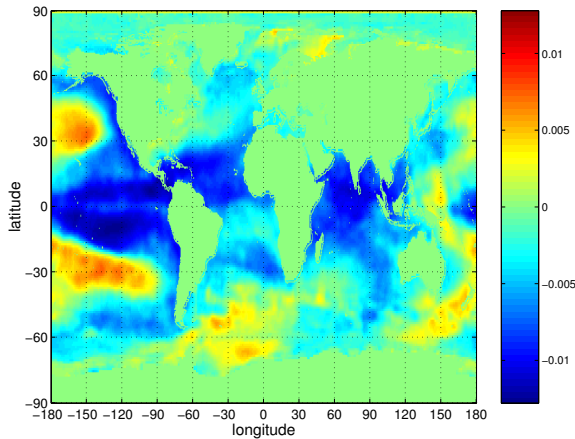
#### An SST Example

To illustrate EOFs and t-EOFs we provide an example using SST data. In the following, we use data that has been pre-processed using the monthly Z-score and detrending. (Note that the rows of this data have a mean of 0 and thus, and SVD analysis is equivalent to an EOF analysis.) Finding the top spatial and temporal patterns via SVD is a simple matter using current mathematics or statistics packages. For example, in MATLAB this requires only the following command:  $[\mathbf{u} \ \mathbf{s} \ \mathbf{v}] = \text{svds}(\mathbf{z\_sst}, \mathbf{30})$ , where  $\mathbf{z\_sst}$  is the SST data matrix which has been normalized using the monthly Z-score. The columns of  $\mathbf{u}$ , the left singular vectors, are the spatial patterns, the diagonal elements of  $\mathbf{s}$  are the singular values, and the columns of  $\mathbf{v}$ , the right singular vectors, are the temporal patterns.

For SST, the strongest temporal pattern (the first column of  $\mathbf{v}$ ) is highly related to the El Nino climate phenomenon as shown in Figure 3. The correlation of the first right singular vector with NINO4 is 0.69. The spatial pattern corresponding to the first column of  $\mathbf{v}$  is the first column of  $\mathbf{u}$ . This pattern, shown in Figure 4, indicates the strongest areas contributing to the pattern. While this gives some indication of the areas involved in El Nino, the pattern is not as well localized as for the clusters used to derive candidate indices later in the paper.



**Figure 3: First right singular vector of SST (blue) plotted against the NINO4 index (green).**



**Figure 4: Strongest spatial pattern of SST.**

### Limitations of SVD Based Approaches

The performance of SVD in the above example, is impressive: a well-known climate index was discovered straightforwardly. However, there are a number of limitations of SVD analysis, most of which are well-known. For example, SVD finds the strongest patterns best since its goal is to provide the best rank  $k$  approximation to a matrix,  $1 \leq k \leq \text{rank}(\text{data matrix})$  [1]. Thus, slightly weaker patterns may not ‘show up’ as well as the stronger patterns. For this reason—as well as the orthogonality constraint described below—only the first few SVD vectors are typically regarded as trustworthy. For example, in Earth science time series, seasonality dominates all other patterns. Thus, for SVD analysis, if the seasonality is not removed from the data, at least the first few strongest patterns will be seasonal patterns of different types. Indeed, other strong patterns, such as El Nino, emerge more clearly once seasonality is removed. However, El Nino and other strong patterns, e.g., AO, may in turn mask yet other patterns. Of course, clustering is somewhat subject to the same problem—small clusters are

hard to detect—but, we would argue, not to the extent of SVD. Indeed, as we show in Section 5, clustering approaches can potentially find many ‘good’ clusters that correspond to weak, as well as strong climate patterns.

Also, the patterns found using SVD, i.e., the singular vectors, are constrained to be orthogonal to each other. While orthogonality may be appealing mathematically, it can also make patterns hard to interpret. Researchers have developed an approach to try to address this problem—‘rotated’ EOFs—but it is somewhat controversial since its application may not provide any improvement or, in some cases, may split a single pattern into two patterns [16].

Yet another limitation of SVD is best illustrated by example. Suppose that we have a number of clusters in two dimensional space, e.g., 10, then SVD cannot find all of these ‘patterns’ because  $u$  and  $v$  consist only of two vectors. More generally, SVD will find patterns if they fall into independent subspaces, but cannot distinguish between patterns that lie within a subspace, and may have problems with patterns in overlapping subspaces.

Finally, while efficiency can be a concern for the SVD approach, we note that even with our biggest current data set—consisting of 70,000 time series of length 492—SVD computation times are still acceptable: less than 10 minutes on a 2.8GHz Intel processor.

## 4. CLUSTERING BASED METHODOLOGY FOR THE DISCOVERY OF INDICES

In this section, we elaborate on two key steps for finding climate indices: find candidate indices using clustering, and then evaluate these candidate indices for Earth Science significance.

### 4.1 Use clustering to find candidate climate indices

Various techniques can be used for clustering Earth Science data. For instance, a mixture model approach has been used to identify the cluster structure in atmospheric pressure data [12], a novel divisive clustering algorithm, ACTS (Automatic Classification of Time Series), was used for land use classification [20], and a modified version of a hierarchical clustering algorithm was used for land use segmentation [19]. Furthermore, we also used the K-means [8] and DBSCAN [5] clustering algorithms in some of our preliminary investigations of Earth Science data.

However, in this work, we will use our SNN clustering approach described in [2, 3, 4]. This technique first finds the nearest neighbors of each data point and then redefines the similarity between pairs of points in terms of how many nearest neighbors the two points share. Using this definition of similarity, our algorithm identifies core points and then builds clusters around the core points. The use of a shared nearest neighbor definition of similarity alleviates problems with varying densities and high dimensionality, and the use of core points handles problems with shape and size. Furthermore, the number of clusters is automatically determined by the location and distribution of core points. Another novel aspect of the SNN clustering algorithm is that the resulting clusters do not contain all the points, but rather, contain only points that come from regions of relatively uniform density. These features allow the algorithm to find clusters that other approaches over-

look, i.e., clusters of low or medium density which represent relatively uniform regions ‘surrounded’ by non-uniform or higher density areas. With respect to Earth Science data, SNN clustering produces high quality clusters, which are almost always geographically contiguous, and automatically selects the number of clusters. Because of space considerations, we omit a detailed description of the SNN algorithm and refer the reader to [2, 4]. Note that [4] contains experimental results indicating that SNN clustering works better than DBSCAN or K-means when the goal is to find clusters of relatively uniform densities.

## 4.2 Evaluate Candidate Indices

To evaluate potential ocean climate indices, we need a numerical measure of the strength of the association between the behavior of a candidate index and land climate. Earth Scientists have traditionally used correlation [9] to measure the degree of association between two time series. In general, we are only interested in using a time series (cluster centroid, or otherwise) as an index if it shows a strong correlation—positive or negative—with the behavior of a well-defined region of the land and if the strength of that correlation is comparable to that of known indices.

To evaluate the influence of climate indices on land, we use *area-weighted correlation*, which is the weighted average of the correlation of the climate index with all land points, where weight is based on the area of the land grid point. Area-weighted correlation is computed as follows. We first compute the correlation of the time series of the candidate climate index with the time series associated with each land point. We then compute the weighted average of the absolute correlations of each land point, where the weight associated with each land point is just its area. (We use absolute correlation because we are interested in the strength of the connections between ocean/atmosphere and land, not the direction.) The resulting area-weighted correlation value can be at most 1 (this would be the case where all land time series have a correlation of 1 or -1 with the candidate index), but is normally much lower. The minimum value of area weighted correlation is 0.

Note that if area weighted correlation of a candidate index is too low, it could mean that either it is not a good index, or it impacts only a small number of land grid points.

### Accounting for Lags

As mentioned earlier, since the same Earth Science phenomenon can occur at different times in different places, we want to use a form of area weighted correlation that takes these lags into account. Thus, it is necessary to compute the area-weighted correlation for various shifts. This involves shifting two time series to simulate leads (lags) of up to six months, computing the correlation, and then taking the ‘best’ (highest positive or negative value) as the correlation. This procedure is used in all of our evaluations.

### A Baseline for Area Weighted Correlation

We need a baseline to compare the values of the area weighted correlations of candidate indices. First we decided to look at the area weighted correlation of random time series since this gives us a means of judging how spurious the area weighted correlation of a time series might be. The histogram of the area weighted correlation of 1000 random

time series is shown in Figure 5. We can see that no random time series have an area weighted correlation greater than 0.09, which indicates that 0.1 might be a good baseline level to use as a threshold for deeming a candidate index worthy of consideration.

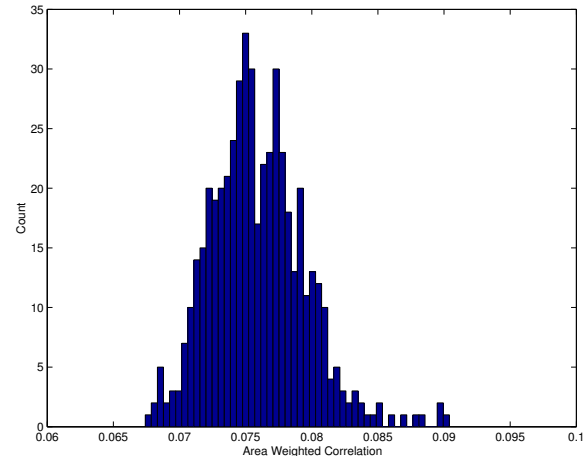


Figure 5: Histogram of area-weighted correlation of random time series vs. land temperature.

To provide a further check as to the usefulness of this threshold, we computed the total area-weighted correlations for 11 well-known indices, which are shown in Figure 6. Note that all indices, except for one, have an area weighted correlation greater than 0.1. Thus, if a candidate index shows an area-weighted correlation that is roughly 0.1 or higher, we investigate it further.

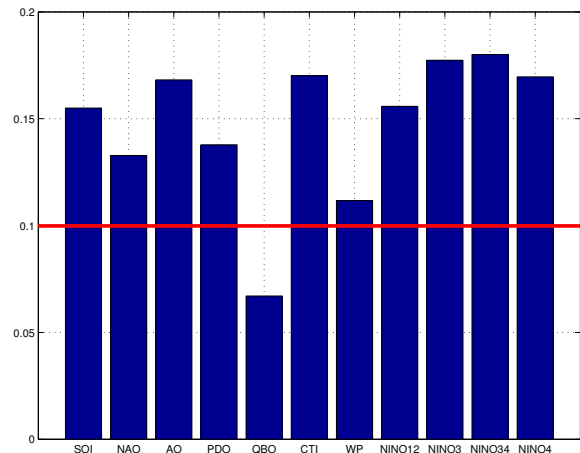


Figure 6: Area weighted correlation of well-known indices to land temperature.

## 5. DISCOVERY AND ANALYSIS OF CLUSTER-BASED INDICES

In this section we describe the results of discovering SST and SLP climate indices using the SNN clustering approach.

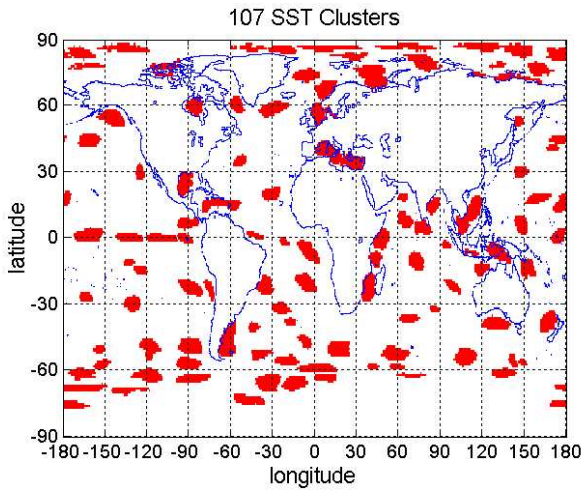


Figure 7: 107 SST clusters.

Because of space constraints, we mostly focus on SST, although some SLP results are also presented. As part of this discussion, we also compare potential climate indices derived by an SVD approach to the cluster-based indices.

### 5.1 SST-based Candidate Indices

We applied SNN clustering on the SST data over the time period from 1958 to 1998. There are 107 clusters found by SNN, as shown in Figure 7. Note that many grid points from the ocean do not belong to any clusters (these are the points belonging to the white background), as these points come from regions that are not relatively uniform and homogeneous. Since we are mainly interested in finding strong candidate indices, we first eliminate all clusters with poor area-weighted correlation, i.e., below the specified baseline of 0.1. The cluster centroids of the remaining clusters are potential candidate indices.

For further evaluation of the candidate indices, we divided the cluster centroids into 4 groups, G0, G1, G2, and G3, depending on the correlation of the cluster centroids to known indices. Cluster centroids (G0) that are very highly correlated to known indices represent a rediscovery of well-known indices and serve to validate our approach. Cluster centroids (G1) that have a high correlation to well-known indices represent variants of existing indices, but can be useful alternatives if they are better predictors of land behavior, at least for some regions of the land. Finally, cluster centroids (G2 and G3) with medium to low correlation may represent potentially new Earth science phenomena. These four groups of clusters are shown in Figures 8-11.

#### Candidate Indices Similar to Known Indices

Figure 8 shows clusters that reproduce some well-known climate indices. In particular, we were able to replicate the four El Niño SST-based indices: cluster 94 corresponds to NINO 1+2, 67 to NINO 3, 78 to NINO 3.4, and 75 to NINO 4. The correlations of these clusters to their corresponding indices are higher than 0.9, as shown in the second and third columns of Table 3. In addition, cluster 67 is highly correlated to the CTI index, which is defined over a wider area

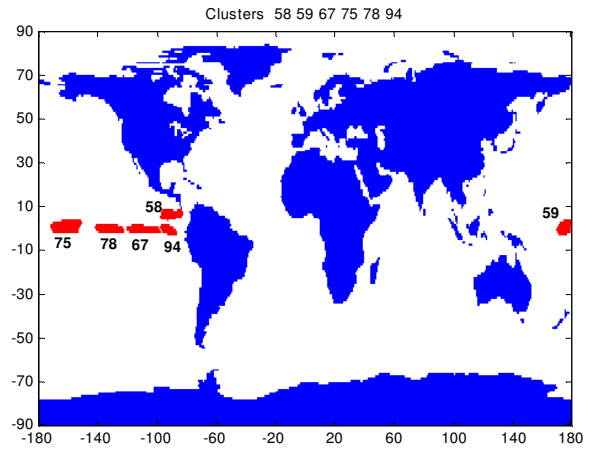


Figure 8: G0: Clusters with correlation to known indices  $\geq 0.8$ .

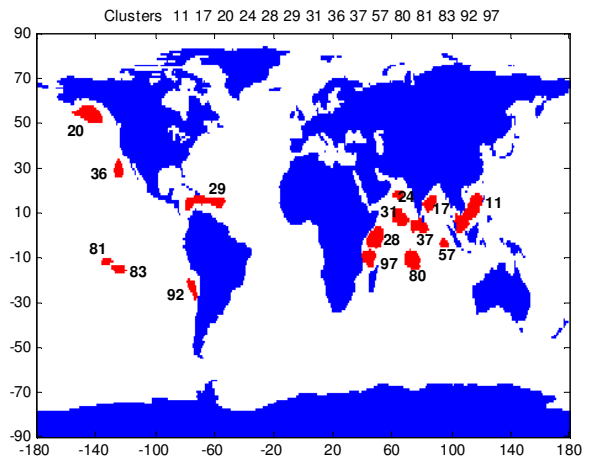


Figure 9: G1: Clusters with correlation to known indices between 0.4 and 0.8.

in the same region. Clusters 58 and 59 are very similar to the other El Niño indices, and correlate most strongly with NINO 3 and NINO 4, respectively. But their correlations to the El Niño indices are not as high as the other 4 clusters.

#### Candidate Indices that Are Variants of Known Indices

Although the clusters that are highly and moderately correlated with known indices probably capture similar Earth science phenomena, there is still benefit to considering such variants of known indices. In particular, some cluster centroids provide better ‘coverage,’ i.e., higher correlation, for some areas of the land. This is illustrated in Figure 12, which compares the El Niño indices to that of cluster 29 (G1). Color indicates the difference in correlation. Areas of yellow indicate where the cluster centroids have higher correlation, while areas of blue indicate where the El Niño indices have higher correlation. It is clear that for these clusters, there are areas of the land where the cluster ‘outperforms’ the known indices. Similar results are obtained

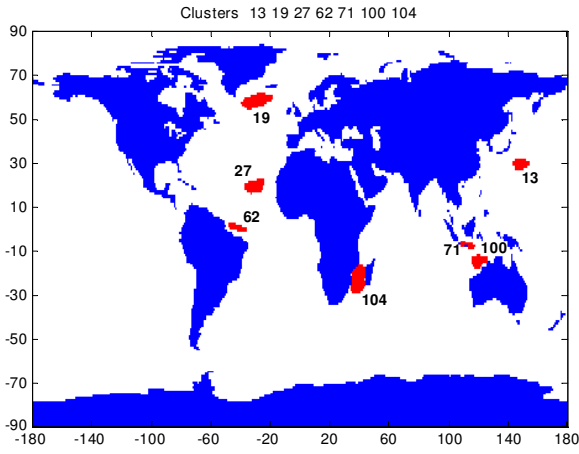


Figure 10: G2: Clusters with correlation to known indices between 0.25 and 0.4.

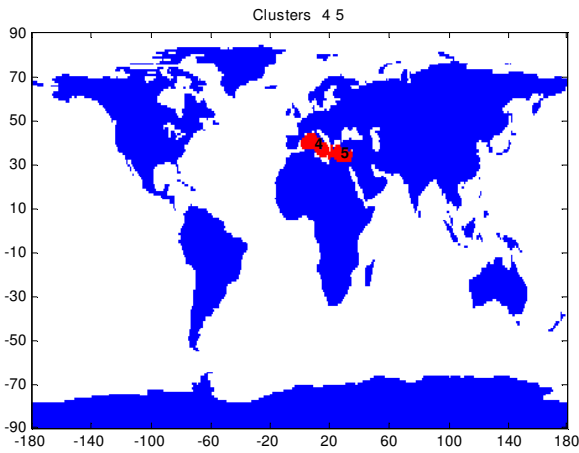


Figure 11: G3: Clusters with correlation to known indices  $\leq 0.25$ .

for other clusters in this group.

### Candidate Indices that May Represent New Indices

The clusters that are less well correlated with known indices (G2 and G3) may represent new Earth science phenomena or weaker versions or variations of known phenomena. Again, some of these cluster centroids provide better ‘coverage,’ i.e., higher correlation, for some areas of the land. This is illustrated in Figure 13, which compares the El Nino indices to that of clusters 62 (G2). Once more, areas of yellow indicate where the cluster centroids have higher correlation, while areas of blue indicate where the El Nino indices have higher correlation. Observe that cluster 62 ‘outperforms’ the known indices for some areas of the land. In contrast, clusters 4 and 5 in the group G3 derive most of their area weighted correlation from the land areas located close to the clusters. Hence, they do not appear to represent significant teleconnections.

Table 3: Correlation between known indices with SST cluster centroids and SVD components.

Climate Indices	Cluster Centroids		SVD Components	
	Best-shifted Correlation	Best Centroid	Best-shifted Correlation	Best SVD Component
SOI	-0.7006	75 (G0)	-0.5427	3
NAO	-0.2973	19 (G2)	0.1774	8
AO	-0.2383	29 (G1)	0.2301	8
PDO	0.5172	20 (G1)	-0.4684	7
QBO	-0.2675	20 (G1)	0.3187	11
CTI	0.9147	67 (G0)	0.6316	3
WP	0.2590	78 (G0)	0.1904	3
NINO1+2	0.9225	94 (G0)	-0.5419	1
NINO3	0.9462	67 (G0)	-0.6449	1
NINO3.4	0.9196	78 (G0)	-0.6844	1
NINO4	0.9165	75 (G0)	-0.6894	1

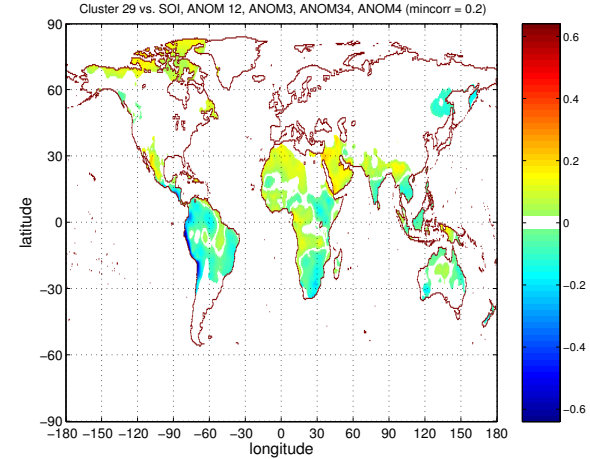


Figure 12: Cluster 29 vs. El Nino Indices. (Best viewed in color.)

### Comparison with SVD based Indices

To compare the performance of our cluster-based candidate indices with the SVD based technique, we calculated the correlation of the first 30 SVD components of SST to the El Nino Indices.

The area-weighted correlations are computed for the SST cluster centroids and the SVD components of SST. The results for the clusters from G0 and G1 are shown in Figure 14. Figure 15 shows the results for the SVD components. The area-weighted coverage baseline 0.10 is shown as a horizontal line in both figures. Many of the cluster-based candidate indices have area weighted correlations significantly above 0.10, while only about six of the SVD components have area weighted correlations above 0.10. Notice that the first SVD component has the best area weighted correlation—better even than the best of the known indices. However, from Table 3, it can be seen that this component correlates with four of the El Nino indices. (El Nino is the most dominant climate phenomenon.) Therefore, despite the large area weighted correlation, this component appears to have mixed the various El Nino based indices.

The correlation of both SST-based clusters and SVD components for NAO and AO is poor. However, the pressure cluster-based candidate indices, described in Section 5.2, have much higher correlation to NAO and AO. For QBO,

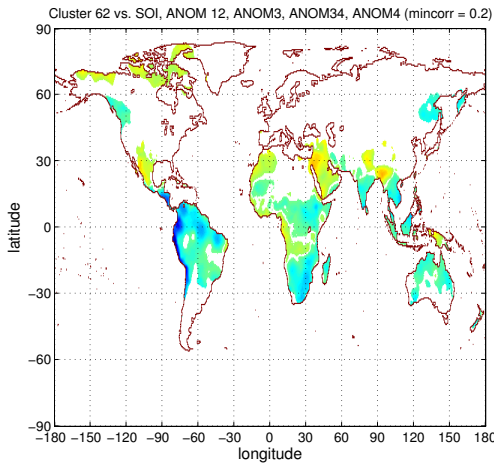


Figure 13: Cluster 62 vs. El Niño Indices. (Best viewed in color.)

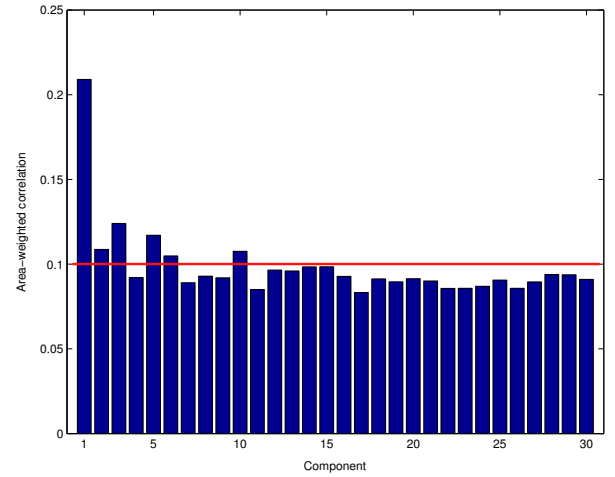


Figure 15: Area weighted correlation of first 30 SVD components of SST to land temperature.

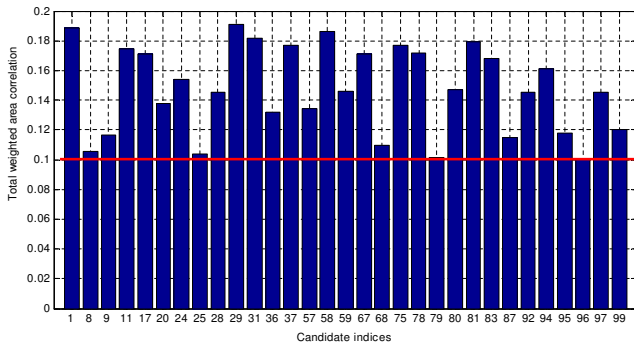


Figure 14: Area weighted correlation of highest SST cluster centroids to land temperature.

Table 4: Area-weighted correlation for known indices with SST cluster centroids and SVD components.

Index Indices	Area Weighted Correlation for		
	Index	Best Centroid	Best SVD Component
SOI	0.1550	0.1768	0.1240
NAO	0.1328	0.1387	0.0929
AO	0.1682	0.1912	0.0929
PDO	0.1378	0.1377	0.0891
QBO	0.0671	0.1377	0.0850
CTI	0.1702	0.1708	0.1240
WP	0.1117	0.1714	0.1240
NINO1+2	0.1558	0.1608	0.2091
NINO3	0.1774	0.1708	0.2091
NINO 3.4	0.1800	0.1714	0.2091
NINO 4	0.1696	0.1768	0.2091

the SVD component is better than the cluster based index, while for WP the reverse is true. However, in both cases, the correlation is poor.

The best correlated SVD component for each well known index is shown in the last column of Table 3. Notice that the candidate indices produced by clustering SST match the El Niño-based indices much more closely than the SVD components.

Finally, Table 4 compares the area weighted correlations of the centroids and SVD components chosen as most highly correlated to the known indices. As can be seen, for the El Niño based indices, the area weighted correlations for the cluster centroids are comparable to or better than the area weighted correlations for the actual indices in most cases.

## 5.2 SLP-based Candidate Indices

Figure 16 shows the clusters we found for sea level pressure (SLP) between the time period of 1958 to 1998. Since pressure differences are important in weather and climate, negative correlations between pairs of clusters are also important. Indeed, most of the well-known pressure-based indices (Table 1) are defined as the difference of two pressure time series. Therefore, for the SLP clusters, we generated

candidate indices by considering the difference of all pairs of cluster centroids. Figure 17 shows the pairwise correlation between SLP clusters. Typically, the pairs that are promising candidate indices are negatively correlated. For example, the negatively correlated pairs shown in Figure 17 serve as potential candidates for pressure-based climate indices. As shown in Table 5, c13-c20 corresponds to SOI, c7-c10 corresponds to NAO, and c10-c18 corresponds to AO.

We also calculated the correlation of the first 30 SVD components of SLP to some of the known indices. Only the second and third components were highly correlated to these indices. (The first SVD component had little correlation with known indices. It may be a mixture of other indices, but more investigation is needed.) We show these correlation results in Table 5. Notice that the candidate indices produced by clustering SLP match the known indices more closely, except for SVD component 2 which matches AO more closely. However, since AO is defined as the first principal component of the region located to the north of 20°N, it is not surprising to find that one of the SVD components of our world-wide SLP data—the second component—is highly correlated with the AO index.

Table 6 summarizes the area weighted correlations of the indices, along with the cluster pairs and SVD components



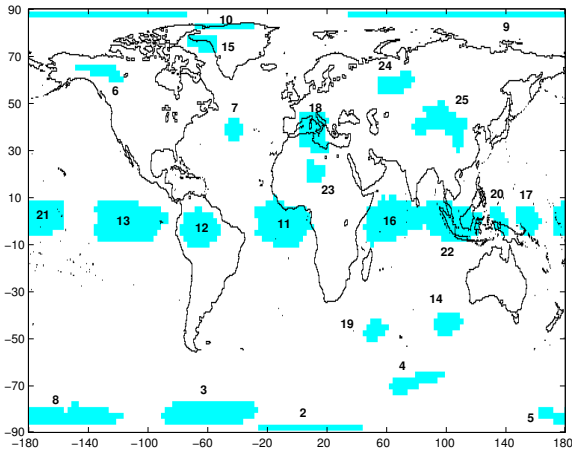


Figure 16: 25 SLP clusters.

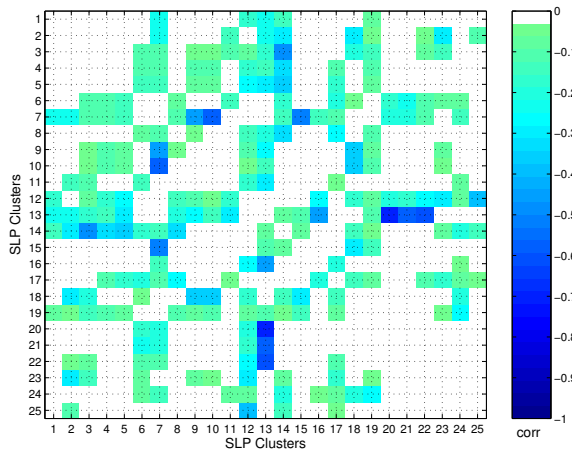


Figure 17: Smoothed pairwise correlation of SLP clusters. (Only negative correlation values shown.)

that have the strongest correlations to the respective indices. For SOI and the El Niño based indices, the area weighted correlation of the cluster pair is significantly higher than the area weighted correlation of the corresponding SVD component. In fact, for SOI and NINO1+2, the cluster pairs have higher area weighted correlations compared to their actual indices. For AO and NAO, area weighted correlations of SVD based indices are higher than the cluster pairs. But note that the same SVD component is chosen for both indices, indicating a possible mixing of the time series.

Recently a new index, the Indian Ocean Dipole Mode index (DMI), has been discovered [11]. (Because the time series for this index is shorter than other known indices, we do not incorporate this index into our analyses above.) This index is defined as the difference in SST anomaly between the region  $5^{\circ}\text{S}-5^{\circ}\text{N}$ ,  $55^{\circ}\text{E}-75^{\circ}\text{E}$  and the region  $0^{\circ}-10^{\circ}\text{S}$ ,  $85^{\circ}\text{E}-95^{\circ}\text{E}$ , and is an indicator of a weak monsoon over the Indian subcontinent and heavy rainfall over East Africa. Figure 18 shows a plot of  $c16-c22$  versus the Indian Ocean Dipole Mode index. Note the close connection between the two time series: the correlation between them is

Table 5: Correlation between known indices with SLP cluster centroids and SVD components.

Climate Indices	Cluster Pairs		SVD Components	
	Best-shifted Correlation	Best Pair	Best-shifted Correlation	Best SVD Component
SOI	-0.7312	c13 - c20	0.6137	3
NAO	0.7519	c7 - c10	0.5979	2
AO	-0.7577	c10 - c18	0.8203	2
PDO	-0.3853	c13 - c22	-0.3399	3
QBO	-0.2153	c2 - c4	0.1480	29
CTI	-0.8463	c13 - c20	-0.6295	3
WP	-0.2857	c13 - c20	0.2731	11
NINO1+2	-0.7468	c13 - c21	-0.4944	3
NINO3	-0.8469	c13 - c20	-0.5899	3
NINO 3.4	-0.8456	c13 - c20	-0.6290	3
NINO 4	-0.7171	c13 - c20	-0.5656	3

Table 6: Area-weighted correlation for known indices with SLP cluster centroids and SVD components.

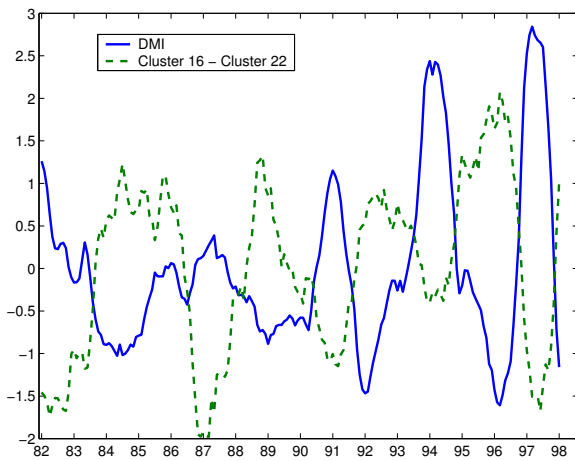
Index Indices	Area Weighted Correlation for		
	Index	Best Centroid	Best SVD Component
SOI	0.1550	0.1715	0.1416
NAO	0.1328	0.1393	0.1658
AO	0.1682	0.1561	0.1658
PDO	0.1378	0.1621	0.1416
QBO	0.0671	0.0871	0.0921
CTI	0.1702	0.1715	0.1416
WP	0.1117	0.1715	0.1020
NINO1+2	0.1558	0.1712	0.1416
NINO3	0.1774	0.1715	0.1416
NINO 3.4	0.1800	0.1715	0.1416
NINO 4	0.1696	0.1715	0.1416

-0.5315. Another interesting observation is that even though the index is defined using SST,  $c16-c22$  is a pressure based index.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper we demonstrated that clustering can provide an alternative approach to eigenvalue analyses—based on PCA or SVD—for finding ocean climate indices. Specifically, we showed that centroids of many clusters of SST and SLP data found using the SNN clustering algorithm correspond to known climate indices and provide a validation of our methodology; other centroids are variants of known indices that may provide better predictive power for some land areas; and still other indices may represent potentially new Earth science phenomena. Indices that represent new Earth science phenomena will require further evaluation by domain specialists. In addition, we compared potential indices derived from using SVD to our candidate indices and to well-known indices, showing that, in general, the SVD derived indices had lower area weighted correlation than many of the cluster-derived candidate indices and the well-known indices. This comparison showed that cluster based indices generally outperform SVD derived indices, both in terms of area weighted correlation and direct correlation with the known indices.

It should be noted that SVD results were obtained by using the data for the entire ocean (SST), or for SLP, for the entire globe. Earth scientists typically apply SVD analysis to a select region. However, it requires a considerable amount of domain knowledge to determine the appropriate area. Clustering on the other hand, provides a means of



**Figure 18: Plot of c16-c22 versus the Indian Ocean Dipole Mode index. (Indices smoothed using 12 month moving average.)**

automatically identifying regions that may be of interest.

In the future, we plan to extend our current work to address several unresolved issues. Specifically, we want to determine if there are any climate indices that cannot be represented using clusters. Also, we have also begun to investigate the effect of eliminating any correlations whose magnitudes are below a certain threshold. The idea is to eliminate noise and to see if looking only at stronger correlations produces better results. Finally, we intend to extend our analyses to other land and ocean variables and to investigate ways of aggregating the data so as to make patterns easier to detect.

## 7. ACKNOWLEDGMENTS

This work was partially supported by NASA grant # NCC 2 1231, NSF ACI-9982274, and by Army High Performance Computing Research Center cooperative agreement number DAAD19-01-2-0014. The content of this work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred. Access to computing facilities was provided by the AHPCRC and the Minnesota Supercomputing Institute.

## 8. REFERENCES

- [1] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, January 1997.
- [2] L. Ertöz, M. Steinbach, and V. Kumar. Finding topics in collections of documents: A shared nearest neighbor approach. In *Proceedings of Text Mine'01, First SIAM International Conference on Data Mining, Chicago, IL, USA, 2001*.
- [3] L. Ertöz, M. Steinbach, and V. Kumar. A new shared nearest neighbor clustering algorithm and its applications. In *Workshop on Clustering High Dimensional Data and its Applications, SIAM Data Mining 2002, Arlington, VA, USA, 2002*.
- [4] L. Ertöz, M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of Second*

*SIAM International Conference on Data Mining, San Francisco, CA, USA, May 2003*.

- [5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD 1996*, pages 226–231, 1996.
- [6] <http://www.cgd.ucar.edu/cas/catalog/climind/>.
- [7] <http://www.cdc.noaa.gov/USclimate/Correlation/help.html>.
- [8] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall Advanced Reference Series. Prentice Hall, Englewood Cliffs, New Jersey, March 1988.
- [9] B. Lindgren. *Statistical Theory*. CRC Press, January 1993.
- [10] C. Potter, S. A. Klooster, and V. Brooks. Inter-annual variability in terrestrial net primary production: Exploration of trends and controls on regional to global scales. *Ecosystems*, 2(1):36–48, August 1999.
- [11] N. Saji, B. Goswami, P. Vinayachandran, and T. Yamagata. A dipole mode in the tropical indian ocean. *Nature*, 401:360–363, 1999.
- [12] P. Smyth, K. Ide, and M. Ghil. Multiple regimes in northern hemisphere height fields via mixture model clustering. *Journal of Atmospheric Science*, 56:3704–3723, 2000.
- [13] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Temporal data mining for the discovery and analysis of ocean climate indices. In *Proceedings of the KDD Temporal Data Mining Workshop, Edmonton, Alberta, Canada, August 2002*.
- [14] M. Steinbach, P.-N. Tan, V. Kumar, C. Potter, and S. Klooster. Data mining for the discovery of ocean climate indices. In *Mining Scientific Datasets Workshop, 2nd Annual SIAM International Conference on Data Mining, April 2002*.
- [15] M. Steinbach, P.-N. Tan, V. Kumar, C. Potter, S. Klooster, and A. Torregrosa. Clustering earth science data: Goals, issues and results. In *Proceedings of the Fourth KDD Workshop on Mining Scientific Datasets, San Francisco, California, USA, August 2001*.
- [16] H. V. Storch and F. W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, July 1999.
- [17] P.-N. Tan, M. Steinbach, V. Kumar, S. Klooster, C. Potter, and A. Torregrosa. Finding spatio-temporal patterns in earth science data. In *KDD Temporal Data Mining Workshop, San Francisco, California, USA, August 2001*.
- [18] G. H. Taylor. Impacts of the el nino/southern oscillation on the pacific northwest. Technical report, Oregon State University, Corvallis, Oregon, 1998.
- [19] J. C. Tilton. Image segmentation by region growing and spectral clustering with a natural convergence criterion. In *Proc. of the 1998 International Geoscience and Remote Sensing Symposium (IGARSS '98), Seattle, WA, 1998*.
- [20] N. Vivoy. Automatic classification of time series (acts): a new clustering method for remote sensing time series. *International Journal of Remote Sensing*, 2000.