Temporal Data Mining for the Discovery and Analysis of Ocean Climate Indices [*]

Michael Steinbach[+] Pang-Ning Tan[+] Vipin Kumar[+]
Steven Klooster[+++] Christopher Potter[++]

[+] Department of Computer Science and Engineering, Army HPC Research Center
University of Minnesota
{steinbac, ptan, kumar@cs.umn.edu}

[++] NASA Ames Research Center
{cpotter@mail.arc.nasa.gov}

[+++] California State University, Monterey Bay
{klooster@gaia.arc.nasa.gov}

ABSTRACT

To predict the effect of the oceans on land climate, Earth Scientists have developed ocean climate indices (OCIs), which are time series that summarize the behavior of selected areas of the Earth's oceans. For example, the Southern Oscillation Index (SOI) is an OCI that is associated with El Nino. 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 ocean climate indices. However, these techniques are only useful for finding a few of the strongest signals and, furthermore, impose a condition that all discovered signals must be orthogonal to each other. We have developed an alternative methodology for the discovery of OCIs that overcomes these limitations and is based on clusters that represent ocean regions with relatively homogeneous behavior. The centroids of these clusters are time series that summarize the behavior of these ocean areas. We divide the cluster centroids into several categories: those that correspond to known OCIs, those that are variants of known OCIs, and those that represent potentially new OCIs. The centroids that correspond to known OCIs provide a validation of our methodology, while some variants of known OCIs may provide better predictive power for some land areas. Finally, we show that, in some sense, our current cluster centroids are relatively complete, i.e., capture most of the possible candidate OCIs.

Keywords

clustering, time series, Earth science data, correlation, scientific data mining

1. INTRODUCTION

Teleconnections are the simultaneous variation in climate and related processes over widely separated points on the Earth. For example, El Nino, the anomalous warming of the eastern tropical region of the Pacific, has been linked to climate phenomena such as droughts in Australia and heavy rainfall along the Eastern coast of South America [Tay98]. For this paper, we will be concerned with teleconnections, such as El Nino, that involve the relationship of the ocean to land climate. To capture these ocean-land teleconnections, Earth Scientists have developed ocean climate indices (OCIs), which are time series that summarize the behavior of selected areas of the Earth's oceans [IND1, IND2].

Our interest in OCIs arises from a desire to use climate variables, such as long term sea level pressure (SLP) and sea surface temperature (SST), to discover interesting patterns relating changes in NPP (“plant growth”) to land surface climatology and global climate. NPP (Net Primary Production) is the net assimilation of atmospheric carbon dioxide (CO_2) into organic matter by plants, and ecologists who work at the regional and global scale have identified NPP as a key variable for understanding the global carbon cycle and the ecological dynamics of the Earth. Terrestrial NPP is driven by solar radiation and can be constrained

[*] This work was partially supported by NASA grant # NCC 2 1231, NSF grant Number EIA 9818338, 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.

by precipitation and temperature. Keeping track of NPP is important because it includes the food source of humans and all other animals, and thus, sudden changes in the NPP of a region can have a direct impact on the regional ecology. Predicting NPP based on, for example, sea surface temperature would be of great benefit given the near real-time availability of SST data and the ability of climate forecasting to anticipate SST El Nino/La Nina events. An ecosystem model for predicting NPP, CASA (the Carnegie Ames Stanford Approach [PKB99]), has been used for over a decade to produce a detailed view of terrestrial productivity. Our goal in the investigations of OCIs is to use an improved understanding of the effect of OCIs on land climate to enhance the CASA model.

Earth scientists have used observation and eigenvalue analysis techniques, such as principal components analysis (PCA) and singular value decomposition (SVD), to discover ocean climate indices [SZ98]. However, these techniques are only useful for finding a few of the strongest signals and, furthermore, impose a condition that all discovered signals must be orthogonal to each other. After a brief overview of the type of data we are working with (Section 2), we present a more complete discussion of the limitations of eigenvalue based approaches and how an alternative approach, based on clustering, can overcome these limitation (Section 3).

In a previous paper, [Ste+01] we described this clustering based approach to the discovery of OCIs, and showed that our approach was capable of discovering some of the well-known OCIs, such as those related to El Nino. For this paper we focus on showing how this methodology can be used to identify clusters that represent potentially useful OCIs that are different from known OCIs. We will use the procedure of [Ste+01] in a slightly simplified form as described below.

1) **Use clustering to find areas of the oceans that have relatively homogeneous behavior.** Each of these clusters can be characterized by a centroid, i.e., the mean of all the time series describing the ocean points that belong to the cluster, and this centroid represents a potential OCI.
2) **Evaluate the influence of potential OCIs on land points.** Specifically, we are only interested in using a time series (cluster centroid, or otherwise) as an OCI if it shows a strong connection (correlation) with the behavior of a well-defined region of the land. One way of evaluating OCI "impact" on the land is to compute the area-weighted* correlation of each cluster centroid (potential OCI) with each land point, where the behavior of a land point is described by a time series which captures the time dependent behavior of some variable, e.g., temperature or precipitation, associated with the land point. The clusters or pairs of clusters which strongly "affect" many land points are potential candidates for climate indices. (Note: some of the words are italicized to indicate that correlation does not imply causality. However, often causal relationships exist in this domain and for simplicity, we may sometimes use causal terminology.)
3) **Compare the influence of candidate OCIs to well-known OCIs.** The cluster centroids which are candidate OCIs can be divided into four categories with respect to known OCI's in terms of correlation: very high, high, medium, and low. Cluster centroids that are very highly correlated to known indices represent a rediscovery of well-known indices and serve to validate our approach. This was the focus of [Ste+01]. Cluster centroids that have a high or medium correlation to well-known indices represent alternatives to current indices in that they may potentially be better predictors of land behavior, at least for some regions of the land. Finally, cluster centroids that are not well correlated with known indices may represent potentially new Earth science phenomena.

More specifically, Section 4 is devoted to the presentation of preliminary results that compare candidate OCIs, which are cluster centroids derived from the clustering of Sea Surface Temperature, with well-known indices in terms of their (area-weighted) correlation to temperature on the land. After briefly mentioning the rediscovery of some well-known indices, we present a number of cluster centroids that have high area-weighted correlation to land temperature. While the coverage, (area of the land for which the correlation is high) is often similar to that of well-known OCIs, in many cases, the cluster centroids have higher correlation than the known indices for certain regions of the land.

Finally, although clustering appears to be doing a good job of finding some regions of the ocean that are highly correlated to land behavior, it is reasonable to ask whether we have missed some points on the ocean that might also be good predictors of land

* Area-weighted correlation is the weighed average of the correlation of the OCI with all land points, where weight is based on the land area of the land grid point.

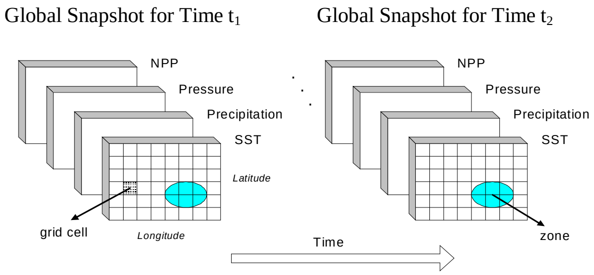

Figure 1: A simplified view of the problem domain.

behavior. To answer this question, we calculated the area-weighted correlation of each ocean grid point and compared the points with high area-weighted correlations to the cluster centroids. What we found was that most, although not all, points that with high area weighted correlation are quite similar to the cluster centroids, and thus, we are not missing many potential OCIs. This analysis is presented in Section 5.

2. Earth Science Data

The Earth science data for our analysis consists of global snapshots of measurement values for a number of variables (e.g., NPP, temperature, pressure and precipitation) collected for all land surfaces or water (see Figure 1). These variable values are either observations from different sensors, e.g., precipitation and sea surface temperature (SST), or the result of model predictions, e.g., NPP from the CASA model, 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., NPP, which is available at a resolution of 0.5° x 0.5°, and sea surface temperature, which is available for a 1° x 1° grid.

Using variables derived from sensor observations, Earth scientists have developed standard ocean climate indices. These indices are useful because 1) they can distill climate variability at a regional or global scale into a single time series, 2) they are related to well-known climate phenomena such as El Nino, and 3) they are well-accepted by Earth scientists. For example, various El Nino related indices, such as NINO 1+2 and NINO 4, have been established to measure sea surface temperature anomalies across different regions of the Pacific Ocean. Some of the well-known climate indices are shown in Table 1 [IND1, IND2]. Figure 2 shows the time series for the SOI index. Note that the dip in 1982 and 1983 corresponds to a severe El Nino event.

For completeness, we mention that there are significant issues related to the spatial and temporal nature of Earth science data: the “proper” measure of similarity between time series, the seasonality of the data, and the presence of spatial and temporal autocorrelation (i.e., measured values that are close in time and space tend to be highly correlated, or similar). For our similarity measure, we use Pearson’s correlation coefficient [Lin98], which ranges between –1 (perfect negative linear correlation) and 1 (perfect positive linear correlation), with a value of 0 indicating no linear correlation. To handle the issues of seasonality and temporal autocorrelation, we pre-process the data to remove seasonality. In particular, we use 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. For further details, we refer the reader to [Ste+01] or [Tan+01].

Climate Index	Description
SOI	Measures the sea level pressure (SLP) anomalies between Darwin and Tahiti
NAO	Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
NINO 1+2	Sea surface temperature anomalies in the region bounded by 80°W-90°W and 0°-10°S
NINO 4	Sea surface temperature anomalies in the region bounded by 150°W-160°W and 5°S-5°N
NP	Area-weighted sea level pressure over the region 30N-65N, 160E-140W

Table 1: Description of well-known climate indices.

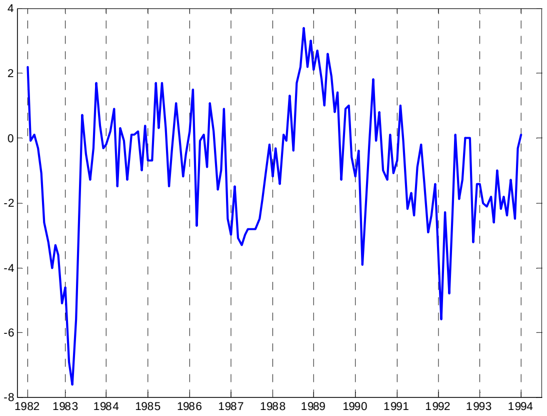

Figure 2: Southern Oscillation Index (SOI)

3. Eigenvalue Approaches vs. Clustering

3.1 Finding Strong Spatial or Temporal Patterns in Earth Science Data Using SVD Analysis

Given a data matrix, whose rows consist of time series from various points on the globe, we would like to discover the strongest temporal or spatial patterns in the data. Earth Scientists have profitably used Empirical Orthogonal Functions (EOF), to find spatial patterns, and temporal patterns.

EOF is just another name for a statistical technique known as Principal Components Analysis (PCA), which, in turn, is equivalent 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 (see Appendix A for a more technical description), 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. (Note that each temporal pattern is a row vector, i.e., a time series, while each spatial pattern is a column vector.)

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 sometimes, 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 in another way, the original data can be approximated as a linear combination of these strongest patterns. Again, see Appendix A for a technical explanation.

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

3.2 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. (Note that the rows of this data have a mean of 0 and thus, and SVD analysis is equivalent to an EOF analysis.) To find the top spatial and temporal patterns is a simple matter using current mathematics or statistics packages. For example, in MATLAB this requires only the following command:

```
[ u s v ] = svds( sparse( z_sst ) , 20 );
```

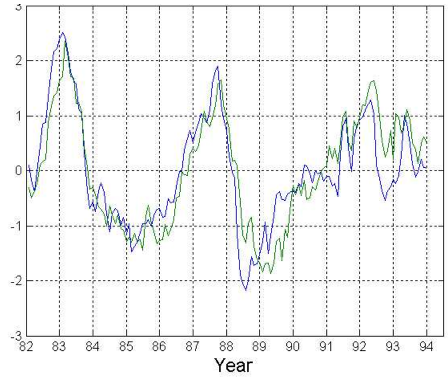

Figure 3: First right singular vector of SST (green) plotted against the NINO 3 index (blue).

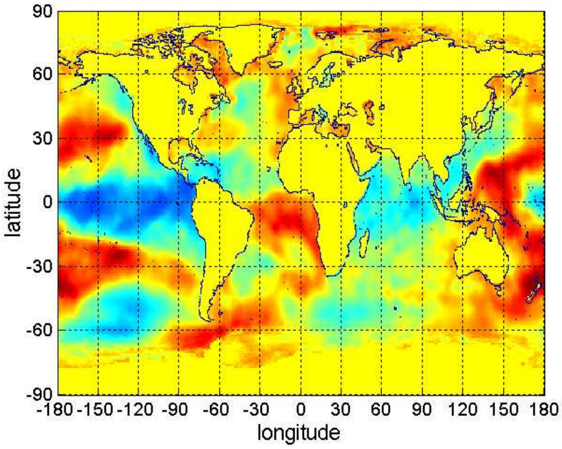

Figure 4: Strongest spatial pattern of SST.

where **z_sst** is the SST data matrix which has been normalized using the montly Z-score. The columns of *u* are the spatial patterns, the diagonal elements of *s* are the singular values, and the columns of *v* are the temporal patterns.

For SST the strongest temporal pattern (the first column of *v*) is highly related to SST as is shown in Figure 3. The correlation of the first right singular vector with Nino 3 is 0.86. The spatial pattern corresponding to the first column of *v* is the first column of *u*, and is shown in Figure 4. Note that to map such a spatial pattern it is necessary to keep track of which grid location that corresponds to each row of data (each row of u).

3.3 Limitations of SVD

The performance of SVD in the above example, is impressive, a well-known OCI was discovered straightforwardly. However, There are a number of limitations of SVD analysis, some 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, $1 \leq k \leq$ rank(data matrix) (see Appendix A). Thus, slightly weaker patterns may

[*] Singular values equivalent to the eigenvalues of PCA.

not show up as well. For example, if the seasonality is not removed from the data, at least the first few strongest patterns will be seasonal patterns of different types. While it is true that we remove seasonality from the data as part of the preprocessing step, this phenomenon can still occur once seasonality is removed. In other words, strong patterns mask weaker ones. Of course, clustering is somewhat subject to the same problem, dominace of strong patterns, but we would argue not to the extent of SVD. Traditionally only the first few SVD vectors are regarded as trustworthy while clustering approaches can find many “good” clusters.

Also, the patterns found using SVD, i.e., the singular vectors, are constrained to be orthogonal to each other. (This is another reason that only the first few singular vectors are “reliable.”) While orthogonality may be appealing mathematically, it can also make patterns hard to interpret. Earth scientists have developed an approach to try to address this problem - ‘rotated’ EOFs [ZS98] - but it is somewhat controversial.

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.

A more subtle limitation of SVD is that the spatial pattern (an EOF) found in a data set corresponds roughly to the pattern of correlation that you would see if you computed the correlation of each original data point with the corresponding right singular vector. However, in the Earth science domain, there is often a lag associated with the impact of various phenomena. SVD analysis cannot take into account any such lag.

Finally, efficiency can be a concern for the SVD approach, although, even with our biggest current data set consisting of ~100,000 time series of length 500, SVD computation times are still acceptable.

3.4 Clustering

Clustering, on the other hand, does not suffer from the limitations mentioned above. In this case the patterns are the cluster centroids. They are not constrained to be orthogonal and are easy to interpret, i.e., they are the representative point of a collection of relatively cohesive points. In our case, they are the representative time series of relatively cohesive sets of time series. Furthermore, in general, clustering does not have any limitations with respect to detecting patterns that lie in one subspace or in overlapping subspaces. Finally, while clustering finds the strong patterns, it can also do a good job on finding weaker patterns.

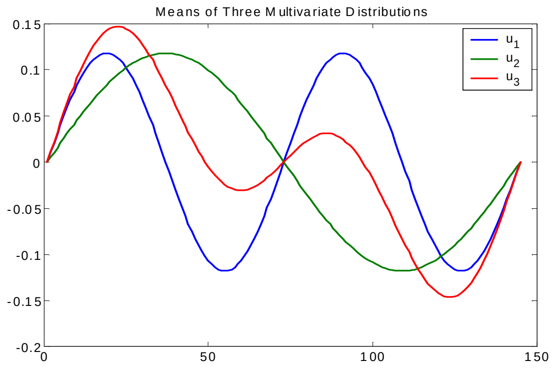

Figure 5: Three Mean Vectors.

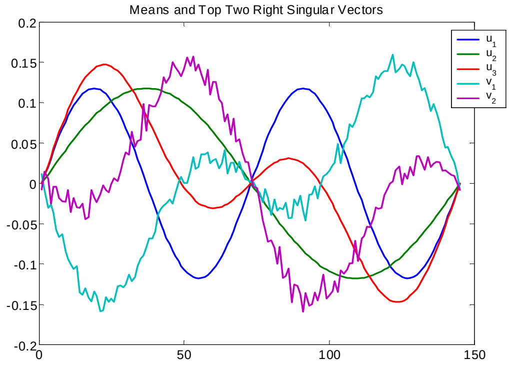

Figure 6: Means and Top Two Right Singular Vectors.

Of course, clustering has a number of limitations of its own. In particular, it is necessary to choose a clustering algorithm that is suitable for the data and the task at hand, and to choose the clustering parameters appropriately. However, there are also choices involved with the use of SVD. In particular Earth Scientists tend to select the areas to which this analysis is applied.

3.5 Another Example

To illustrate these ideas, consider a simple example. Assume that we have three sets of multivariate normal data of size 100 each and dimension 144, which are distributed as **N**($\mathbf{u}_1$, Σ_1), **N**($\mathbf{u}_2$, Σ_2), and **N**($\mathbf{u}_3$, Σ_3), where

$\mathbf{u}_1 \perp \mathbf{u}_2$, and are *sin(2t)* and *sin(t)*, respectively
$\mathbf{u}_3 = \mathbf{u}_1 + \mathbf{u}_2$, normalized to have unit L_2 norm
$\cos(\mathbf{u}_1, \mathbf{u}_2) = 0$,
$\cos(\mathbf{u}_1, \mathbf{u}_3) = 0.7071$,
$\cos(\mathbf{u}_2, \mathbf{u}_3) = 0.7071$
$\Sigma_1 = \Sigma_2 = \Sigma_3 = 0.01*\mathbf{I}$,
$\mathbf{I}$ = the identity matrix

Figure 5 shows a plot of $\mathbf{u}_{1,}$ $\mathbf{u}_{2,}$ and $\mathbf{u}_{3.}$

If we perform an SVD analysis on the combined data – after removing the mean to make the results equivalent to a PCA analysis – then we get the following singular values:

14.3114 10.2865 2.9116 2.8749 … 0.6273 0.5919 0.5658 0.0000

Clearly only the first two singular values are significant. Table 2 shows the cosine similarity for the first two right singular vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, vs. the three means. While $\mathbf{v}_1$ reflects $\mathbf{u}_3$ and the third cluster, $\mathbf{v}_2$ mixes up clusters 1 and 2. Figure 6, which plots the three means and first two right singular vectors, also indicates this.

	$\mathbf{v}_1$	$\mathbf{v}_2$
$\mathbf{u}_1$	-0.7203	-0.6868
$\mathbf{u}_2$	-0.6880	0.7178
$\mathbf{u}_3$	-0.9958	0.0219

Table 2: Cosine similarity of means and right singular vectors.

However, if K-means is asked to find three clusters, it will, with a simple K-means algorithm and on the first try, find exactly the right clusters, i.e., every cluster consists of points generated from the same mean, indicating that the grouping in the data is quite strong.

4. Discovery and Analysis of OCIs

4.1 Background

If we apply a clustering algorithm [JD88, KR90] to cluster the temperature time series associated with points on the ocean, we obtain clusters that represent ocean regions with relatively homogeneous behavior. The centroids of these clusters are time series that summarize the behavior of these ocean areas, and thus, represent potential OCIs. Consequently, clustering is an initial and key step in using data mining for the discovery of OCIs.

The SNN clustering approach produces high quality clusters, which are almost always geographically contiguous, and automatically discovers the “correct” number of clusters. Because of space considerations, we omit a detailed description of the SNN algorithm and refer the reader to [ESK01]. We used SNN clustering on sea surface temperature (SST) over the time period from 1958 to 1998. Note that the monthly Z score transformation has been used, thus removing seasonality and putting the focus on the anomalies in the time series. Figure 7 shows the ocean clusters for the long-term data.

One approach to evaluating potential ocean climate indices is to look at the *area-weighted correlation* of the time series representing an a candidate OCI with the time series associated with land points, e.g., temperature or precipitation time series. A higher value indicates a stronger impact on the land.

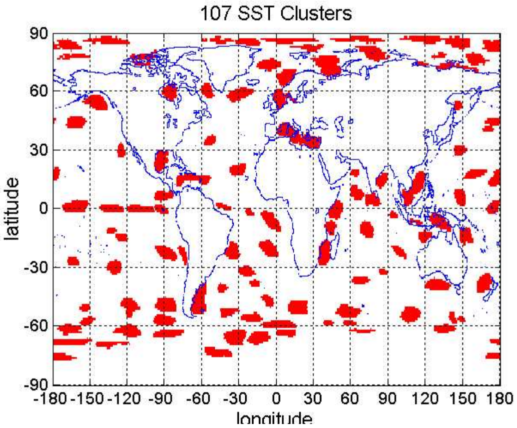

Figure 7: 107 SST clusters.

The details of computing the area-weighted correlation are as follows. We first compute the correlation of the time series of the candidate OCI 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 used absolute correlation because we are interested in the strength of the connections between ocean 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 OCI), but is normally much lower. The minimum value is 0.

One variation of this procedure is to eliminate any correlation whose magnitudes are below a certain threshold. The idea is to see if looking only at stronger correlations produces different results and to eliminate noise. Another variation is to compute the area-weighted correlation for various shifts, i.e., the correlation between each OCI and each land point is calculated using the maximum shifted correlation, where the possible shifts range from 0 to 6 months. While other variations are possible, we observed similar results and will not discuss them further here. Indeed, we mostly focus on results using the maximum shifted correlation and no threshold.

Regardless of exactly which version of area-weighted correlation is used, we need a baseline to compare against. For this, we can use already existing OCIs. Thus, if a candidate shows an area-weighted correlation that is roughly as good or better than that of an existing OCI, it might be worth investigating. If the area-weighted correlation is lower, then either the candidate OCI is not a good index or its connection

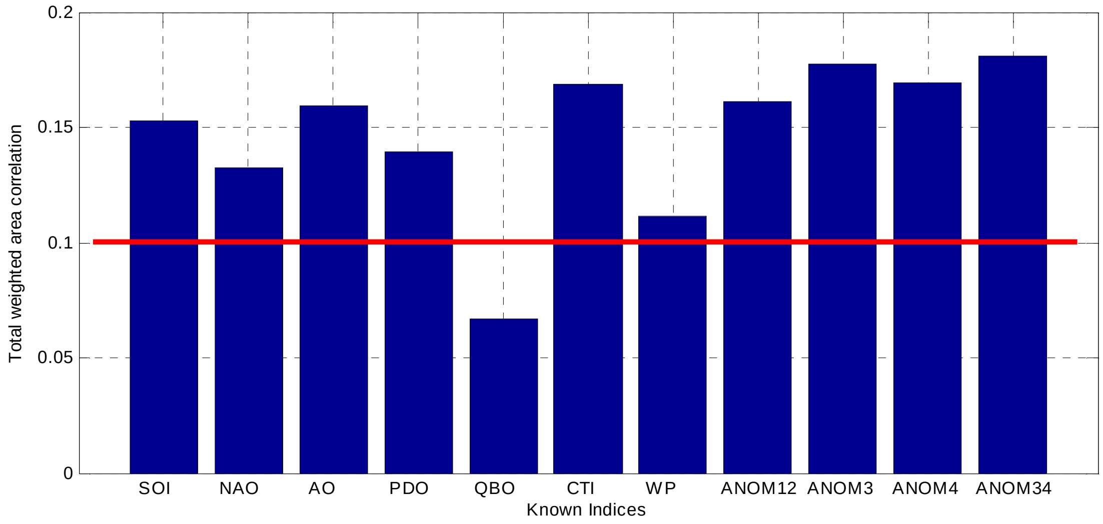

Figure 8: Area weighted correlation of well-known indices.

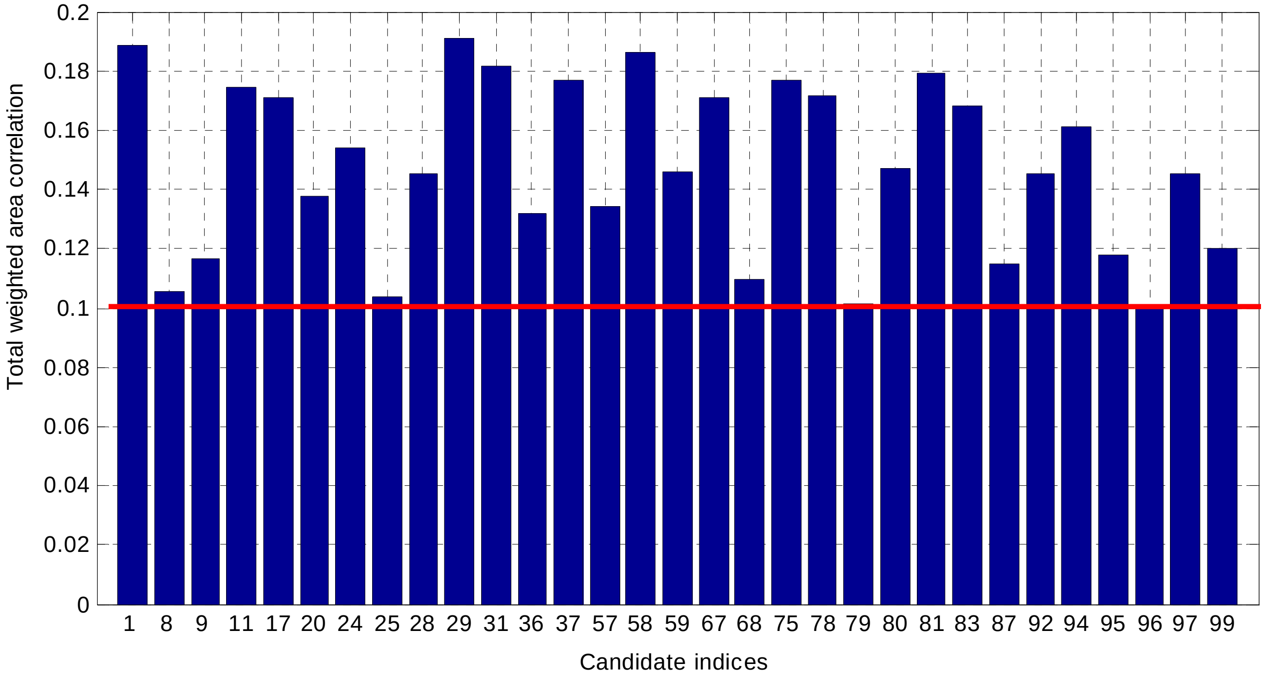

Figure 9: Area weighted correlation of highest SST cluster centroids.

with the land may be very localized.

4.2 Baseline: Area-weighted Correlation of Well-Known Climate Indices

Thus, to get our baseline we computed the total are area-weighted correlations for 11 well-known OCIs. These results are shown in Figure 8. The ENSO indices have total area-weighted correlation above 0.15, while the NAO and AO indices have area-weighted correlation greater than 0.13. Other indices have much lower area-weighted correlation.

The area-weighted correlations were also computed for the SST cluster centroids. These results are shown in Figure 9. Only the cluster centroids that are close to the area-weighted coverage limit of interest, 0.11, are shown.

The following index labels are for the x-axis of the plots. For more information see Table 1 or [IND1, IND2]. Note that 1, 6, and 8-11 are all El Nino related indices.

1. SOI (Southern Oscillation Index)
2. NAO (North Atlantic Oscillation)
3. AO (Artic Oscillation)
4. PDO (Pacific Decadel Oscillation)
5. QBO (Quasi-Biennial Oscillation Index)
6 .CTI (Cold Tongue Index)
7. WP (Western Pacific)
8. ANOM12 (Normalized version of NINO12)
9. ANOM3 (Normalized version of ANOM3)
10. ANOM4 (Normalized version of NINO4)
11. ANOM34 (Normalized version NINO34)

4.3 Weighed Area Correlation of Cluster Centroids

For the analysis we divided the cluster centroids with high area-weighted correlation (> 0.1) into 4 groups depending on the correlation of the cluster centroids to know OCIs.

1. G0: correlation to known OCIs ≥ 0.8.
2. G1: correlation to known OCIs between 0.4 and 0.8.
3. G2: correlation to known OCIs between 0.25 and 0.4.
4. G3: correlation to known OCIs ≤ 0.25.

Note that we studied the area-weighted correlation for each group separately. For each group, we kept only clusters whose area-weighted correlations satisfy all the following conditions: (1) greater than 0.1 (at min correlation = 0), (2) greater than 0.05 (at min corr = 0.1), (3) greater than 0.03 (at min corr = 0.2), (4) greater than 0.02 (at min corr = 0.25), (5) greater than 0.01 (at min corr = 0.3).

Figure 10 shows clusters that reproduce some well-known OCIs. In particular, cluster 54 corresponds to ANOM 1+2, 67 to ANOM 3, 73 to ANOM 3.4, and 75 to ANOM 4.

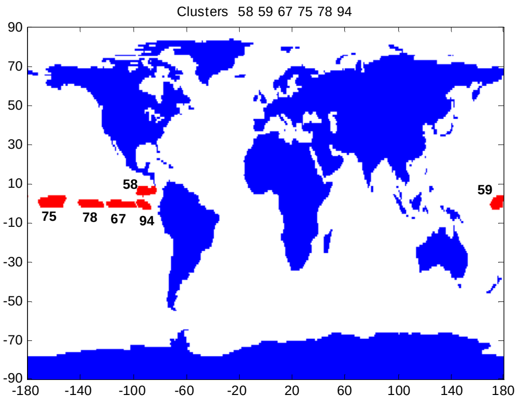

Figure 10: **G0:** Clusters with correlation to known OCIs ≥ 0.8.

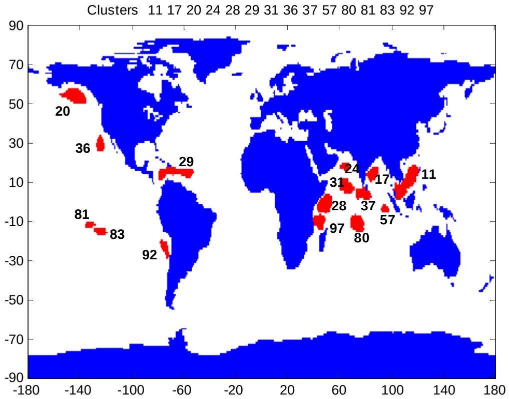

Figure 11: **G1:** Clusters with correlation to known OCIs between 0.4 and 0.8.

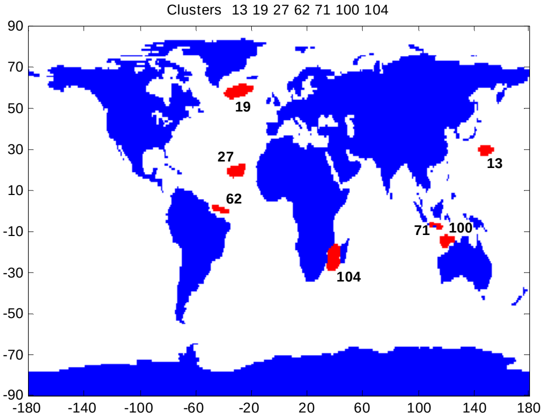

Figure 12: **G2:** Clusters with correlation to known OCIs between 0.25 and 0.4

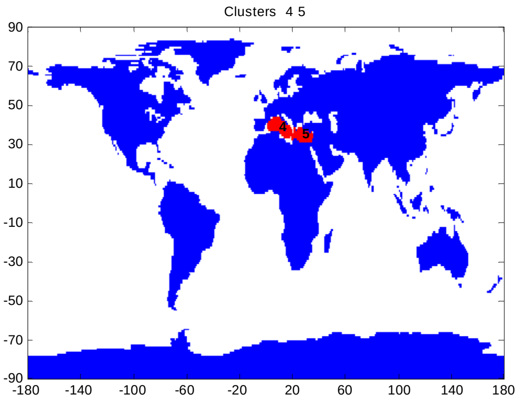

Figure 13: **G3:** Clusters with correlation to known OCIs ≤ 0.25.

4.4 Weighed Area Correlation of Cluster Centroids

While the clusters that are highly and moderately correlated with know OCIs probably capture similar Earth science phenomena, there is still benefit as known OCIs they may still provide some benefits. In particular, some cluster centroids provide better "coverage," i.e., higher correlation, for some areas of the land. This is illustrated in Figures 14 and 15, which, respectively, compare the El Nino OCIs to that of clusters 62 (G2) and 29 (G1). Areas of yellow indicate where the cluster have higher correlation, while areas of blue indicate where the El Nino indices have higher correlation. It is clear that for both these clusters there are areas of the land where the cluster "outperforms" the known OCIs.

5. Completeness Analysis

Finally, although clustering appears to be doing a good job of finding regions of the ocean that are highly correlated to land behavior, it is reasonable to ask whether we have missed some points on the ocean that might also be good predictors of land behavior. Ideally, to answer this question, we should find the area-weighted correlation of each ocean grid point and compare the points with high area-weighted correlation to find our cluster centroids. Hopefully all or most points with high area-weighted correlation will be similar to the cluster centroids.

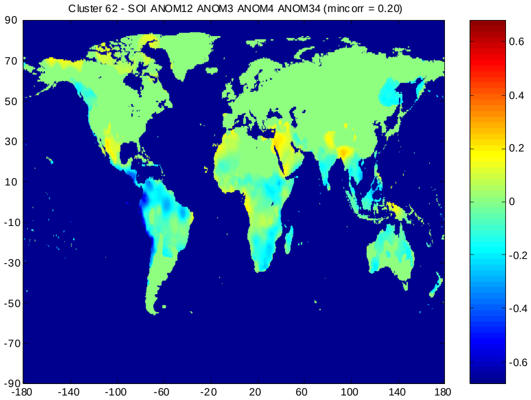

Figure 14: Cluster 29 vs. El Nino Indices

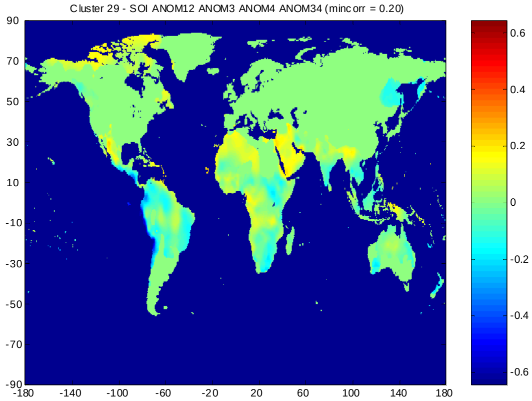

Figure 15: Cluster 29 vs. El Nino Indices

Thus, we calculated the area-weighted correlation of each ocean grid point. Note that this is very computationally intensive calculation. Figure 16 shows the area-weighted correlation (SST vs. land temperature) for all 43,614 ocean grid points. Redder areas indicate points with the strongest relationship to land temperature.

For those points that have an area-weighted correlation greater than 0.14, we found the points that are not similar to the SST cluster centroids for a similarity (correlation) threshold of 0.5, 0.7, and 0.9. These figures are shown below. The threshold of 0.14

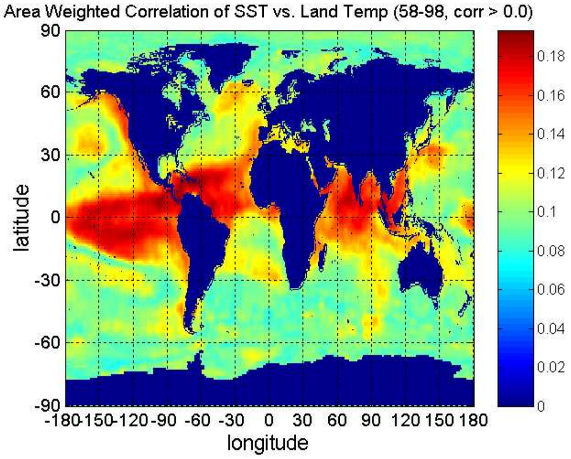

Figure 16: Area-weighted correlation of ocean points vs. land temperature.

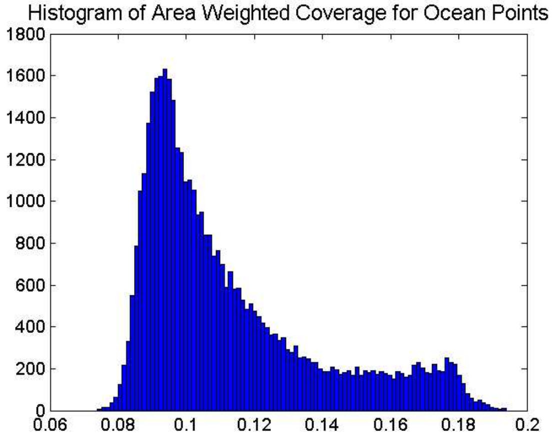

Figure 17: Histogram of area-weighted correlation of ocean points vs. land temperature.

was chosen by looking at the histogram in Figure 17.

While most points are not being "missed", there seems to be a very specific pattern of points that are being missed and it may be worthwhile to investigate the behavior of our SNN clustering algorithm. While we do not want to go into the details of the algorithms here, we show the

6. Conclusions and Future Work

In this paper we have argued that clustering can provide an alternative approach to eigenvalalue based analyses based on PCA or SVD for finding ocean climate indices. To that end we illustrated some of the limitations of eigenvalue analysis, i.e., that it only reliably finds a few of the strongest patterns and that these patterns are constrained to be orthogonal to one another. Clustering does not suffer from these limitations, although it of course, has its own issues.

We then illustrated the use of clustering by showing how clusters of SST could be found and evaluated with respect to their impact on land temperature. To measure that impact a new measure, area-weighted correlation was introduced. We investigated those clusters with relatively high area-weighted correlation and divided them into four groups: those that are very highly correlated with well-known indices and represent a rediscovery of such indices and those that are highly, moderately, or poorly correlated with known indices. The indices that are highly or moderately correlated may still represent the same phenomenon as well-known indices, but may provide better predictive power for some land areas. The indices that are poorly correlated may represent new Earth science phenomena.

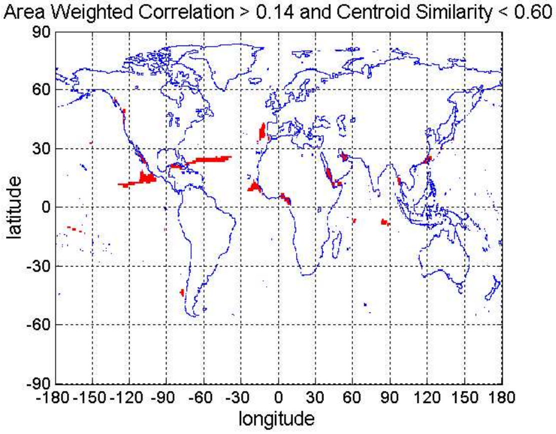

Figure 18: Ocean points with area-weighted correlation > 0.14 and similarity of SST centroids less than 0.60.

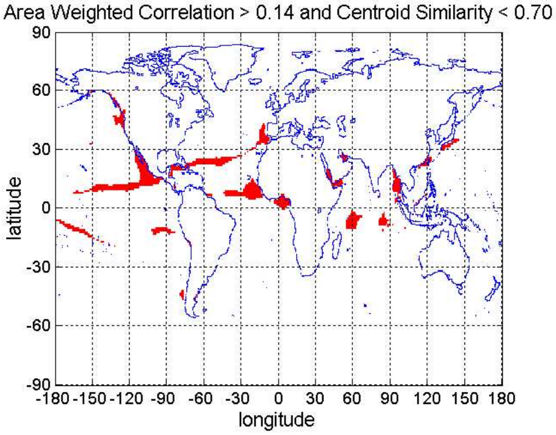

Figure 19: Ocean points with area-weighted correlation > 0.14 and similarity of SST centroids less than 0.70.

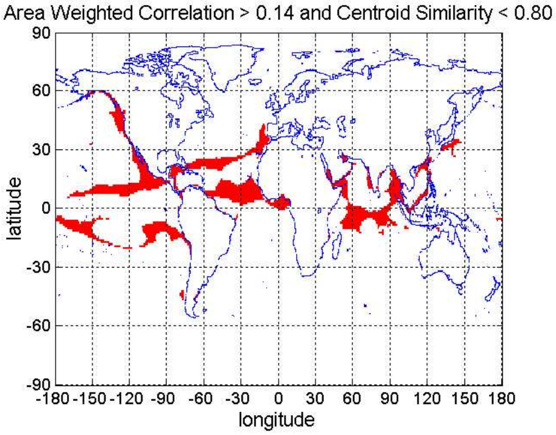

Figure 20: Ocean points with area-weighted correlation > 0.14 and similarity of SST centroids less than 0.80.

Finally, we looked at the relationship between the density of a point on the ocean and its area-weighted correlation. What we found, is that most points that have high area-weighted correlation tend to be well correlated with cluster centroids. However, it does appear as though we may be missing some points of interest and further investigation seems indicated.

In the future, 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.

References

[JD88] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data,* Prentice Hall (1988).

[ESK01] L. Ertöz, M. Steinbach, and V. Kumar, "Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach," Text Mine '01, Workshop on Text Mining, First SIAM International Conference on Data Mining, Chicago, IL, (2001).

[GRS99] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim, (1998*),* "ROCK: A Robust Clustering Algorithm for Categorical Attributes," In Proceedings of the 15th International Conference on Data Engineering, 1999.

[IND1] http://www.cgd.ucar.edu/cas/catalog/climind/

[IND2] http://www.cdc.noaa.gov/USclimate/Correlation/help.html

[KR90] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley and Sons (1990).

[Lin98] B.W. Lindgren, "Statistical Theory", Fourth Edition, Chapman & Hall/CRC (1998).

[NASA] http://earthobservatory.nasa.gov/Library/

[PKB99] C.S. 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 (1999).

[Ste+01] M. Steinbach, P. N. Tan, V. Kumar, C. Potter, S. Klooster, A. Torregrosa, "Clustering Earth Science Data: Goals, Issues and Results", In *Proc. of the Fourth KDD Workshop on Mining Scientific Datasets* (2001).

[Ste+02] Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Steven Klooster, and Christopher Potter, "Data Mining for the Discovery of Ocean Climate Indices," Proc of the Fifth Workshop on Scientific Data Mining at 2nd SIAM International Conference on Data Mining (2002).

[Tan+01] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Steven Klooster, Christopher Potter, Alicia Torregrosa, "Finding Spatio-Termporal Patterns in Earth Science Data: Goals, Issues and Results**,**" Temporal Data Mining Workshop, KDD2001 (2001).

[Tay98] G. H. Taylor, "Impacts of the El Niño/Southern Oscillation on the Pacific Northwest" (1998) http://www.ocs.orst.edu/reports/enso_pnw.html

[SZ98] *Statistical Analysis in Climate Research*, Hans Von Storch And Francis W. Zwiers, Cambridge University Press, 1998

Appendix A. Principal Components Analysis (PCA) and Singular Value Decomposition (SVD)

This is a slightly more technical description of principal components analysis (PCA) and singular value decomposition (SVD) [WSB92]. We focus first on SVD and then briefly describe PCA in terms of SVD. We will illustrate our discussion through the use of SVD approach for removing seasonality from sea surface temperature (SST), where our data matrix is *M*, whose rows consist of the collection of time series that are of interest, i.e., in this case, the matrix rows consist of the sea surface temperature time series for a large number of points on the ocean (~150,000 points). A singular value decomposition expresses an *m* by *n* matrix, *M*, as the sum of simpler rank 1 matrices as follows:

$\boldsymbol{M} = \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{s_i}\vec{\boldsymbol{u}}_{\boldsymbol{i}}\vec{\boldsymbol{v}}_{\boldsymbol{i}}{}'$, where s_i, a scalar, is the i^{th} singular value of *M*, $\vec{u}_i$ is the i^{th} left singular vector, and $\vec{v}_i$ is the i^{th} right singular vector. All singular values beyond the first *r*, where *r* = rank(*M*) are 0 and all left (right) singular vectors are orthogonal to each other and are of unit length. Also, the singular values are in order of decreasing magnitude.

Thus, a matrix can be approximated by omitting some of the terms of the series that correspond to non-zero singular values. Indeed, if *k* terms are retained, then this approximation has rank k and is the best possible approximation as measured by the Frobenius matrix norm. Furthermore, as should be clear from the series formulation of SVD, terms with small magnitudes do not contribute much, and thus, SVD is often used for dimensionality reduction. In some cases the first few singular values are much larger than the remaining ones, and the reduction in dimensionality is very significant.

Furthermore, if a characteristic of the data corresponds to a particular term (singular value), then this characteristic can be removed by eliminating the corresponding term. For example, removing the first term, which corresponds to the largest singular value, removes a constant component from the SST data, i.e., after removing the first term the maximum mean value of any times series from is 0.02. (Before there was a wide distribution of mean values, e.g., many time series in the tropics had means in 20's.) Thus, in this case,

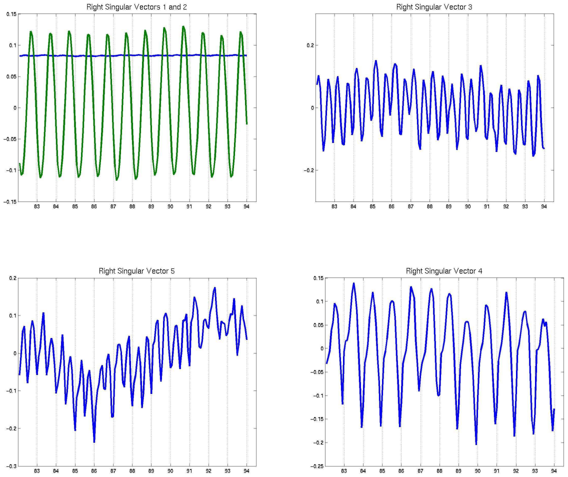

Figure 21: First five right singular values of SST data. (In top left plot, second right singular vector is green.)

removing the first term is roughly equivalent to normalizing each time series to have a mean value of 0.

The nature of each term can be analyzed by looking at the associated right singular vector, which, in this case, can be interpreted as a time series. Figure 21 shows the first five right singular vectors for the SST matrix. From the first plot we see that the 1st and 2nd right singular vectors, correspond, respectively, to a constant and a 12-month seasonal component.

If the first five right singular vectors are removed, then most of the seasonality is removed, and the resulting data is much the same as if it had been processed by using a Fourier transform or the monthly Z score. However, the SVD approach for removing seasonality is more computationally intensive than the other approaches and, the other approaches seem more "direct." But again, this example is just for illustration.

Finally, PCA is essentially SVD except that the mean of the data is removed first. The more traditional computational approach is to find the eigenvalues of the covariance matrix of *M*, where *M* is the data matrix. These eigenvalues are the square of the singular values found by SVD and, the eigenvectors are the right singular vectors of the SVD decomposition. Furthermore, the magnitude of each eigenvalue represents the variance of data captured by the dimension defined the corresponding eigenvalue, i.e., by each pair of singular vectors.