# Data Mining for the Discovery 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
Ocean climate indices (OCIs), which are time series that summarize the behavior of selected areas of the Earth's oceans, are important tools for predicting the effect of the oceans on land climate. In this paper we describe the use of data mining to discover Ocean Climate Indices (OCIs). In particular, we apply a shared nearest neighbor (SNN) clustering algorithm to cluster the pressure and temperature time series associated with points on the ocean, yielding 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. To evaluate cluster centroids for their usefulness as potential OCIs, we must determine which cluster centroids significantly influence the behavior of well-defined land areas. For this task, we use a variety of approaches that analyze the correlation between potential OCIs and the time series (e.g., of temperature or precipitation) which describe the behavior of land points. Based on these approaches, we have identified some cluster centroids that are almost identical to well-known OCIs, e.g., the Southern Oscillation Index (SOI) and the North Atlantic Oscillation (NAO). We also introduce two strategies for validating potential OCIs which do not correspond to well-known (and probably "stronger" OCIs), namely, focusing on the correlation between "extreme" events on the ocean and land and looking for more persistent patterns of correlation.

## Keywords
clustering, shared nearest neighbor, time series, Earth science data, correlation, scientific data mining

## 1. INTRODUCTION
The climate of the Earth's land surface is strongly influenced by the behavior of the Earth's oceans. 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]. To investigate such land-sea connections, Earth scientists often use ocean climate indices (OCIs), which are time series (of sea surface temperature or air pressure) that summarize the behavior of selected areas of the Earth's oceans [IND1].

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 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

* This work was partially supported by NASA grant # NCC 2 1231 and by Army High Performance Computing Research Center contract number DAAH04-95-C-0008. 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.

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.

This paper outlines a data mining approach for the discovery of OCIs which consists of five key steps.

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. (Actually, as we will see later, an OCI can correspond either to a single cluster centroid or to a pair of cluster centroids.) In previous Earth science work [Ste+01], we used K-means clustering, but for work reported here, we use a shared nearest neighbor clustering approach [ESK01]. While K-means produces clusters of "reasonable" quality, SNN clustering is better at finding high quality clusters in noisy data.
2) **Analyze the correlation between the clusters we have found.** Many cluster centroids are highly correlated with one another since regions of the oceans are highly coupled to one another and are part of a global climate system. In particular, we need to understand which clusters belong to groups of clusters, which together represent a single phenomenon. Also, pairs of clusters which are negatively correlated sometimes correspond to a single OCI, and this can be identified by examining pairwise cluster correlations.
3) **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 can be used to explain the behavior of a well-defined region of the land. One way of evaluating OCI impact on the land is to compute the 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. In this fashion, we can determine, for each land point, the cluster centroid with which it is most highly correlated. We can also investigate, for each land point, what the top two centroids are. The clusters or pairs of clusters which strongly affect many land points are potential candidates for climate indices.
4) **Determine if the potential OCI matches a known OCI.** We show that some of the cluster centroids which are the best OCI candidates are almost identical to well-known OCIs, specifically, the Southern Oscillation Index (SOI) and the North Atlantic Oscillation (NAO).
5) **For potential OCIs that are not well-known, conduct further analysis.** Only the strongest indices are likely to be discovered by the techniques in step 3, which rely on analyzing the raw correlation between two time series. In large part this is because the impact of an ocean area on the land may only be significant for extreme events, e.g., an anomalously high temperature in an ocean region may produce anomalously low precipitation in a land area. In such cases, it is better to look at only the anomalous portion of the potential OCI. In addition, we have noticed that OCIs typically show patterns of correlations that persist in time, and thus, comparing the correlation patterns in successive months is useful for OCI validation.

The basic outline of this paper is as follows. Section 2 provides a description of the Earth science data that we use in our subsequent analyses; Section 3 briefly discusses our clustering technique; and Section 4 shows the ocean clusters produced by our SNN clustering approach. Section 5 investigates the inter-cluster correlation for clusters derived from pressure time series, discovering a) groups of related clusters and b) pairs of clusters that are candidate OCIs. Section 6 describes how analysis of the correlation between ocean clusters can be used to identify clusters (or pairs of clusters) that are promising candidates for OCIs. Based on the results of sections 5 and 6, Section 7 shows that some of our clusters (or pairs of clusters) correspond to well-known OCIs such as SOI and NAO. Section 8 describes a methodology for validating "weaker" candidate OCIs, while Section 9 is a conclusion and an indication of future directions.

## 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

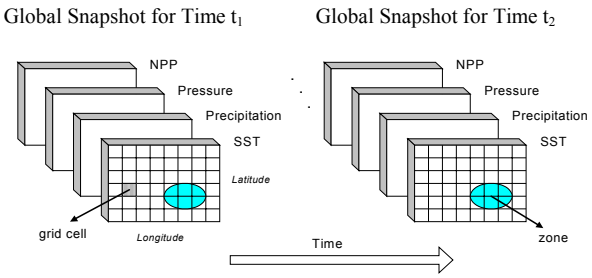


**Figure 1:** A simplified view of the problem domain.

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.

| 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.

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

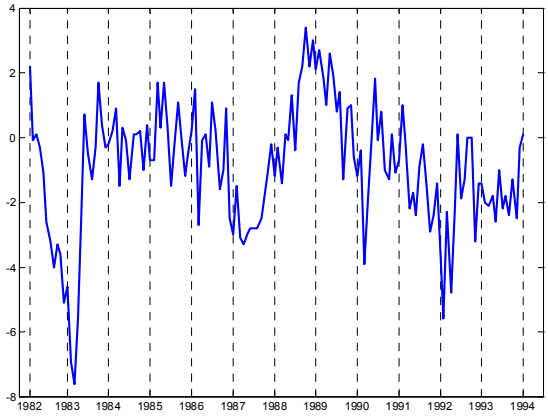


**Figure 2:** Southern Oscillation Index (SOI)

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].

## 3. An SNN Based Clustering Approach

If we apply a clustering algorithm [DJ88, KR90] to cluster the pressure and 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.

For our initial exploration of Earth science data, we employed the widely used K-means clustering algorithm [DJ88], which is simple and efficient. (Because of space considerations we omit a detailed description of the K-means algorithm and refer the reader to [DJ88] for a general description, and to [Stei+01] for a description of K-means in the context of the Earth science data that we are discussing.) While we did find some interesting results using K-means clusters [Ste+01], we decided to switch to a shared nearest neighbor (SNN) clustering approach [ESK01]. K-means tries to cluster all the data, and because of this, cluster quality suffers greatly, particularly if the data is noisy, as with Earth science data. Also, for K-means, the number of clusters needs

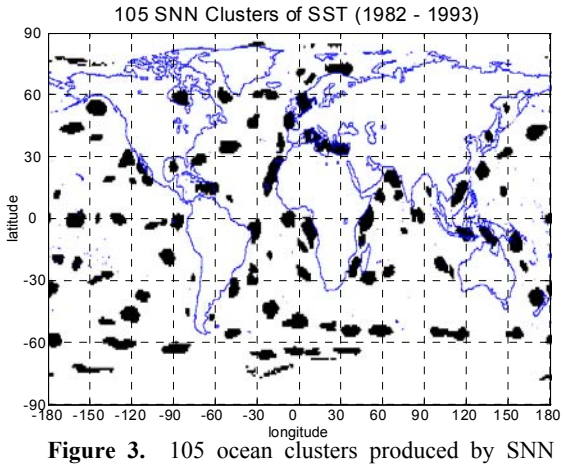


**Figure 3.** 105 ocean clusters produced by SNN clustering of sea surface temperature (1982-1993).

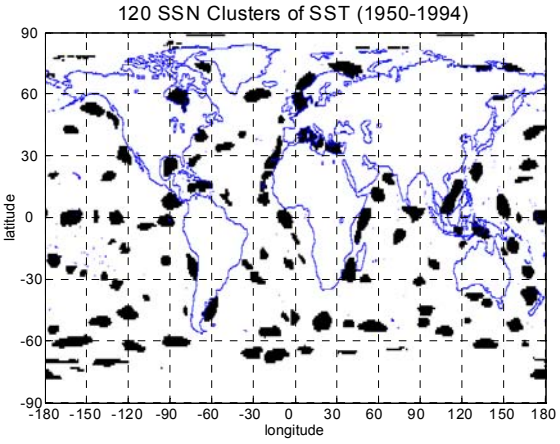


**Figure 4.** 120 ocean clusters produced by SNN clustering of sea surface temperature (1950-1994).

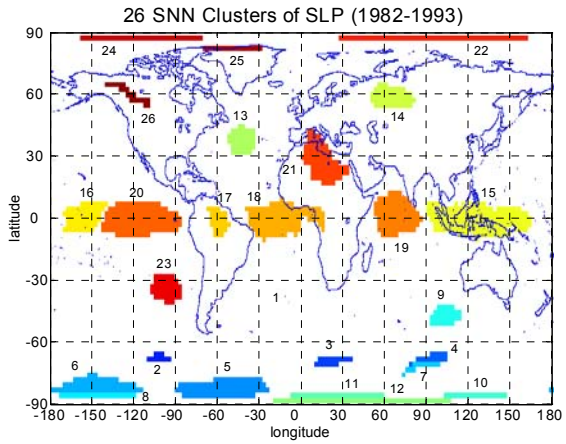


**Figure 5.** 26 ocean clusters produced by SNN clustering of sea level pressure (1982-1993).

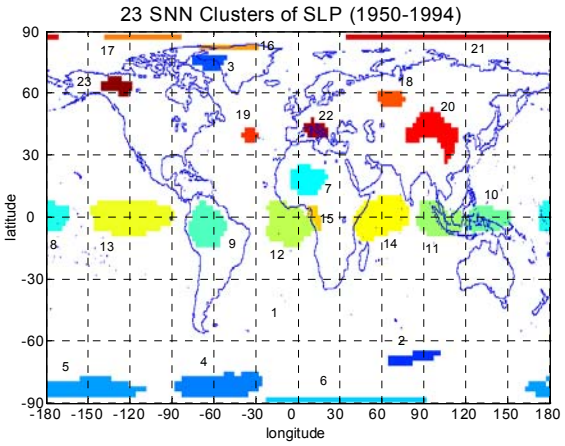


**Figure 6.** 22 ocean clusters produced by SNN clustering of sea level pressure (1950-1994).

to be specified in advance. Furthermore, the clusters produced by K-means sometimes consist of "chunks" which are geographically widely separated. While this can be interesting and useful, for our work in detecting OCIs, we wanted clusters that are geographically contiguous, or nearly so. 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].

## 4. SNN Clustering of Ocean Data

We used SNN clustering on the two sets of data that we have for the ocean, sea level pressure (SLP) and sea surface temperature (SST). For each of these data sets we clustered over two different time periods, from 1950 through 1994, and from 1982 though 1993. The second, shorter time frame was chosen because it matches the time frame for our NPP data, while the longer term data represents the full set of ocean data available to us.

We first present the clusters that resulted from clustering SST. Figure 3 shows the ocean clusters (black regions) found for the short term SST data, while Figure 4 shows the ocean clusters for the long term data. Notice that while there are some differences, e.g., a cluster has disappeared from off the tip of eastern Brazil and the shapes and sizes of some clusters have changed, there are many similarities between the two figures. It is possible that some of the differences in the two sets of clusters are related to climate change. However, we do not pursue that issue here.

Figures 5 and 6 show the ocean and land clusters (colored regions) for the short term and long term SLP data, respectively. (Notice that the pressure data is actually for the entire globe and thus, we have some clusters on the land.) We have numbered these clusters for easy reference since they will be referred

to in the following sections. Please note that the numbering and colors are not the same between Figure 5 and Figure 6. Finally, in order to keep the focus on the oceans, we will mostly ignore land clusters in the subsequent analysis.

Once again, there are some differences between the long and short term data, e.g., clusters 2 and 23 from the short term figures are absent from the set of long term clusters. However, there are also many similarities, and the clusters that we will discuss in detail are present in both the short and long term data.

In the following, we will concentrate on the SLP clusters since there are fewer, and thus, it is easier to deal with them.

## 5. Cluster Correlations

In this section we consider only the SLP clusters for the short term data (1982-1993). Our goal is to identify groups of related clusters and to identify pairs of clusters that are negatively correlated. Identifying groups of related clusters is important when evaluating a cluster centroid as a potential OCI because some ocean phenomena, e.g., El Nino, may involve several different, but related, ocean areas. Since pressure differences are important in weather and climate, negative correlations between pairs of clusters are also important, and indeed, some well-known OCIs that are based on pressure are defined as the difference of two pressure time series.

Figure 7 shows the correlation matrix for the 26 clusters found by clustering the sea level pressure for the time period 1982-1983. First, notice that the clusters fall into three groups of clusters whose members are relatively highly (negatively or positively) correlated to each other. The largest group is clusters 3-12, which are clusters near Antarctica.

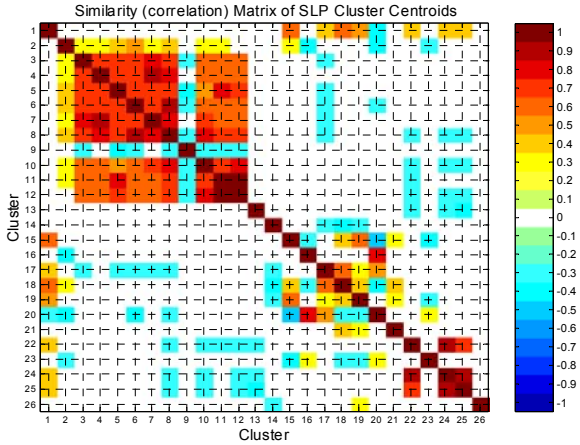


**Figure 7.** Pairwise correlation of SLP cluster centroids. (1982-1993). Correlations with absolute value < 0.2 are been omitted for

Similarly, the Artic clusters 22, 24, and 25 also form a group. Finally, the equatorial clusters, 15 through 20, show a noticeable pattern of correlation, although this pattern is not as strong, or consistent as with the polar groups.

Since some OCIs, e.g., SOI and NAO, are defined by the pressure differences between two points on the Earth, Figure 7 can also be used to identify cluster pairs that correspond to potential OCIs. For example, the negative correlation between clusters 15 and 20 is clear, as is the negative correlation of clusters 13 and 25. We will investigate these two pairs of clusters and their relationship to known OCIs (see Section 8) after we first investigate another technique for identifying potential OCIs. We plan to investigate the other pairs of negatively correlated clusters, e.g., 15 and 23, but will not discuss them further in this paper.

## 6. Correlation of Ocean Clusters to Land Points

As mentioned previously, a cluster centroid (or pair of centroids) is an interesting OCI only if it strongly influences well-defined regions on the land. Our first approach for determining which clusters are most influential counts the number of land points for which a given cluster centroid is the most highly correlated centroid. Note that the maximum correlation between time series may occur when the time series are shifted, and thus, we take the maximum value of the correlations found by considering shifts between 0 and 6 months.

There are a number of different land variables that can be used for this investigation. To cut down on the number of figures displayed, we use only NPP, temperature, and precipitation. For these variables, respectively, figures 8, 9, and 10, show the number of land points for which each SLP cluster centroid is the most highly correlated centroid. Notice that results are omitted for SLP clusters that lie entirely on the land and that the y-scale, the number of points, is different for each plot.

Clusters 15 and 20 have the largest impact on land points in terms of NPP, and are also among the highest for precipitation. Cluster 15 has, by far, the highest impact for temperature, and although 20 is not nearly as influential with respect to temperature, it is higher than average. Clusters 13 and 25 are also among the clusters showing a widespread impact on the land.

For land points, it is also possible to compute the pairs of clusters that occur most frequently as the two most highly correlated clusters. Figure 11, shows the number of land points (on a log base 10 scale) for which each pair of pressure clusters are the two most

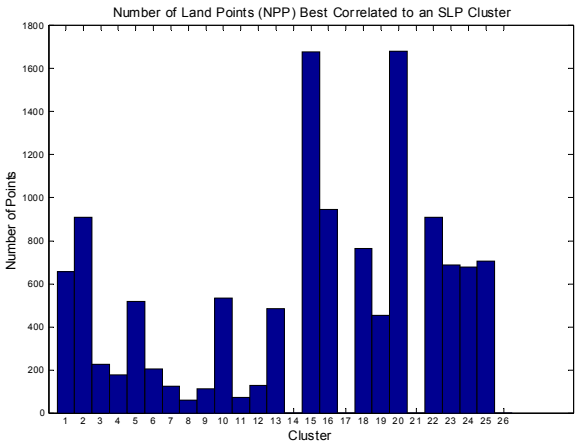


**Figure 8:** Number of land points best correlated to an SLP cluster for NPP.

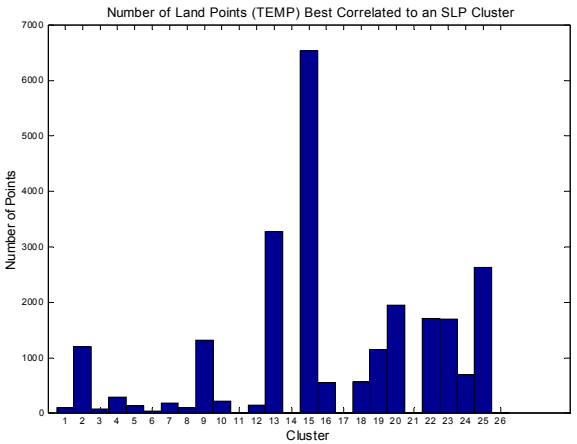


**Figure 9:** Number of land points best correlated to an SLP cluster for temperature.

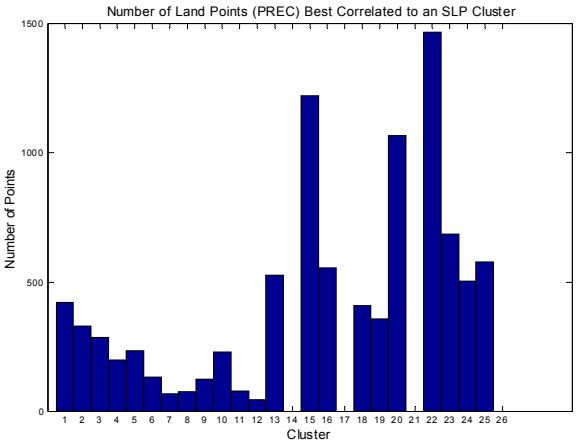


**Figure 10:** Number of land points best correlated to an SLP cluster for precipitation.

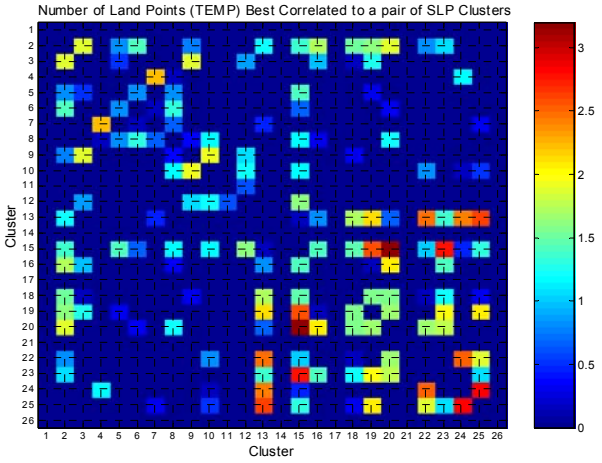


**Figure 11:** Number of land points best correlated for pairs of SLP clusters for Temperature.

correlated clusters with respect to land temperature. Clusters 15 and 20 have a strong co-occurrence relationship for all the variables. Clusters 13 and 25 also have a noticeable co-occurrence relationship for all variables, although it is weaker than that of clusters 15 and 20. From figure 11, and from similar plots (for other land variables) which are not shown, we can identify other pairs of clusters that may possibly be new climate indices: (2, 20), (15, 23), and (13, 22). Notice that in all cases, these pairs of clusters are negatively correlated. In the next section, we show that two of these pairs of clusters, (20, 15) and (13, 25), correspond to well known OCIs.

## 7. Replicating Current Climate Indices

There are two basic types of ocean climate indices: OCIs based on pressure and OCIs based on temperature. The OCIs based on pressure are more complicated, since they are often defined as the difference of the anomalous pressure readings at two different locations on the Earth's surface. For example, as mentioned in Table 1, the Southern Oscillation Index (SOI) measures the sea level pressure anomalies between Darwin and Tahiti, while the North Atlantic Oscillation (NAO) measures normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland.

If our approach for discovering ocean climate indices via clustering works, then we should be able to discover some of the well known indices. Thus, in this section we use our approach to show that there are pairs of clusters (15 and 20, 13 and 25 in Figure 5) which correspond to well-known OCIs (SOI and NAO) created by Earth scientists. We also show that some SST clusters correspond to well-known OCIs which are based on temperature.

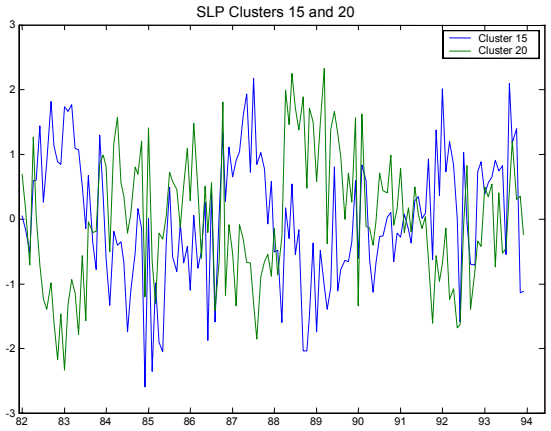


**Figure 12.** Centroids of SLP clusters 15 (near Darwin, Australia) and 20 (near Tahiti) 1982-1993.

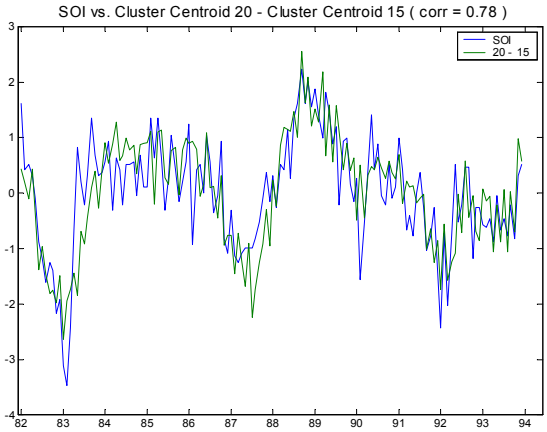


**Figure 13.** Difference of SLP cluster centroids 20 and 15 versus the SOI index. (1982-1993)

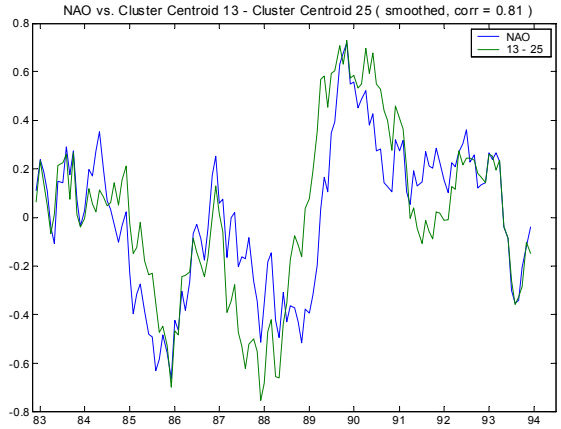


**Figure 14.** Smoothed difference of SLP cluster centroids 13 and 25 versus NAO. (1982-1993)

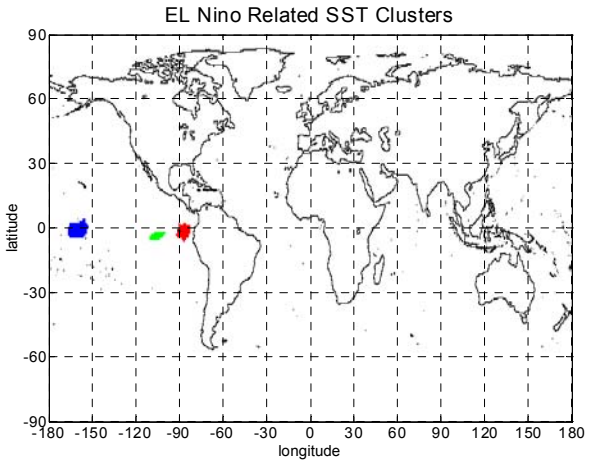


**Figure 15:** SNN clusters of SST that are highly correlated with El Nino indices.

For the 1982-1993 SLP data, we can reproduce the SOI index by taking the difference of the centroids of clusters 20 (near Tahiti) and 15 (near Darwin, Australia). Figure 12 shows, the plot of SLP cluster 15 versus SLP cluster 20. These two time series are very similar, but of opposite phase. Figure 13 shows time series for cluster centroid 20 minus cluster centroid 15 versus the SOI index.. The degree of correlation, 0.78, is statistically very significant (>> 0.01) and the visual match is striking

We can perform the same sort of analysis for NAO. When we do, we find that for short term SLP data, the difference of clusters 13 and 25 is highly correlated with NAO. These clusters correspond, respectively, to clusters near the Azores and Greenland. Figure 14 shows the smoothed version (12-month moving average) of NAO versus cluster centroid 13 minus cluster centroid 25. (NAO is very spiky and an un-smoothed figure is hard to evaluate.) Coincidentally, the correlation for the un-smoothed time series (not shown) is 0.81, the same as for the smoothed series.

We obtain similar results if we use the long terms SLP, from Figure 6. Clusters 13 and 10 correspond to SOI with a correlation of 0.77, while clusters 19 and 16 correspond to NAO with a correlation of 0.76.

To conclude this section, we provide Figure 15, which shows three SST clusters – red, green, and blue (from the 1982-1993 SST data, Figure 3), which are highly correlated ( ≈ 0.94 for all cases) with the El Nino indices. In particular, Nino 1+2 corresponds to the red cluster, Nino 3 to the green, Nino 3+4 to the red, and Nino 4 also to the red. These clusters

correspond well with the El Nino region definitions given in Table 2 [IND1].

| Niño Region | Range Longitude | Range Latitude |
|---|---|---|
| 1+2 | 90°W-80°W | 10°S-0° |
| 3 | 150°W-90°W | 5°S-5°N |
| 3.4 | 170°W-120°W | 5°S-5°N |
| 4 | 160°E-150°W | 5°S-5°N |

**Table 2:** El Nino Regions

## 8. Validation of Candidate OCIs

Some of the cluster centroids discovered using the previous approaches do not correspond to known OCIs, but are potential candidates for new OCIs. Earth scientists are more interested in candidates that exhibit strong correlation with some of the land climate variables. However, quite often, a straightforward correlation between a candidate OCI and the time series of the land climate variable yields poor results for the following reasons:

a) **The impact of an OCI on the land climate variable is often more pronounced at its extreme (high and low) values as compared to its moderate values**. For example, Figure 16 shows the correlation between SOI and precipitation in the United States between January, 1958 and December, 1994. We can divide the SOI time series into 2 disjoint segments: one that corresponds to the months for which the value of SOI is anomalously high or low (this is called the HI & LO series) and another that corresponds to the months for which SOI is moderate (this is known as the NULL series). The results of this figure suggest that much of the correlation between SOI and precipitation can be explained by the HI & LO series (figures in the second column), and there is very little correlation

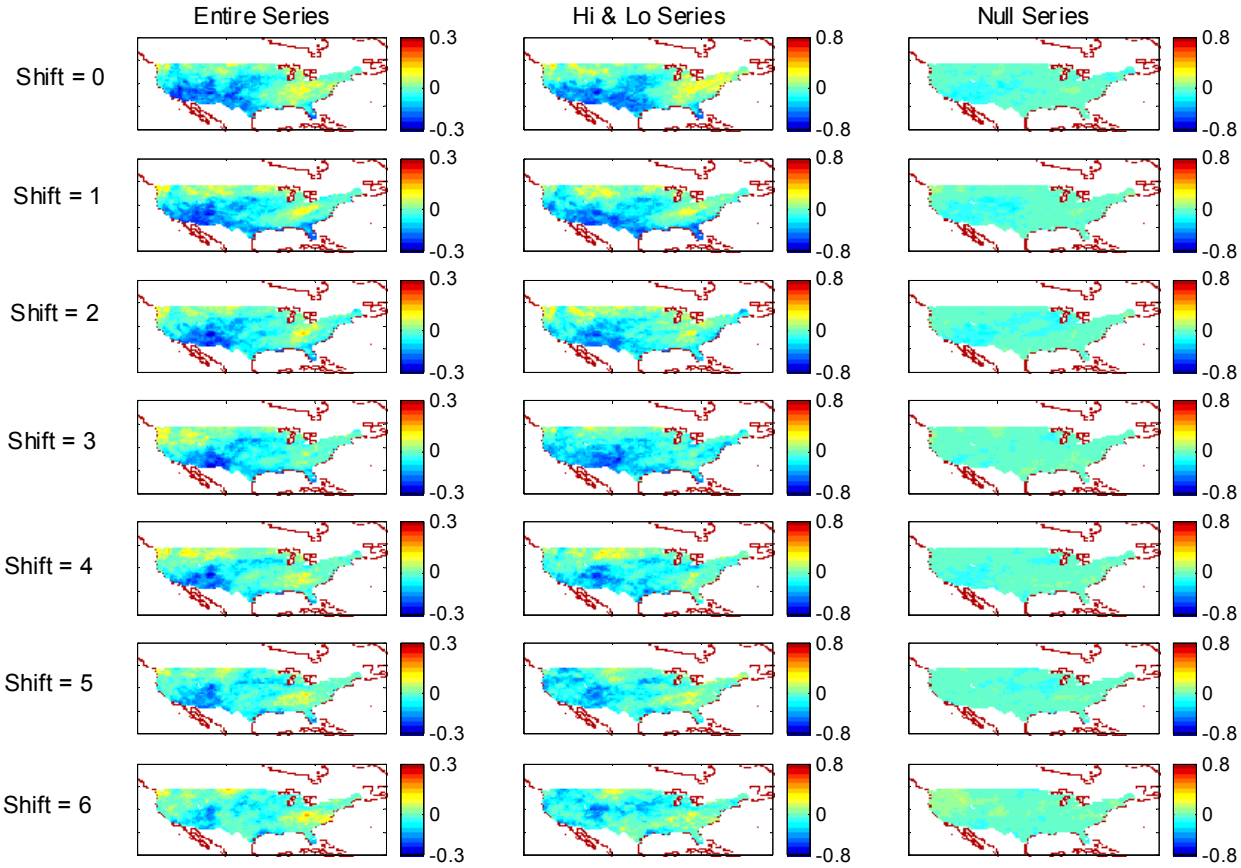


**Figure 16:** Correlation between SOI and precipitation in the United States (from Jan 1958 – Dec 1994). The first column corresponds to correlation for the entire SOI time series. The second column corresponds to the anomalously high/low SOI segment ($Z \geq 1.5$ or $Z \leq -1.5$) while the third column corresponds to the moderate SOI segment ($-1 \leq Z \leq 1$), where Z is the standardized value of SOI.

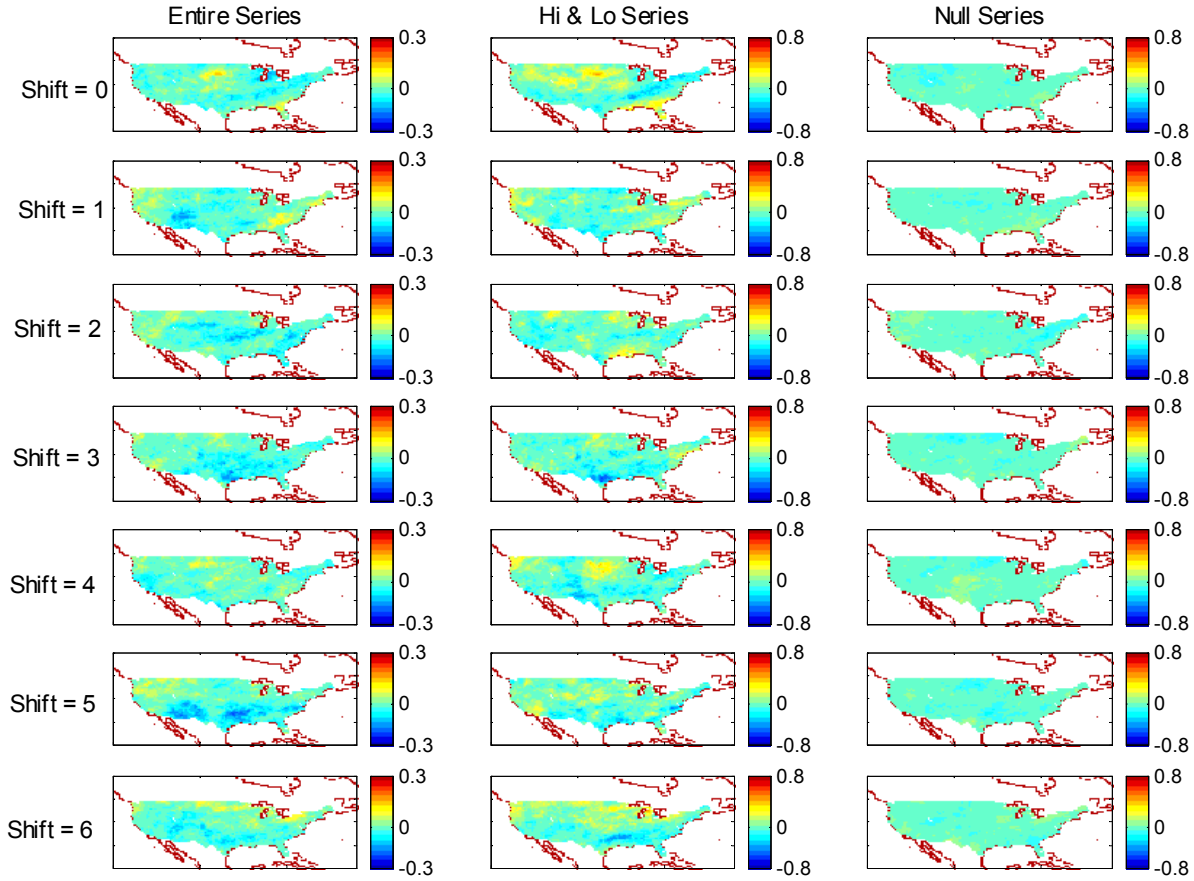


**Figure 17:** Correlation between random noise and precipitation in the United States.

between precipitation and the NULL series (figures in the third column). Furthermore, the correlation values computed using the HI & LO segment are significantly higher than the correlation values computed using the entire time series**.**

b) **The impact of OCI on the land climate variable may not be immediate.** As a result, the direct correlation between them could be poor unless time shifts are taken into account. The rows in Figure 16 illustrate the values of the correlation between SOI and precipitation for various lags of the precipitation time series.

Figure 17 shows a similar plot for correlation between precipitation in the United States and a randomly generated time series. The following results are observed upon comparing figures 16 and 17:

a) The magnitude of the correlation is weaker for random noise compared to the SOI time series, especially for the HI & LO segment of the series.

b) The land region that attains the highest correlation at different shifts appears to be quite similar for the SOI time series but is significantly different for the random noise time series.

These results suggest that one way to identify a reliable OCI is to look for candidates that have a strong correlation, especially for extreme events, with some land climate variables. However, another requirement is that their correlation maps should be relatively stable, i.e., should not change dramatically when we shift the time series of the land climate variable. To demonstrate this point, we have computed an overall similarity measure between the correlation maps (at consecutive shifts) of a given OCI. For a fixed region, let $M_i$ denote the correlation map between an OCI and the time series of a land variable (shifted by $i$ months). Also, let $corr(M_i, M_{i+1})$ denote the correlation between two consecutive maps $M_i$ and $M_{i+1}$. Then the overall similarity measure, $S$, between the shifted correlation maps is:

$$S = \frac{1}{p}\sum_{i=0}^{p-1} corr(M_i, M_{i+1})$$

where $p$ is the maximum shift. Figure 18 shows a histogram of $S$ for the shifted correlation maps of 1000 randomly generated time series with the United States precipitation time series. About 95% of the values lie between –0.2 and 0.25. Figure 19 shows the corresponding values of $S$ for several known OCIs.

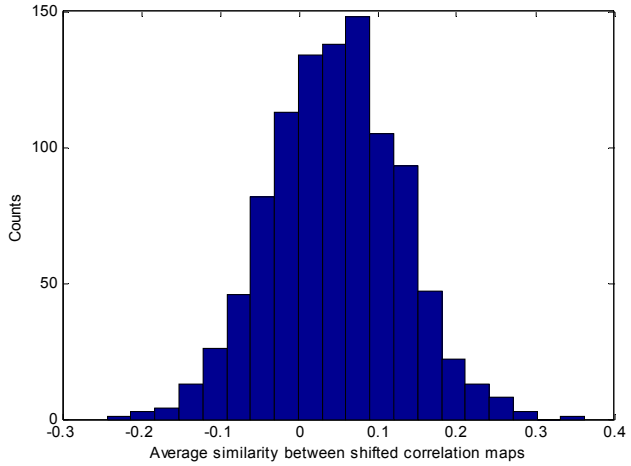


**Figure 18:** Average similarity of shifted correlation maps for 1000 randomly generated time series.

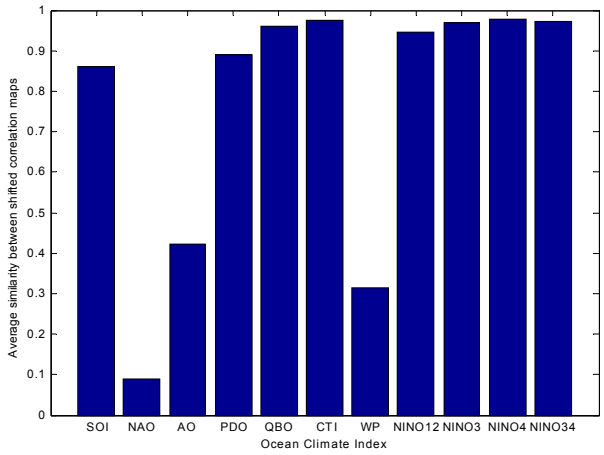


**Figure 19:** Average similarity of shifted correlation maps for various OCIs.

Apart from NAO and WP, many of the known OCIs have higher *S* values than those for random noise. The value of *S* for NAO could be poor because the NAO time series fluctuates rapidly. Furthermore, we have only examined precipitation in the United States. If we consider the precipitation for the entire globe, the *S* value for NAO increases to 0.2, while the range of *S* values for the noise time series is between –0.1 and 0.25. Thus, more study is needed to distinguish NAO-like candidate OCIs from random noise.

## 9. Conclusion and Future Work

In this paper we described the use of data mining to discover ocean climate indices. In particular, we illustrated how SNN clustering can be used to find ocean clusters when each point in the ocean is described either by a temperature or a pressure time series. The centroids of these clusters are time series that summarize the behavior of these areas, and thus, represent potential ocean climate indices. We also described some ways of evaluating which cluster centroids might be good candidates for climate indices, i.e., by looking at the correlation matrix of the cluster centroids or looking at the number of land points which have a cluster or a pair of clusters as their top centroid(s).

For SLP and SST ocean clusters, we then showed that some of the promising ocean clusters (or pairs of ocean clusters in the case of pressure) correspond to well-known ocean indices. However, for unknown and potentially "weaker" OCIs, we need additional strategies for validating whether the potential OCI is really useful, and we discussed two approaches: focusing on extreme values of a potential OCI and looking for correlation patterns that persist over several months.

A task for future research is to use regression or other statistical models to quantitatively evaluate the effect of multiple OCI's on land points. A key issue here is the lag between an OCI and its effect on or correlation with a land area. It is not clear how to perform multiple regression for thousands of land points in cases where there are different "best" shifts for different OCIs and different land points.

## References


DJ88] R. C. Dubes and A. K. Jain, *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).

[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," Submitted to KDD 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