Discovering Dynamic Dipoles in Climate Data

Jaya Kawale † Michael Steinbach † Vipin Kumar †

Abstract

Pressure dipoles are important long distance climate phenomena (teleconnection) characterized by pressure anomalies of opposite polarity appearing at two different locations at the same time. Such dipoles have proven important for understanding and explaining the variability in climate in many regions of the world, e.g., the El Niño climate phenomenon is known to be responsible for precipitation and temperature anomalies worldwide. This paper presents a novel approach for dipole discovery that outperforms existing state of the art algorithms. Our approach is based on a climate anomaly network that is constructed using the correlation of time series of climate variables at all the locations on the Earth. One novel aspect of our approach to the analysis of such networks is a careful treatment of negative correlations, whose proper consideration is critical for finding dipoles. Another key insight provided by our work is the importance of modeling the time dependent patterns of the dipoles in order to better capture the impact of important climate phenomena on land. The results presented in this paper show that these innovations allow our approach to produce better results than previous approaches in terms of matching existing climate indices with high correlation and capturing the impact of climate indices on land.

1 Introduction

Teleconnections, i.e., long distance connections between the climate of two places on the globe, have proven important for understanding and explaining the variability in climate in many regions of the world. Typically, these teleconnections are represented by time series known as *climate indices* [20], which are often used in studies of the impact of climate phenomena on temperature, precipitation, and other climate variables. One important class of climate indices are pressure dipoles,* which are characterized by pressure anomalies of opposite polarity appearing at two different locations at the same time.

Scientists have known of the existence of such dipoles for about a century. Two of the best known pressure dipoles are the North Atlantic Oscillation (NAO) and the Southern Oscillation (SO). NAO, which measures the difference in anomalies in pressure between Akyureyri in Iceland and Ponta Delgada in the Azores, captures the large scale atmospheric fluctuations between Greenland and northern Europe. A positive NAO index, which involves higher than normal pressure in northern Europe and lower than normal pressure around Iceland, is believed to be connected to warm and wet winters in Europe and cold and dry winters in northern Canada and Greenland. Conversely, a negative NAO index is associated with colder conditions in Europe and milder winters in Greenland. Figure 1 shows the time series of pressure anomalies for both Ponta Delgada (measured at 37.5N, 25W) and Akyureyri (measured at 65N, 17.5W).

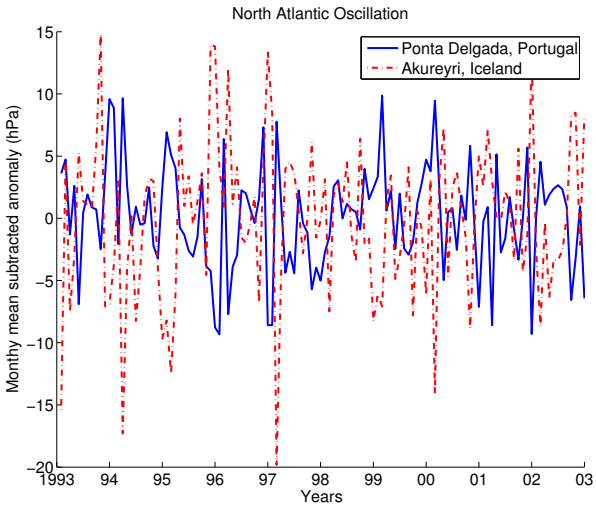

Figure 1: Pressure anomaly time series for the North Atlantic Oscillation. Note that these anomaly time series were constructed from the raw data by removing the monthly means from each time series.

The Southern Oscillation index (SOI) is measured as the difference in the pressure anomalies at

†Department of Computer Science, University of Minnesota, {kawale, steinbac, kumar}@cs.umn.edu

*Climate variables other than pressure can be involved in dipoles. For example, the Dipole Mode Index (DMI) [4], which has been investigated in relation to the Indian Monsoon.

Table 1: List of major pressure based climate indices.

Dipole	Climate Variable	Description
North Atlantic Oscillation (NAO)	Sea Level Pressure, Air Temperature	Characterized by the pressure anomalies at Ponta Delgada and Akyureyri at Iceland.
Southern Oscillation Index (SOI)	Sea Level Pressure, Air Temperature and Precipitation	Defined by pressure anomalies in Tahiti and Darwin, Australia
Pacific/North American Index (PNA)	Sea Level Pressure	Anomalies at the North Pacific Ocean and the North America
Antarctic Oscillation (AAO)	Sea Level Pressure	The first leading mode of the EOF analysis of pressure anomalies from 20°S poleward
Arctic Oscillation (AO)	Sea Level Pressure	The first leading mode of the EOF analysis of pressure anomalies from 20°N poleward
Western Pacific (WP)	Sea Level Pressure	Low frequency variability over the North Pacific with one center located over the Kamchatka Peninsula and another broad center of opposite sign covering portions of southeastern Asia and the low latitudes of the extreme western North Pacific

Tahiti and Darwin, Australia and captures fluctuations in pressure around the tropical Indo-Pacific region that correspond to the El Niño Southern Oscillation (ENSO) climate phenomenon [13]. A high value of SOI indicates higher pressure anomalies in the eastern tropical Pacific around Tahiti and lower pressure anomalies around Indonesia and northern Australia, while a low value of SOI is associated with the reverse conditions. Figure 2 shows the time series of pressure anomalies at Tahiti (measured at 17.5 S, 150W) and Darwin (measured at 12.5S, 130E).

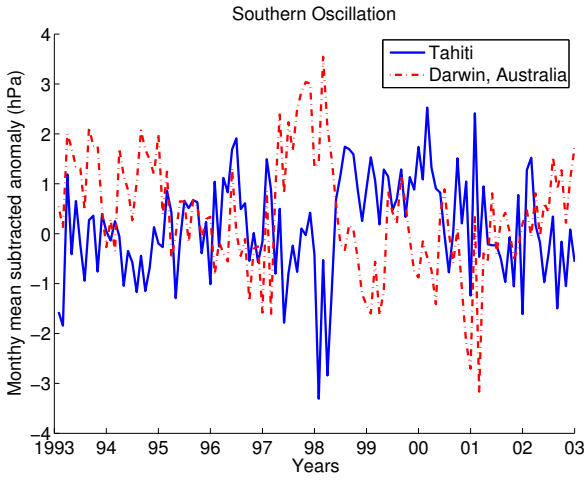

Figure 2: Pressure anomaly time series for the Southern Oscillation

As mentioned, climate indices, including dipoles, are of great importance in understanding climate variability. Table 1 lists some dipoles that are well known to climate researchers. These dipoles have been discovered by observation, e.g., SOI and NAO, or by EOF analysis [12], e.g., AO. However, all these discoveries have required considerable research and insight on the part of the domain experts involved. Because of the amount of effort involved and the possibility of missing indices, an automated approach to climate index discovery could be quite useful.

One of the first attempts in this direction was Steinbach et al [8, 9, 10]. The approach used a shared nearest neighbor (SNN) [2] clustering approach to find climate indices. More specifically, it built a graph of all locations on a latitude-longitude grid based on the positive pairwise correlations between the anomaly time series of temperature or pressure at these locations and then found clusters in this graph. The centroids of these clusters or the differences between two centroids were then used as candidate climate indices. Many of the resulting candidate indices showed a high correlation with known climate indices and were similar in their level of impact on land climate variables such as temperature.

Tsonis et al. [14] pioneered the use of complex networks to study climate systems. The authors constructed networks using nodes on a 5°x 5°grid on the globe, where the edges of the network were defined in terms of the (absolute) correlation values between the anomaly time series of climate variables (SST, SLP) of all the pairs of nodes. From this complete

correlation graph, only the edges with significant correlation (> 0.5) were retained. In the tropics, the network had very high connectivity and resembled a complete graph, while away from the equator, the network showed characteristics typical of a scale free network. The authors further showed that the supernodes in the scale-free network corresponded to major climate indices such as NAO and PNA [14, 15]. Other researchers have also applied complex networks to climate for examining the structure of the climate system Donges et al. [1], analyzing hurricane activity Forgarty et al. [3], and finding communities in climate networks and how they correspond to known climate patterns Steinhaeuser et al. [11].

In our work, we present comprehensive techniques to systematically find climate indices that are dipoles from the climate data. In the other approaches, negative correlations have often been ignored [8] or only absolute values of correlations have been considered [14]. However, as we show in Section 3, negative correlations are key for detecting dipoles, and thus, must be preserved in both sign and magnitude. In addition, a threshold is often used to eliminate spurious correlations, but using the same threshold for positive and negative correlations is not appropriate since negative correlations are usually weaker and many nearby locations have high positive correlation. We also study the change in climate indices over time unlike [8]. Although some of the approaches based on complex networks have taken time into consideration, we go further, defining *dynamic* climate indices and evaluating the improvement that results in terms of evaluating the impact on land.

1.1 Our Contributions: More specifically, the contributions of our paper are as follows:

1. We show the importance of treating negative correlations in climate data differently than positive correlations unlike [8, 14, 1].

2. We present algorithms for discovering dipoles from climate data that are cognizant of the positive and negative nature of correlation. This includes an algorithm based on discovering communities in complex networks. These algorithms are able to identify most of the major existing dipoles in climate data with higher correlation than current techniques.

3. Our approach provides a novel framework for studying the change in the dipoles across both space and time. Investigations using this framework reveal that the area weighted impact on the land is higher if the dipole climate indices are defined by moving rather than fixed locations. The utility of dynamic dipoles has been discussed by Portis et. al in [7]. However, to the best of our knowledge, this paper is the first one to compute dynamic climate indices automatically and study their impact on the local climate variables.

1.2 Organization of the paper The paper is organized as follows: Sections 2 and 3 describe the data and the preliminary analysis needed for the network construction, respectively. Our algorithms for dipole detection are presented in Section 4, while Section 5 discusses the dipoles discovered from the data. Section 6 provides an evaluation of the results with respect to existing climate indices and other work. Conclusions and possibilities for future work are presented in Section 7.

2 Dataset

For our analysis, we use pressure climate data from the NCEP/NCAR Reanalysis project provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA [17]. The NCEP/NCAR reanalysis project has data assimilated from 1948 – present which is available for public download at [18]. We focus on sea level pressure, which consists of monthly mean values at a grid resolution of 2.5°longitude x 2.5°latitude on the globe. In all, we have 62 years of data (corresponding to 744 monthly values) for 10512 grid locations on the globe. We chose pressure because it is an important climate variable and many of the well known climate indices are based upon it. Air temperature, although also important, is locally correlated with pressure.

3 Network Construction Method

We use a network construction method similar to [14] and [1] except that we do not threshold the networks by taking the absolute value of correlation and using a single threshold. Instead, we define separate thresholds for positive and negative correlations as is further discussed in section 3.3. We also use smoothing. The details of network construction are provided in the following subsections.

3.1 Data Smoothing The NCEP/NCAR Reanalysis data consists of monthly mean values for each of the climate variables. When we consider pairwise linear correlation between two time series, an anomalous peak or a valley around a month can distort the correlation value. In order to remove such inconsistencies, we smooth the data by considering the moving average of three months.

3.2 Seasonality Removal Generally, climate data has a strong seasonality signal due to the Earth's revolution. The seasonality component is typically uninteresting and masks other more interesting sig-

nals. In order to handle this problem, we preprocess the raw data by removing the monthly means in order to obtain anomaly values for each month. The data normalization for every location is performed as follows:

$$\mu_m = \frac{1}{end - start + 1} \sum_{y=start}^{end} x_y(m), \forall_{m \in \{1..12\}}$$

$$x_y(m) = x_y(m) - \mu_m, \forall_{y \in \{1948..2009\}}$$

In this formula, start and end represent the start and end years to consider for the mean and define the base for computing the mean for subtraction (in our case 1948 and 2009). μ_m is the mean of the month m and $x_y(m)$ represents the value of pressure for the month m and year y.Once we remove the monthly means, the resulting values are the anomaly time series for that location.

3.3 Edge Weight Estimation After we get the anomaly values for every node, the networks are constructed by looking at the similarity values between the anomaly timeseries of two nodes. We compute the similarity between two nodes by taking the Pearson correlation between the two time series at the nodes. Pearson correlation is a linear measure of similarity and is expressed as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where $\bar{x}$ and $\bar{y}$ are the mean of the two series X and Y, and s_x and s_y are the standard deviations of the two series.

We do not want to consider all the edges in the complete graph formed to be a part of the network for analysis as there are about 100 million edges in our case and most of them are uninteresting. Thus, the correlation threshold plays an important role in defining the network and must be chosen appropriately. Fig 3 shows the distribution of the correlations in the pressure network. Due to autocorrelation in space, the positive correlation goes as high as 1. However the negative correlation between any two nodes does not go as high. If we threshold the graph using a single absolute value (for e.g. 0.5) we will be using a very harsh filter for negative correlations but a weak filter for positive correlations that allows many spurious values to pass through. Fig 4 shows the distribution of edges which are greater than 5000km away. Note that most of the high positive correlation edges have disappeared and the distribution of the positive and negative correlation is now quite similar. In particular, the distribution of negatively connected edges is quite similar in both Fig. 3 and 4. This gives credence to our assumption that most of the very high positive correlation comes from nearby links. This also makes it harder to prune edges due to positive correlation since the pruning threshold needs to be cognizant of the physical distance between the nodes.

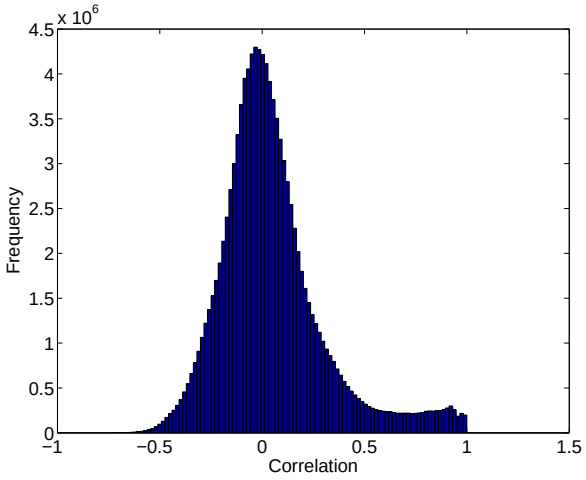

Figure 3: Distribution of correlation

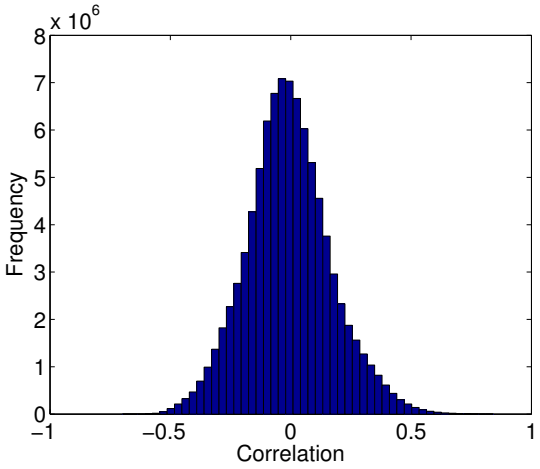

Figure 4: Distribution of correlation after filtering edges < 5000 km

We formally define the network or the graph to be an ordered pair G = (V, E) where V = {1,2.. N} is the set of nodes on the globe grid and E is the set of undirected edges (i, j) such that the r_{ij} is significant or above the threshold, which is different for negative and positive correlations. We construct networks for the pressure variable using a 20 year long window and slide the window by 5 years at a time so as to get 9 separate networks spanning 20 years each for our 62

years of data. Constructing such networks enables us to study the changes over time and is important in understanding the dynamics of the climate network.

4 Our Approach to Discover Dipoles

Automatic discovery of candidate dipoles faces several challenges. First, a formal definition is needed. For this work will will define dipoles as *pairs of regions whose locations have strong negative correlation with locations in the other region and strong positive correlation with locations in the same region.* The reason to look at regions instead of single locations is that the correlation between single locations can easily be spurious. On the other hand, if the size of the regions gets too large, the climate phenomenon will be diluted or disappear, so a careful balance needs to be striked out.

In the previous section, we presented a method to construct a weighted undirected graph where the nodes are location on the earth and the edge weights represent the strength in correlation in the time series of climate data at the two end points of the edge. The scale of the correlation ranges from -1 to 1, where 1 means perfect correlation and -1 indicates perfect anti-correlation. From the definition of a dipole, two locations within the same region of the dipole should share a strong positive correlation with each other while two locations in different regions of the dipole should have a strong negative correlation between them. Additionally, each region of the dipole should be geographically contiguous.

4.1 Algorithm for Constructing Dipoles(A1)

Our algorithm captures the essential characteristics of the dipole by a very simple mechanism. It first picks a negatively weighted edge from the graph and builds regions around it. This edge can be picked in several ways and the results of the algorithm depend on this choice. We use a simple approach that starts with the most negative edge in the network. The two end points, say $pt1$ and $pt2$, of the starting edge constitute two points of the dipole (of opposite polarity). Consider two sets of K number of locations, $N1$ and $N2$, that have the most negative correlations with $pt1$ and $pt2$, respectively. Similarly consider two sets of K number of locations, $P1$ and $P2$, that have most positive correlation with $pt1$ and $pt2$, respectively. If a node in $N1$ which belongs to the list of most negative edges on $pt1$ is in $P2$, i.e. it is also very highly positively connected to node $pt2$, then it becomes part of the dipole region consisting of $pt2$. Similarly if a node in $N2$ is also in $P1$ then it becomes part of the region consisting of $pt1$. In other words, if a node is highly negatively connected with node $pt1$ and highly positively connected with node $pt2$ then it becomes a part of the dipole region defined by $pt1$ and $pt2$. The details of the algorithm are presented below.

Algorithm 1 A1: Nearest neighbor approach to find dipoles

Require: Two starting points pt1, pt2 of the dipole, K the number of nearest neighbors to examine
  $Region1 \Leftarrow pt1$
  $Region2 \Leftarrow pt2$
  $P1 \Leftarrow K$ number of positive nearest neighbors of $pt1$
  $N1 \Leftarrow K$ number of negative farthest neighbors of $pt1$
  $P2 \Leftarrow K$ number of positive nearest neighbors of $pt2$
  $N2 \Leftarrow K$ number of negative farthest neighbors of $pt2$
  for $i = 1$ to K **do**
    {For every node in $N1, N2$ check if it is in $P2, P1$ respectively}
    $ind \Leftarrow find(N1(i), P2)$
    if $ind \neq 0$ **then**
      $Region2 \Leftarrow Region2 \cup N1(i)$
    end if
    $ind \Leftarrow find(N2(i), P1)$
    if $ind \neq 0$ **then**
      $Region1 \Leftarrow Region1 \cup N2(i)$
    end if
  end for
  if $size(Region1) \geq$ MIN-DIPOLE-SIZE **then**
    if $size(Region2) \geq$ MIN-DIPOLE-SIZE **then**
      return $(Region1, Region2)$
    end if
  end if
  return "no dipoles found"

It could happen that the starting edge has a spurious correlation and the regions around it do not lead to a dipole. In order to verify this, we check whether the size of the two regions has grown to be sufficiently large enough and only then label the two regions as a dipole. After finding a dipole, we remove the edges of the dipole from the network and continue finding further dipoles by picking up the next most negative edge in the graph until the resultant graph becomes very sparse or the most negative edge in the graph falls below a threshold. We used -0.4 as the threshold a negative edge must have in order to be considered by the dipole algorithm. There are about 1.5 million edges in the graph that have a negative correlation lower than -0.4.

Apart from the choice of starting node, this algorithm also depends on the value of K. We studied the impact of several different choices for the value of K (100, 300, 500, 1000, 2000). If we choose K to be very large, the regions of dipoles are very large (and often non-contiguous) whereas for a very small value of K the size of dipoles is small and might not be a good representative of the actual dipole.

Based on these empirical observations, we set K to be 300. This choice of K ensures that the two regions of dipoles are of a reasonable size. However we evaluate our results for different values of K.

Another key point to observe about this algorithm is that we do not explicitly check if region 1 or region 2 are contiguous or not. Since we consider the top 300 positive neighbors of a point due to spatial autocorrelation they will very likely be contiguous, however this is not guaranteed.

4.2 Community Based Method In the previous subsection, the method presented considered all the locations on the earth for building the dipole regions but the regions around the dipoles could be non-contiguous. In order to overcome this limitations, we consider partitioning the network before running $A1$. Network paritioning makes the resulting process more robust by constraining the search of dipoles within a much smaller region than the whole Earth, i.e., within a community consisting of positively and negatively correlated nodes.

We use a community detection algorithm for partitioning the network. The main goal of a community based approach is to partition the network into several smaller subsets of nodes and make the algorithm $A1$ less sensitive to the variability at smaller non contiguous locations that should not be a part of the dipole. The aim of clustering is to find regions containing nodes that are highly positively or negatively correlated. We can achieve this goal by guiding the community detection algorithm into partitioning the network into appropriate clusters by choosing the correlation thresholds as discussed in 3.3. Using the histogram of thresholds and empirical evaluation, we set the positive threshold to be very high (close to 0.85) so that only nearby contiguous locations form an edge and the negative threshold to be lower (close to -0.4) so that the significant dipoles are still captured.

For community detection, we chose Walktrap algorithm [6] which is based upon random walks. This algorithm is based on the fact that random walks tend to become trapped in dense part of the network corresponding to communities (clusters). Once we get communities from the entire network, we find dipoles within a community by using algorithm A1 and picking up the most negative edge within the community as the starting edge.

4.3 Summary: Thus to summarize, our algorithm for dipole detection consists of the following steps -

1. Construct the anomaly series from the smoothed data by removing the seasonality as mentioned in section 3.2.
2. Construct the network using the value of correlation from the anomaly data for different windows of time periods and for each network, threshold it to retain only the edges with significant correlations as described in section 3.3.
3. Generate clusters from the network data using clustering/community detection as mentioned in 4.2. (Alternatively we can run $A1$ directly on the entire graph.)
4. Using algorithm A1 within a cluster, separate the two ends of all negatively correlated edges within it into two buckets such that nodes within a bucket are positively correlated and the nodes in opposite buckets are negatively correlated .
5. The algorithm returns the two buckets formed from step 4 as dipoles if the size of each bucket is greater than a threshold.

5 Results

We ran our dipole detection algorithm $A1$ and its community version on the pressure data set from the NCEP/NCAR. We constructed anomaly data using the base for mean to be the entire 62 year duration. The networks were constructed for a period of 20 years each with a sliding window of 5 years so as not to introduce abrupt changes between networks. Thus we had 9 networks spanning 20 years each. Before running the community detection algorithm, we threshold the graph using a 0.85 value for positive correlation and -0.4 value for negative correlation. We use these thresholds with the intuition that it helps us include all the significantly negative edges in the graph but we are still very strict for the positive edges since we need them only to construct homogenous regions around each end of the dipoles. This threshold helped us get all the major dipoles as a part of a community each. In order to find communities, we used walktrap version 2 with default parameters (random walk length=4). The following sections describe some of the well known dipoles that we discovered from the data.

- Southern Oscillation: The SO is one of the most important dipole in climate. It is clearly seen in all the networks with correlations close to 0.9.
- North Atlantic Oscillation: NAO is a well established fluctuation in opposite phase of the climate of Greenland/Iceland and northern Europe. From the data we see that NAO is one of the strongest signals. When we pick up the most negative edge in the graph, most of the time it is

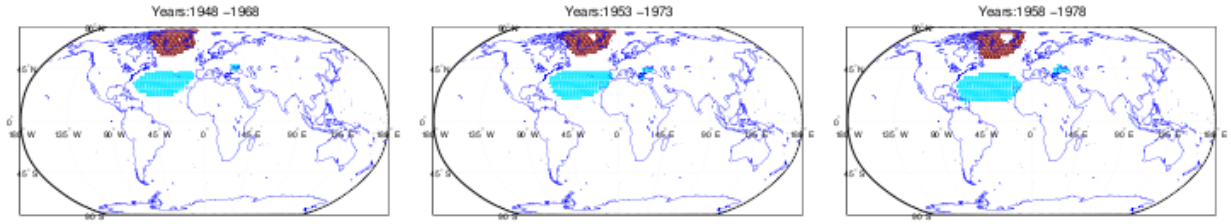

Figure 5: Different phases of NAO seen in pressure data

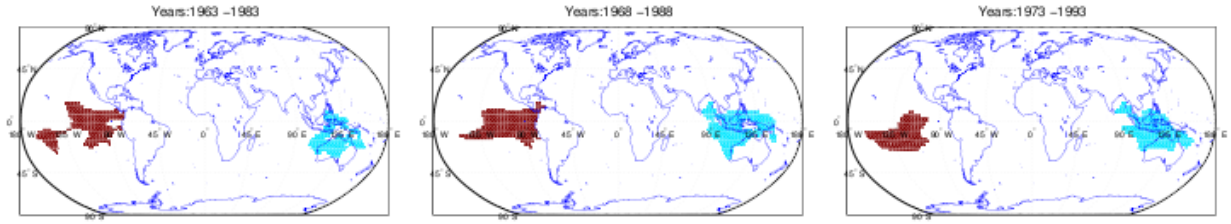

Figure 6: Different phases of SOI seen in pressure data

in the NAO region. The North Atlantic Oscillation is seen very clearly in all the 9 networks of 20 year periods.

- Arctic Oscillation: The Arctic Oscillation is the pressure anomaly around the North pole and is defined on the basis of the first leading component of an EOF analysis using the region north of 20N latitude. It does not have a pair of physical locations associated with it. However using our method we are able to find it in all the 9 networks with a very high correlation.

- Antarctic Oscillation: The Antarctic Oscillation measures the anomaly of pressure around the Antarctic region. This oscillation is the analog of the Arctic oscillation in the southern hemisphere and is also defined by EOF analysis of locations south of 20S. We see the Antarctic Oscillation in all the climate networks. However the climate indices data from the Climate Prediction Center is defined from 1979 onwards[19]. Hence we can only compare its correlation with known climate indices for the last two networks.

- Western Pacific Index: The Western Pacific index is north south dipole around the western Pacific with one end located over the Kamachatka peninsula and the other end in southeastern Asia and the subtropical north Pacific.

6 Experimental Evaluation

In order to evaluate the goodness of the dipole regions generated, we look at three things -

1. Strength of the negative correlation between the two regions of the dipole. Higher negative correlation implies a stronger dipole.

2. Correlation with known dipole indices. This highlights the ability to reproduce known dipoles.

3. Impact of the dipole indices on land by computing an area weighted correlation of land temperature anomalies with the dipole indices. This highlights the ability of data driven dipoles to potentially outperform known dipoles.

6.1 Negative Correlation within regions of Dipole From the definition of the dipole, the two regions forming a dipole should be negatively correlated with each other. To compute the strength of the negative correlation across the two regions, we look at three values -

1. The mean value of the correlation between all the locations pairs across two regions constituting the dipoles. We call this value mean of all pairs.

2. The best correlation in the two regions of the dipole represented by the most negative edge in the two regions. We call this value the best pair.

3. Compute the mean of the anomalies of all the locations at each region and then take the correlation between them. We call this pair of means.

Table 2 shows the three correlation values of the dipole regions discovered by our algorithms. The table reports the mean values for all the 9 networks.

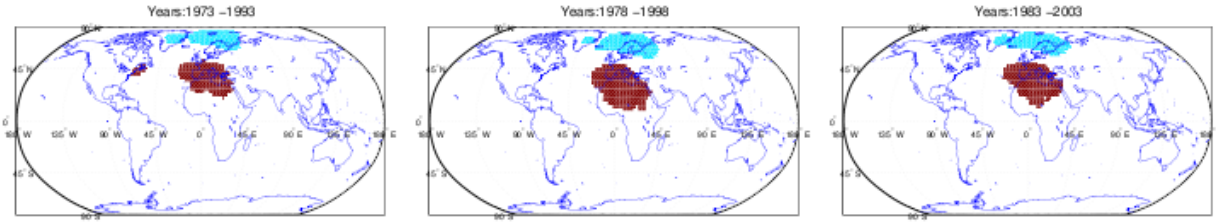

Figure 7: Different phases of AO seen in pressure data

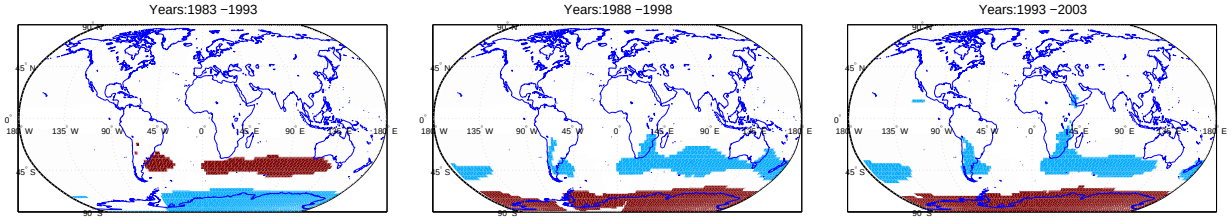

Figure 8: Different phases of Antarctic Oscillation seen in pressure data

From the table it can be seen that all the regions are strongly negatively correlated, indicating that the regions indeed consist of strong opposing pressure polarities.

We performed a further analysis of the SOI region and found that the negative correlation between Tahiti and Darwin is not as strong as several other location pairs. Fig 9 shows the correlation between Tahiti and Darwin as well as the best pair results from our two dipole finding algorithms. This results indicates that the underlying phenomenon leading to the negative correlation is not fixed at Tahiti and Darwin and that SOI and other climate indices are perhaps better captured with dynamic clusters.

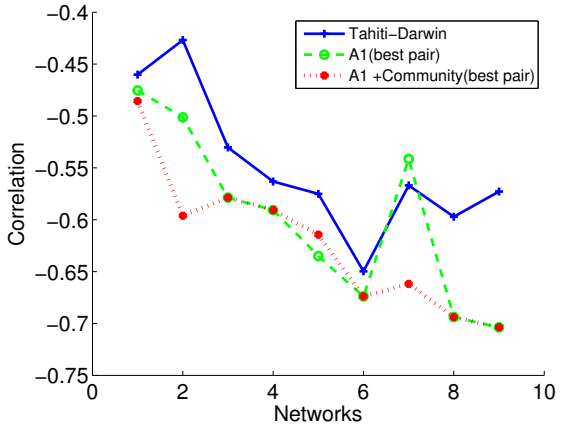

Figure 9: Best negative correlation for the SOI cluster. Lower curves are better.

6.2 Comparison with known Climate Indices

In order to evaluate the goodness of the dipole clusters found, we compared them with some well known climate indices. For each of the 9 network periods, we generated a set of dipoles from the corresponding network. For every dipole belonging to a time period, we took the two clusters belonging to the dipole and computed their centroids by taking the mean of the anomaly at those locations during that time period. We computed the difference in between the two cluster centroids to create a time series which is then compared with all the climate indices over that period using linear correlation. We kept track of the best correlation to the climate indices during the period and recorded the dipole cluster that best matched each climate index. We performed this step for all the time periods. Table 3 shows the the best correlation to each climate index of the dipoles found using the two variations of algorithm $A1$ with a bin size of 300. Although $A1$ + community shows weaker correlation than $A1$ in a number of cases, the impact of $A1$ + community is sometimes still better as is shown later in the paper.

From Table 3 we see that using our algorithm for with $A1$ with or without a community, we are able to match that with an average precision of 0.88 and 0.86, respectively to find SOI. Another important dipole that we find with very high correlation is the Arctic oscillation. Climate scientists define the Arctic Oscillation as the first leading component of an EOF analysis and thus it does not have interpretation in terms of pairs of locations. However, using our method we are able to find a pair of negatively

Dipole	A1			A1 + community		
	Mean of all pairs	Best pair	Pair of means	Mean of all pairs	Best pair	Pair of means
SOI	-0.4425	-0.5993	-0.3658	-0.4482	-0.6221	-0.3426
NAO	-0.4584	-0.7171	-0.6997	-0.4598	-0.7170	-0.7019
AO	-0.5071	-0.7405	-0.6950	-0.5063	-0.7405	-0.6974
AAO	-0.4187	-0.5478	-0.4988	-0.4173	-0.5639	-0.4345
WP	-0.4000	-0.5139	-0.4424	-0.4069	-0.5301	-0.4635

Table 2: Strength of negative correlation of identified dipoles using our algorithm.

correlated clusters whose difference correlates very well with the AO climate index (as high as 0.85).

To evaluate the sensitivity of our analysis to the choice of the dipole bin size, K, we looked at the mean correlation with the climate indices using different values of K. These results are shown in Fig 10. As expected a small value of K gives very focused patterns for SOI and NAO and leads to better correlation because they are actually defined by single point locations. However we see that at very small values of K, the correlation of the dipole cluster with AO or AAO is not as high as for larger values of K. This is because the AO and AAO patterns are not defined using single point locations, but instead are defined as a summary of the behavior of a large region.

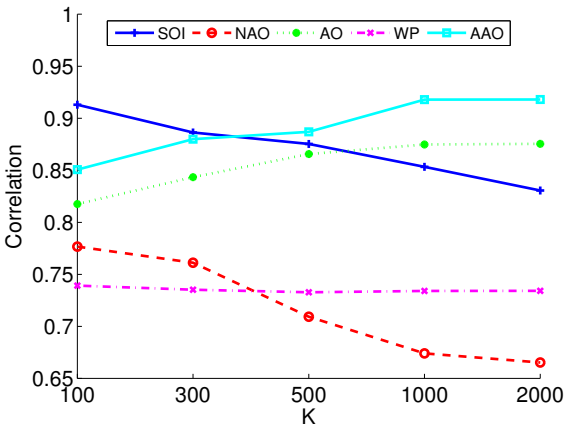

Figure 10: Effect of varying the region size K on A1

6.2.1 Comparison with existing approaches

We also compare our approach for finding dipoles with existing approaches to find them. In [8], Steinbach, et al present a SNN clustering based approach to find climate indices. We use the numbers as reported in [8] for our comparison. The SNN clustering numbers are shifted correlations, however we for our numbers we use linear correlations only. Computing shifted correlations will only improve the numbers further. From 4 we see that our algorithm is better than the existing approaches to find climate indices. Note that for $A1$, we report the mean values that we got from choosing K=300 as shown in Table. 3.

6.3 Area weighted correlation with land temperature From the previous sections we see that we can generate dipoles that dynamically change over time and from the results we see that their correlation with known climate indices is very high. In order to study the changes in the dipole clusters over time, we take their centroids and plot them on the globe. Fig.11 shows the plot of moving centroids of the Arctic Oscillation dipole. Fig 12 shows the plot of moving centroids of the North Atlantic Oscillation cluster.

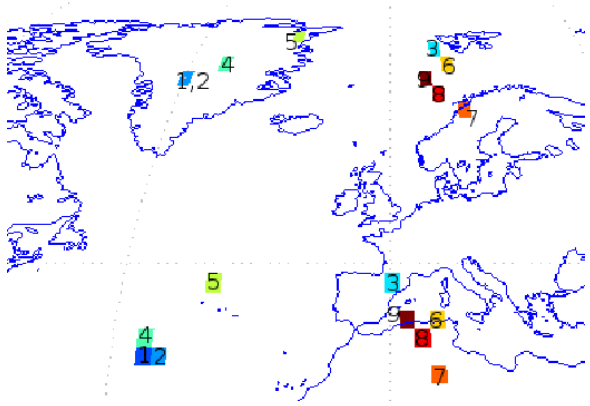

Figure 11: Moving Centroids of Arctic Oscillation

We hypothesize that the climate indices are better captured by considering them to be moving.To verify this hypothesis, we compute the area weighted correlation with the land temperature anomalies for both static and the dynamic index for each of the 9 different network periods. Our dynamic index for SOI is computed by taking the mean of the two regions of the dipole and their difference. We also generated a random baseline and compute the correlation of land

Table 3: Correlation of our dynamic indices with known climate indices (K=300)

Network	**A1**					**A1 + community**				
	SOI	NAO	AO	AAO	WP	SOI	NAO	AO	AAO	WP
1	0.8885	0.7686	0.8665	-	0.7166	0.8761	0.7686	0.8665	-	0.7163
2	0.8696	0.7729	0.8506	-	0.7231	0.7378	0.7711	0.8529	-	0.7232
3	0.9012	0.7312	0.8560	-	0.7400	0.8952	0.7317	0.8580	-	0.7399
4	0.8895	0.8044	0.8353	-	0.7306	0.8828	0.8043	0.8353	-	0.7298
5	0.8983	0.7279	0.8037	-	0.7523	0.8540	0.7283	0.8037	-	0.5017
6	0.9214	0.7498	0.8648	-	0.7602	0.9195	0.7488	0.8702	-	0.7408
7	0.8387	0.7769	0.8137	-	0.7604	0.8318	0.7819	0.8137	-	0.7318
8	0.8946	0.7581	0.8407	0.8763	0.7240	0.8933	0.7582	0.8407	0.8797	0.4369
9	0.8746	0.7609	0.8597	0.8835	0.7103	0.8737	0.7621	0.8597	0.8809	0.7095
Mean	0.8863	0.7612	0.8434	0.8799	0.7353	0.8627	0.7608	0.8445	0.8803	0.6642

Climate Indices	**SNN Clustering**	**Our approach**	
	(Shifted Correlation)	A1	A1 + community
SOI	0.7312	0.8863	0.8627
NAO	0.7519	0.7612	0.7608
AO	0.7577	0.8434	0.8445
AAO	-	0.8799	0.8803
WP	0.2857	0.7353	0.6642

Table 4: Comparison with existing approaches.

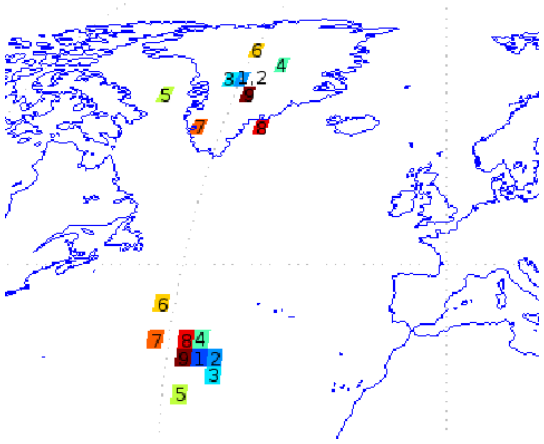

Figure 12: Moving Centroids of North Atlantic Oscillation

temperature anomalies using any two random locations. For each network period we picked 100 pairs of random locations from the globe having a correlation > 0 among them and computed the difference between their anomalies and calculated their average impact on land temperature anomalies. This constitutes the random baseline. The mean of the random baseline is shown in Fig.13 and 14 and the box around the mean shows the interquartile range and the median. Fig.13 shows the comparison of impact on land of our index, SOI, and the random baseline. These results also show that both algorithms $A1$ and $A1$ + community have a stronger impact on land temperature anomalies than SOI—sometimes up to 90% better. In fact, $A1$ + community gives a better performance for all the network years. Note that $A1$ + community gives better results than $A1$ even though $A1$ showed slightly better correlation with static SOI as shown in Table 3.

Instead of just looking at the centroid we also compared the index generated from the best correlation pair. The best correlation pair is the edge with the strongest negative correlation in the dipole cluster. Fig. 14 shows the area weighted correlation from the best pair. We see that the best pair does not always perform better than the mean which provides more stability to the index. The best pair still has a better impact on land temperature anomalies as compared to the SOI index. Fig. 15 and 16 show the correlation with land temperature anomalies for a single network. From the figures it can be seen that the dynamic index is similar in pattern to the static index but has a much stronger correlation.

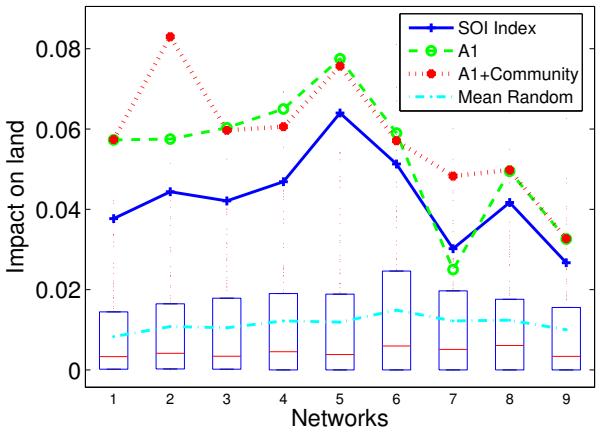

Figure 13: Area weighted correlation of land temperature anomalies using SOI index vs our indices generated from the cluster centroids

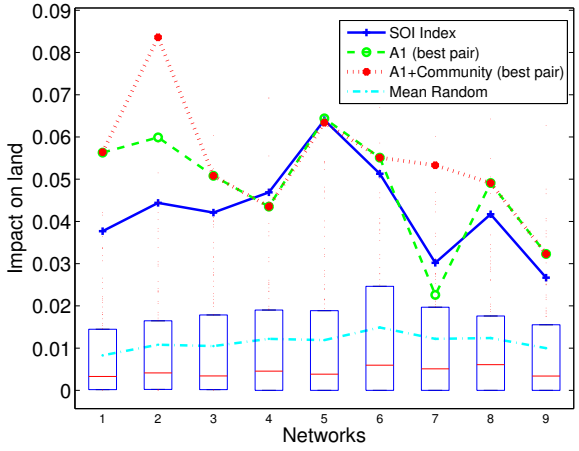

Figure 14: Area weighted correlation of land temperature anomalies using SOI index vs best correlation pair in our dipole cluster given by the algorithms

7 Conclusion and Future Work

This paper presents a novel approach to find dipoles using the climate data. The problem of finding dipoles has been of key interest to climate scientists as it helps in a greater understanding of the teleconnections and several important extreme phenomenons. Finding dipoles has been particularly interesting to the data mining community as the underlying data is not only large but also has a spatio-temporal nature presenting challenges such as seasonality, high variability, autocorrleation, etc. In this setting, we propose a method based on greedy heuristics to identify dipoles. Our methodology seems to produce considerably better results than the current state-of-art algorithms.

The algorithm $A1$ proposed in the paper and it's community version is effective and efficient to implement. Our community based approach to first partition the large network of all locations on the globe narrows the search space for $A1$ algorithm, generates fewer candidate dipoles, removes spurious connections and is able to match the performance of $A1$. However, further investigation is needed to determine if one of these algorithms is to be clearly preferred to the other.

A larger significance of this work, which might impact how climate scientists perceive the climate indices, is that it shows climate indices are better explained as centroids of dynamic clusters. So far, climate scientists have mostly considered climate indices to be fixed. The evidence that supports our claim is that the area weighted correlation of the SOI index with land temperature anomalies is improved by up to 90% by capturing the index as a centroid of moving clusters rather than fixed locations. Given the importance of the Southern Oscillation on the climate of the globe, this result has significant impact in terms of predictions in climate science. The Southern Oscillation is closely tied with the El Niño phenomenon which drives the extreme weather events like tropical cyclones, droughts, hurricanes, etc. A thorough evaluation of this is part of future work.

In addition to further evaluation and improvement of the approaches presented in the paper, we need to go beyond comparisons to current climate indices to see if any novel dipoles can be discovered. Although it is unlikely that any of these would be as significant as NAO or SOI, such dipoles could still be of great regional importance.

Acknowledgement

This work was supported by NSF grants III-0713227, IIS-0905581, and IIS-1029711. We also thank Dr. Stefan Liess and Dr. Shyam Boriah for their comments and feedback.

References

[1] Donges, J. F., Zou, Y., Marwan, N. Complex networks in climate dynamics. In *European Physical Journal Special Topics*, 174 (1), pp. 157–179, 2006.

[2] Ertoz, L., Steinbach, M., Kumar. V. A new shared nearest neighbor clustering algorithm and its appli-

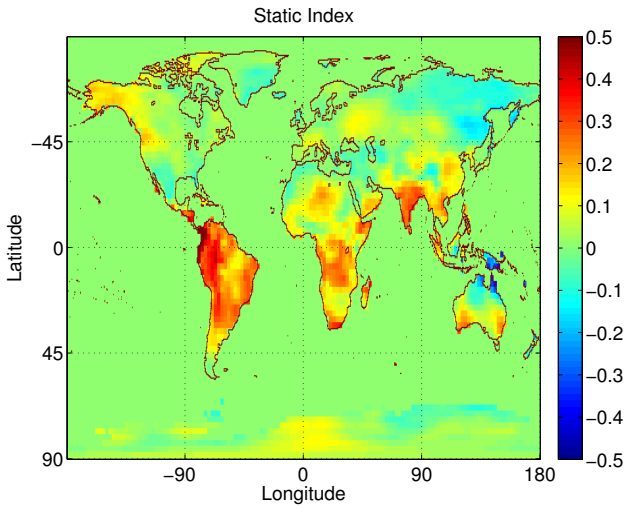

Figure 15: Area weighted correlation of land temperature anomalies using SOI index for network 2

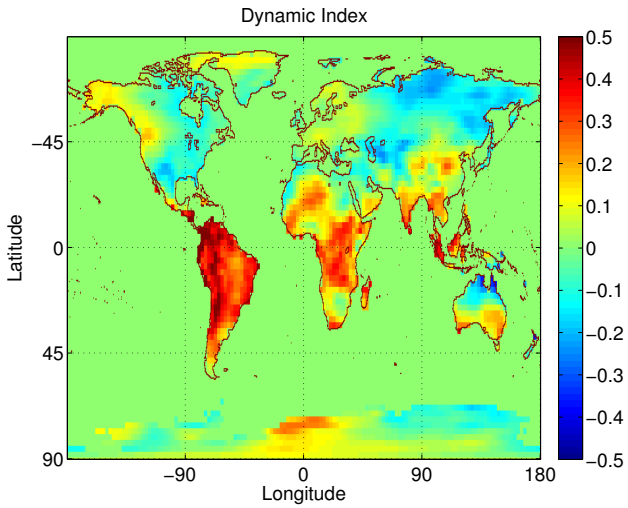

Figure 16: Area weighted correlation of land temperature anomalies using our dynamic index generated from $A1$ + Community for network 2.

cations. In *Workshop on Clustering High Dimensional Data and its Applications*, SIAM Data Mining, 2002.

[3] Fogarty, E. A., Elsner, J. B., Jagger, T. H., Tsonis, A. A. Network Analysis of U.S. Hurricanes, *Hurricanes and Climate Change*, 1-15, 2009.

[4] Gadgil, S. and Vinayachandran, PN and Francis, PA and Gadgil, S. Extremes of the Indian summer monsoon rainfall, ENSO and equatorial Indian Ocean oscillation. In *Geophysical Research Letters*, 174 (1), pp. L12213–1, 2004.

[5] Gozolchiani, A., Yamasaki, K., Gazit, O., Havlin S. Pattern of climate network blinking links follows El Niño events. In *Europhysics Letters*, vol 83, issue 2, 2008.

[6] Pons, P., Latapy, M. Computing Communities in Large Networks Using Random Walks. *Journal Graph Algorithms Applications*. 10(2): 191-218, 2006.

[7] Portis, D. H., Walsh, J. E., El Hamly, Mostafa and Lamb, Peter J., Seasonality of the North Atlantic Oscillation, *Journal of Climate*, vol. 14, pg. 2069-2078, 2001.

[8] Steinbach, M., Tan, P., Kumar, V., Klooster, S., and Potter, C. 2003. Discovery of climate indices using clustering. In *SIGKDD international Conference on Knowledge Discovery and Data Mining*. KDD, pg. 446–455, 2003.

[9] Steinbach, M., Tan, P., Kumar, V., Potter, C and Klooster, S. Data mining for the discovery of ocean climate indices. In *Mining Scientific Datasets Workshop*, 2nd Annual SIAM International Conference on Data Mining, 2002.

[10] Steinbach, M., Tan, P., Kumar, V., Potter, C., Klooster, S. Clustering earth science data: Goals, issues and results. In *Proceedings of the 4th KDD Workshop on Mining Scientific Datasets*, 2001.

[11] Steinhaeuser, K., Chawla, N. V., Ganguly, A. R. An exploration of climate data using complex networks. *KDD Workshop on Knowledge Discovery from Sensor Data*, pp. 23–31, 2009.

[12] Storch, H. V. and Zwiers, F. W. Statistical analysis in Climate Research. Cambridge University Press, 1999.

[13] Taylor, G. H. Impacts of the El Niño/southern oscillation on the pacific northwest. *Technical report*, Oregon State University, USA, 1998.

[14] Tsonis, A. A., Swanson, K. L., Roebber, P. J. What Do Networks Have to Do with Climate. In *Bulletin of the American Meteorological Society*, vol. 87, no. 5, pg. 585–595, 2006.

[15] Tsonis, A. A., Swanson, K. L., Wang, G. On the role of atmospheric teleconnections in climate. In *Bulletin of the American Meteorological Society*, vol. 21, issue 12, 2008.

[16] Tsonis, A. A. and Swanson, K. L., Topology and Predictability of El Niño and La Niña Networks. In *Physics Review Letters*, vol. 100, no. 22, 2008.

[17] E. Kalnay, et al, 1996. *The NCEP/NCAR 40-Year Reanalysis Project* Bulletin of the American Meteorological Society, Vol. 77, No. 3. (1 March 1996), pp. 437-470.

[18] http://www.esrl.noaa.gov/psd/data/

[19] http://www.cpc.ncep.noaa.gov/

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