Mining Relationships in Spatio-temporal Datasets

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jaya Kawale

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Vipin Kumar

January, 2013

Acknowledgements

I am grateful to my advisor Prof. Vipin Kumar for his kind support and guidance throughout my PhD. Vipin provided me invaluable support and was truly inspiring and many of his direct or indirect teachings will stay with me for a long time to come. I greatly admire his commitment towards work and students, his infectious enthusiasm, his passion towards research and science and his down to earth disposition and I wish to imbibe that myself in my life. It was a great pleasure and honor to work with him and I thoroughly enjoyed the journey.

I also thank Prof. Shashi Shekhar, Prof. Arindam Banerjee and Prof. Peter Snyder for kindly agreeing to be a part of my thesis committee and advising me through the process. I also want to thank all my collaborators without whom I would not have achieved much success in my research especially Dr. Michael Steinbach, Dr. Stefan Liess, Prof. Auroop Ganguly and Prof. Snigdhansu Chatterjee. I also want to thank the students who have been a part of the dipole project who helped me learn in the process especially Dominick, Arjun, Saurabh and Luchiana. I want to thank my labmates for the in-numerous laughs that we shared together especially Varun, Xi, Ashish, Yashu, James, Lydia, Arjun, Saurabh and Shyam.

Finally, I want to thank my husband Aditya who is the superhero of my life. He is not just the love of my life but a friend, a collaborator, a supporter and a critique. I am indebted to my parents for their love, support and sacrifice throughout my life. They

have been my strongest pillars of support during my difficult times. Last but not the least I want to thank our sweet friends Amber and Shruti who provided color to our lives outside school.

Dedication

To my parents and late grandparents - there's a part of something you in everything in me.

Abstract

The generation of spatio-temporal datasets has seen a phenomenal growth in the past few years with the advances in remote sensing and location sensing devices. Data in many domains like climate, remote sensing, mobile computing, network monitoring, etc. are characterized by spatial and temporal dimensions and have features like spatial and temporal autocorrelation, complex dependence structures such as non linear associations, time lagged associations and long-range spatial dependencies (also known as teleconnections). Traditional methods of data mining usually handle spatial and temporal dimensions separately and thus are not very effective to capture the dynamic relationships and patterns in spatio-temporal datasets. In this thesis, we present methods and algorithms to analyze the spatio-temporal datasets and to discover patterns.

In particular, the focus of the thesis is on finding a key kind of patterns known as *teleconnections*. Teleconnections are recurring patterns in climate anomalies connecting two regions that are far apart from each other. They have been a subject of interest to climatologists due to the possibility of the linkages in the changes in weather at one location to the changes at another distant location. Two of the most important teleconnection patterns are the El Nino Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO). These teleconnections are important to study in climate because they are known to impact precipitation and temperature anomalies at large parts of the globe.

Scientists have known of the existence of a number of these relationships and historically they have been discovered by human observation or by using pattern analysis techniques such as the Empirical Orthogonal Function (EOF) over a limited region. However there are several limitations of the existing methods of finding these relationships, and they required considerable research and insight on the part of the domain experts involved.

This thesis aims to provide systematic data guided approaches to find such relationships in spatio-temporal data. Discovery of relationships or dependencies among climate variables involved is extremely challenging due to the nature and massive size of the data. In this thesis, we provide a graph based approach to find these patterns in climate datasets. Our approach generates a single snapshot picture of all the teleconnections on the globe and hence it enables us to precisely study the interactions and changes in behavior over time. We are able to identify most of the known connections with a high precision. Further, we show that some of the indices discovered using data guided approaches can capture the impact on temperature anomalies with a much higher correlation as compared to the static indices used by climate scientists. We also extend the algorithm to find time lagged relationships in climate data which are important to study as they add predictive power to these relationships. We further provide an algorithm to test the significance of the teleconnection patterns. Significance testing in a spatio-temporal setting needs to take into account the various characteristics of spatio-temporal datasets like non-i.i.d. data, trends, seasonality, etc. Our approach takes into account all these aspects of the datasets and helps us to remove spurious edges thus potentially enabling us to find new connections. This thesis shows that data guided approaches offer a huge potential for characterizing and discovering unknown relationships and thus advancing climate science.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Data in many scientific domains including climate are characterized by both spatial and temporal dimensions. Some of the key features of such dataset includes spatial and temporal autocorrelation, complex dependence structures such as non linear associations, time lagged associations and long-range spatial dependencies (also known as teleconnections). Mining relationships in spatio-temporal datasets is challenging due to the nature and size of these datasets. Traditional methods of data mining generally treat the spatial and temporal dimensions separately. However, there is a need to study these dimensions in conjunction to better understand the dynamic relationships as both "spatial" and "temporal" nature of the dataset adds substantial complexity to the data mining task. In this thesis, we present novel algorithms and approaches to handle some of the challenges of these datasets. In particular, we focus on the problem of finding "teleconnections" in climate datasets.

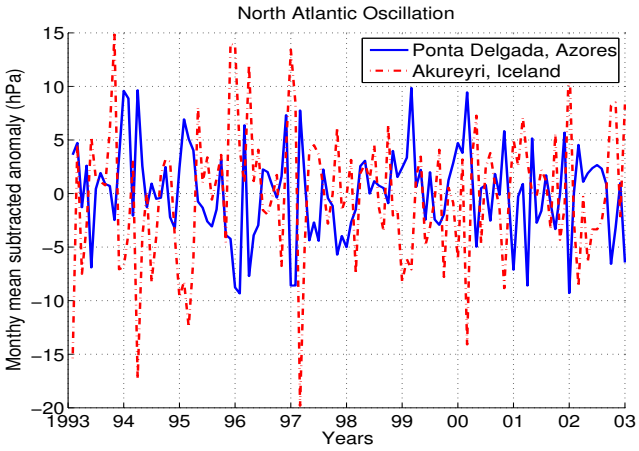
Teleconnections are recurring patterns in climate anomalies connecting two regions that are far apart from each other. These teleconnections represent a persistent and large scale temporal correlation in a given climate variable between two distant geographical locations. They have been a subject of interest to climatologists due to the possibility

of the linkages in the changes in weather at one location to the changes at another distant location. They reflect an important variability in the atmospheric and ocean circulation and are a key to understanding the physical relationships and processes in the atmosphere. They are known to impact and explain the climate variability in many regions of the world and form a critical missing link in the understanding of the atmosphere-ocean interactions and the global climate system. Teleconnections research is of a considerable importance due to the possible predictions that can be based upon such relationships and these relationships are so prevalent that their study forms a subfield in atmospheric sciences [14].

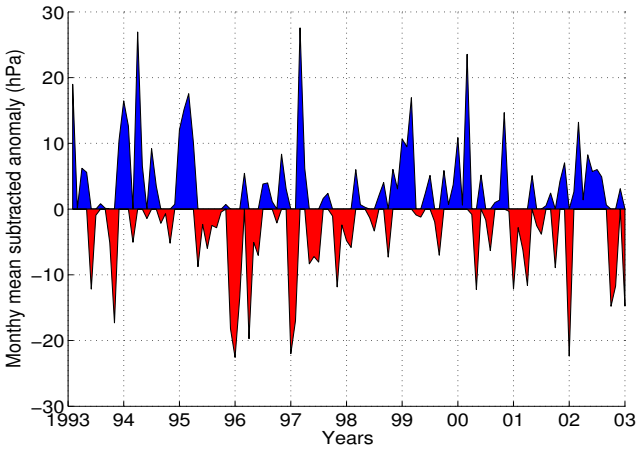
Typically, these teleconnections are represented by time series known as *climate indices* [4], which are often used in studies of the impact of climate phenomena on temperature, precipitation, and other climate variables. One important class of climate indices are pressure dipoles,[1] which are characterized 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 [77, 59, 78]. Two of the best known pressure dipoles are the North Atlantic Oscillation (NAO) and the Southern Oscillation (SO). NAO, which is represented by 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 central Europe. A positive NAO index, which involves higher than normal pressure in central 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 1.1(a) shows the time series of pressure anomalies for both Ponta Delgada and Akyureyri in the NCEP reanalysis dataset [43].

[1] Climate variables other than pressure can be involved in dipoles. For example, the Dipole Mode Index (DMI) [30], which has been investigated in relation to the Indian Monsoon, is a temperature based dipole.

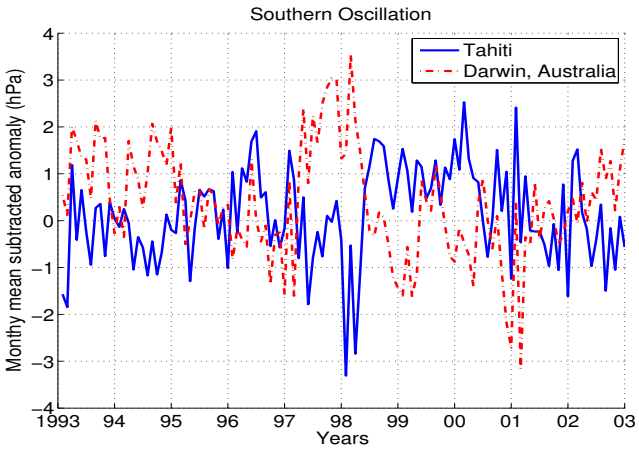


(a) Pressure anomaly time series at Ponta Delgada (measured at 37.5N, 25W) and Akyureyri (measured at 65N, 17.5W) using the NCEP dataset.

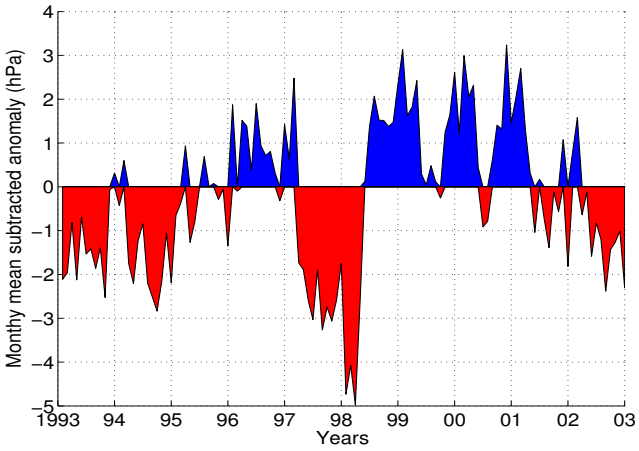


(b) NAO index formed by the subtraction of the two anomaly series. The red and blue shaded regions represent negative and positive phases of the dipoles respectively.

Figure 1.1: North Atlantic Oscillation



(a) Pressure anomaly time series at Tahiti (measured at 17.5 S, 150W) and Darwin (measured at 12.5S, 130E) using the NCEP dataset.



(b) SO index formed by the subtraction of the two anomaly series. The red and blue shaded regions represent the El Niño and the La Niña phase respectively.

Figure 1.2: Southern Oscillation

The Southern Oscillation index (SOI) is measured as the difference in the pressure anomalies at 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 [73]. 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. The pressure anomaly time series of the two locations Tahiti and Darwin is shown in the Figure 1.2(a).

Some of these dipoles have been discovered by human observation (for e.g. SO and NAO). However, later on with the availability of large amount of satellite data, climate scientists started to use slightly more sophisticated techniques like the EOF analysis to define these dipoles. However there are several limitations of the existing methods of finding these relationships, and they required considerable research and insight on the part of the domain experts involved. Knowledge of these teleconnections and their interactions is particularly important for predicting climate extreme events. For example, while the cold winter over Europe in 2010 could be largely explained by the North Atlantic Oscillation (NAO) which is another teleconnection, and other local indices, the cold winter over North America at the same time is largely due to a combination of NAO and ENSO [31]. Further, the ability to address important questions like the degree of climate change and its potential impacts requires a deeper understanding of the behavior and interactions of these atmospheric processes as well as to capture them precisely. This thesis aims to provide systematic data guided approaches to find such relationships in spatio-temporal data. Discovery of relationships or dependencies among climate variables involved is extremely challenging due to the nature and massive size of the data. Data guided approaches offer a huge potential for characterizing and discovering unknown relationships and advancing climate science.

Apart from these instantaneous relationships, we also present algorithms to capture the

time lagged relationships between different variables. Relationships in Earth science are often delayed as a phenomenon occurring at a place occurs at another after some point of time (e.g. extreme precipitation). These lagged relationships are important to study and understand from the perspective of prediction.

Another important challenge in a spatio-temporal setting is testing the significance of the generated patterns. Once we have algorithms to generate patterns from spatio-temporal datasets, we need a significance testing method to rule out spurious patterns. We provide an algorithm to test the significance of the generated patterns.

The research goals of this thesis are mentioned in the following section.

1.1 Research Goals

The main objective of this thesis is to develop models and algorithms for the analysis of spatio-temporal data and mine patterns and associations from it. To realize this goal, we have outlined the following sub-tasks:

- **Discovering dipoles**: Dipoles represent a class of teleconnections characterized by pressure anomalies of opposite polarity appearing at two different locations at the same time. The Southern Oscillation described earlier is one of the most well known dipoles. Our goal is to develop approaches to systematically capture all the dipoles in a given dataset.
- **Finding time lagged relationships**: Spatio-temporal data are often characterized by a delayed relationship between different variables. These lagged relations signify the time lag between the cause and the effect and are important to study and understand. E.g., the ENSO signal originates over the East Pacific but its remnants propagate for months to impact remote locations [26]. Our goal is to develop algorithms to capture the lagged relationships in a given dataset.

- **Testing the significance of spatio-temporal patterns**: Once we have algorithms to analyze spatio-temporal datasets and extract patterns from them, we need to have a mechanism to test the significance of the generated patterns in order to rule out spurious patterns that could be generated by random chance. Given the importance of dipoles in understanding the global climate system, a significance based testing algorithm can help us identify a potentially new connection from the generated patterns.

- **Understanding data preprocessing** : One of the main issues with climate datasets is the understanding of the various preprocessing steps generally applied to the datasets, for e.g., anomaly construction and detrending. Our goal is to analyze how anomaly construction affects the nature of the generated patterns and a different choice of a reference base biases the outcome.

1.2 Organization and Contribution

Our main contributions and the organization of the chapters are as discussed below.

Chapter 2 describes the dataset and the preprocessing that we have used for most of our analysis. An important component of Earth Science data is the seasonal variation in the time series. The seasonality component is the most dominant component in the Earth science data. The seasonal patterns even though important are generally known and hence uninteresting to study. In order to remove seasonality from the raw data, climate scientists generally remove the monthly mean value from the raw data and generate "anomalies". Another important preprocessing step is detrending the anomaly time series to remove long term trends present in the data.

Chapter 3 presents data guided approaches to find dipoles in climate data. We use a

graph based representation of climate data like many other previous researchers [69, 61, 62, 19]. However, unlike the previous research we do not ignore the negative correlations or consider the absolute value of the correlations. We show the importance of treating negative correlations differently and their key utility in finding dipoles . We present a novel graph based approach to find dipoles in the climate data that overcomes the shortcomings of the previous approaches [61]. An important utility of the dynamic dipoles defined using our approach is that they have a much higher correlation with temperature anomalies as compared to the static indices used by the climate scientists, for e.g., the ones available at the Climate Prediction Center (CPC) [1]. Our approach allows us to have a single snapshot of all the dipoles on the globe. This ability enables us to discover new dipoles and comprehensively study the behavior, their interaction, and the movement of the various dipoles in a more precise manner. Our approach of dipole discovery is able to find a pair of regions (dipoles) that highly correlate with the dipoles defined earlier using EOF analysis and have a number of well known limitations [18]. One of the major limitations of the EOF analysis is the orthogonality constraint of the principal components because of which it is difficult to physically interpret the discovered patterns. Further the centers of action using the EOF analysis do not need to correspond to the actual physical process. Also the EOF analysis is very sensitive to the choice of spatial and time domain. We also discuss some of the applications of dipole analysis. Our approach's ability to detect and visualize all the dipoles in a given dataset empowers our understanding of climate data in many ways. For example, it allows us to study the changes in the dipole behavior. Understanding the dipole movements throughout the different time periods enables us to have a better prediction.

Chapter 4 extends our graph based approach to find time lagged relationships in climate data. Time series data in climate are often characterized by a delayed relationship between two variables, for example precipitation and temperature anomalies occurring at a place might also occur at another place after some time. These lagged relations

generally signify the time lag between the cause and the effect or the spread of a common cause and are important to study and understand as they can aid in prediction. In this chapter, we present a general framework for finding all pairs of lagged positive and negative relations that can exist in a given spatio-temporal dataset. We use a graph based approach based upon the concept of *shared reciprocal nearest neighbor* to generate cluster pairs of locations sharing similar or opposing behavior for every time lag. Our framework can be generalized to extract multivariate lagged relationships across different variables thus can be used to understand the lagged response of one variable on another. We show the utility of our approach by extracting some of the known delayed relationships like the Madden Julian Oscillation (MJO) and the Pacific North American (PNA) pattern at different lags using the sea level pressure dataset provided by the NCEP/NCAR. Our approach can be broadly applied to other problems in spatio-temporal domain to extract lagged relationships.

Chapter 5 presents a novel method for testing the statistical significance of the dipoles. So far, we have seen that systematic approaches for dipole detection generate a large number of candidate dipoles, but there exists no method to evaluate the significance of the candidate teleconnections. One of the most important challenges in addressing significance testing in a spatio-temporal context is how to address the spatial and temporal dependencies that show up as high autocorrelation. We present a novel approach that uses the wild bootstrap to capture the spatio-temporal dependencies, in the special use case of teleconnections in climate data. Our approach to find the statistical significance takes into account the autocorrelation, the seasonality and the trend in the time series over a period of time. This framework is applicable to other problems in spatio-temporal data mining to assess the significance of the patterns.

Chapter 6 discusses some of the challenges in anomaly construction in climate datasets. Anomaly construction is an important preprocessing step applied to most of times to before analyzing the climate datasets. Often, climate scientists only take 30 years as

a reference interval and construct anomalies with respect to that interval. There are several important results and implications derived from the anomalies constructed using a short reference base. The choice of the base significantly impacts the patterns that can be discovered from it and some really important climate phenomenons are computed using a fixed base. We show how an arbitrary choice of base can skew the results and lead to favorable outcome which might not necessarily be true. We examine four simple criterions for base selection and empirically evaluate the differences in them. Our empirical evaluation of the different measures reveals that the z-score measure is quite different from the other measures like mean, median and jackknife. We further study the impact of using different base period to highlight that the outcome of further analysis can be sensitive to the choice of the base. We present a case study of the Sahel region to show that the dipole in precipitation in the region moves around and even disappears with the choice of a different base. Finally, we propose a generalized model for base selection which uses Monte-Carlo based sampling methods to minimize the expected variance of the underlying datasets. Our research can be especially instructive to climate scientists in helping them construct a generalized anomaly that does not create a bias in their analysis. Further, other researchers in temporal domain can also benefit from our work and it will enable them to choose a bias-free base.

Chapter 7 discusses the conclusions and future work of the thesis.

Chapter 2

Dataset and Preprocessing

In this chapter, we describe the dataset and the preprocessing that we used for our analysis.

2.1 Dataset

We use sea level pressure (SLP) data to find the dipoles because most of the important climate indices are based upon pressure variability. Apart from SLP, we have also studied dipoles in other variables like precipitation, sea surface temperature (SST) and geopotential height (HGT). We also use the variables surface temperature and precipitation to understand the impact of the discovered dipoles globally.

We analyze three *reanalysis* datasets. Reanalysis projects create gridded datasets for all the locations on the globe by assimilating remote and in situ sensor measurements using a numerical climate model to achieve physical consistency and interpolation for global coverage. In the absence of a global data of observations, the reanalysis datasets are considered the best available proxy for global observations. Such datasets are produced by modeling groups around the world, including the NCEP/NCAR Reanalysis project [43],

the European Reanlysis project [72], and the Japanese Reanalysis project [57]. Table 2.1 shows the summary of the details of the three reanalysis datasets. We present most of our results using the NCEP data as it is the longest reanalysis dataset. The NCEP data spans 1948—present . The NCEP/NCAR reanalysis is provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA [43], available for public download at [6]. The ERA project contains the global atmospheric reanalysis data produced by European Centre for Medium-Range Weather Forecasts (ECMWF). The ERA-40 dataset covers the time period mid-1957 to mid-2002 and is available for public download at [5]. The Japanese Reanalysis project (JRA) has dara assimilated from 1979 to present and is based on the collaboration between Japan Meteorological Agency (JMA) and Central Research Institute of Electric Power Industry (CRIEPI). It is available at [7]. All of the three datasets are present in the 2.5 degree resolution using which there are about 10512 grid points on the Earth.

Model & Institute	**Grid Size**	**Available years**
NCEP/NCAR	144 x 73	1948 - present
JRA	320 x 160	1979-present
ERA-40	288 x 145	1958 - 2002

Table 2.1: Details of the Reanalysis datasets used for our study.

We use two temporal resolutions, i.e. monthly and daily for most of our analysis. We use monthly mean values for the 60 years of data (corresponding to 720 monthly values) for most of our analysis. For the lag analysis mentioned in Chapter 4 we use the daily resolution data. We use the data from the past 30 years (1982-2011) for our lag experiments as it is more reliable due to the availability of satellites. We first aggregate the daily resolution data into five day means to create pentads. This is also done in order to reduce the impact of noise in daily resolution. For a leap year, we consider 6 day mean for 12th pentad (which contains Feb 29). Overall, we have 73 pentads corresponding to the 365 days in a year and for the 30 year data we have overall 2190 pentads.

2.2 Pre-processing

Climate datasets are generally characterized by annual seasonality and long term trends which form the most dominant signal in the data. Before we do further analysis on the data, we need to remove the seasonal and trend patterns as described in the subsections below.

2.2.1 Seasonality Removal

An important component of Earth Science data is the seasonal variation in the time series. The change in seasons brings about annual changes in the climate of the Earth such as increase in temperature in the summer season and decrease in temperature in the winter season. The seasonality component is the most dominant component in the Earth science data. For example, consider the time series of monthly values of air temperature at Minneapolis from 1948-1968 as shown in the Figure 2.1. From the figure, we see that there is a very strong annual cycle in the data. The peaks and valleys in the data correspond to the summer and winter season respectively and occur every year. The seasonal patterns even though important are generally known and hence uninteresting to study. Mostly, scientists are interested in finding non-seasonal patterns and long term variations in the data. As a result of the effect of seasonal patterns, other signals in the data like long term decadal oscillations, trends, etc. are suppressed and hence it is necessary to remove them. Climate scientists usually aim at studying deviations beyond the normal in the data.

In order to take care of the seasonality, we construct anomaly time series from the raw data by removing the monthly mean values of the data. This is done as follows:

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

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

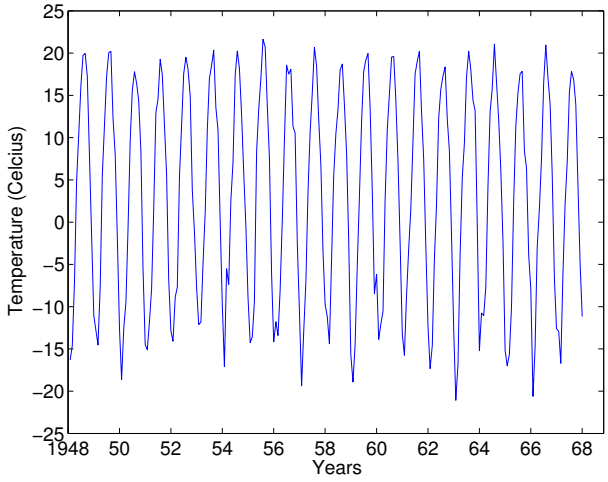


Figure 2.1: The figure shows the monthly mean air temperature at Minneapolis for a 20 year period.

In this equation, start and end represent the start and end years to consider for the mean and define the base for computing the mean for subtraction (for example 1948 and 2009 for the NCEP data). 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.

Although, removal of monthly means is the most popularly used approach to construct anomalies, they can also be constructed by alternate measures like using the z-score. Further, only a part of the data can be used construct the mean values and thus the anomalies. The several issues in anomaly construction and their impact on the discovered dipoles is further discussed in Chapter 6.

In the case of pentad data, we construct anomaly time series from the raw data by removing the mean values of the data for every pentad. This is done as follows:

$$\mu_p = \frac{1}{end - start + 1} \sum_{y=start}^{end} x_y(p), \forall_{p \in \{1..73\}}$$

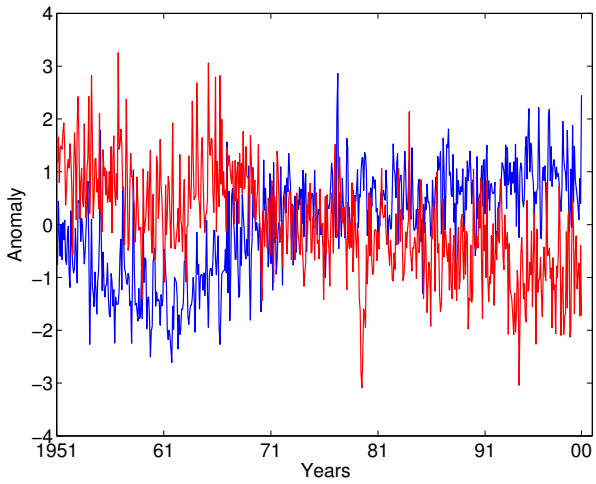
$$x_y(p) = x_y(p) - \mu_p$$

In this equation, start and end represent the start and end years to consider for the mean. μ_p is the mean of the pentad p and $x_y(p)$ represents the value of the raw data for the pentad p and year y.

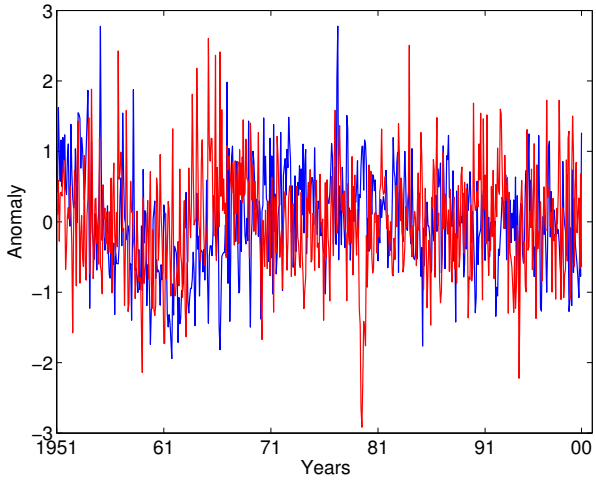
2.2.2 Detrending

Another important aspect of climate data is the presence of long term increasing or decreasing trends in the data. If there are two regions with strong trends in the opposite direction then it could result in spurious negative correlations between the two regions. For, example consider the pressure anomaly time series at two locations as shown in the figure 2.2(a) in the NCEP data during the time period 1951-2000. From the figure, we see that the two locations have trends in the opposite direction and hence as a result have a high negative correlation between their anomalies (-0.46 in this case).

A possible approach to handle the trends in climate data that is widely used by climate scientists is to detrend the data before any possible analysis. If we detrend the data in the example above, we no longer see the high negative correlation between the two time series as shown by the figure 2.2(b). For most of the results in the thesis, we use detrended data to present our results. But, we acknowledge that detrending of non-stationary time series data itself has several issues and may result in removing connections or adding spurious ones, which might require a detailed investigation. Results may also depend upon the nature of trends, whether unit roots are present or not, and the nature of possible co-integrating relations, see Engle and Granger [23, 34] for further details. However, further detrending analysis is beyond the scope of this thesis.



(a) Raw anomaly time series at the two locations showing trends in the opposite direction and a possible dipole having a correlation -0.4616



(b) Anomaly time series of the two locations after detrending has a correlation -0.0470

Figure 2.2: Trends in climate data

Chapter 3

Data Guided Discovery of Dipoles in Climate Data [1]

3.1 Introduction

As mentioned earlier, dipoles are of great importance in understanding climate variability. Table 3.1 lists some dipoles that are well known to climate researchers. These dipoles have been discovered by observation, e.g., SOI and NAO, or by Empirical Orthogonal Function (EOF) analysis [75], e.g., AO and AAO. However, all these discoveries have required considerable research and insight on the part of the domain experts involved. Also, some of these dipoles are defined by static locations (e.g., SO, NAO) but the underlying phenomenon itself is dynamic. Many of the dipoles (e.g., SO, NAO) have been discovered by examining the local data and correlation maps at specific locations. Such manual discovery can miss many dipoles. Ever since the satellite data became widely

[1] **This chapter is based on the work [49] published in the 11th SIAM International Conference on Data Mining, (SDM 2011), [47] published in the proceedings of the Conference on Intelligent Data Understanding, CIDU 2011 where it won the `best student paper award` and [48] to appear in the Statistical Analysis and Data Mining Journal 2013.**

available in the early 1970s, pattern analysis techniques such as the EOF [75] have been used to identify individual dipoles and the climate indices over a limited region, such as Arctic Oscillation (AO) index. However, there are several limitations associated with EOF and other types of eigenvector analysis; namely, it only finds a few of the strongest signals and the physical interpretation of such signals can be difficult [18]. Because of the amount of effort involved and the possibility of missing indices, an automated approach to climate index discovery could be quite useful.

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

Table 3.1: List of major pressure based dipoles.

One of the first attempts in this direction was by Steinbach et al. [61]. They constructed a network using climate data in which each node in the graph represented a region on the Earth and an edge between a pair of nodes represented pairwise correlation between the anomaly time series of the corresponding regions. They used a shared nearest neighbor (SNN) [24] clustering approach to find clusters in the climate graph

that has edges corresponding to only positive correlations. Many of the centroid of the clusters corresponded to known climate indices. Further some pairs of discovered clusters also showed high correlation with many SLP based climate indices defined as dipoles. Tsonis et al. [68] were the first ones to study climate graphs as complex networks. They constructed similar networks using absolute correlation between pairs of nodes and showed that in the tropics, the network had a 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 super nodes in the scale-free network corresponded to major climate indices such as NAO and PNA [69, 70]. Other researchers have also studied climate graphs as complex networks for examining the structure of the climate system (Donges et al. [19]), analyzing hurricane activity (Forgarty et al. [22]), and finding communities in climate networks and how they correspond to known climate patterns (Steinhaeuser et al. [62, 63]).

In our work, we present a comprehensive technique to systematically find climate indices that are dipoles from the climate data. In the other approaches, negative correlations have been ignored [61] or only absolute values of correlations have been considered [69, 19, 62]. However, as we show in Section 3.3, negative correlations are the 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 both positive and negative correlations is not appropriate since the negative correlations are usually weaker and many nearby locations have high positive correlation. We also study the change in the climate indices over time unlike [61]. 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. Our approach discovers dipoles using a Shared Reciprocal Nearest Neighbor (SRNN) algorithm and has a number of advantages. The approach allows us to detect all dipoles represented in an individual global

dataset within the selected time frame and to determine their individual strengths. It makes it possible to discover new dipoles that may not have been seen. It enables tracking the movements of these dipoles and studying their interactions due to the seasonal cycles and other changes in local climate in a much more systematic way.

3.1.1 Our Contributions

Since teleconnections are crucial in the understanding of the climate system, there is a pressing need to better understand the behavior and interactions of these atmospheric processes as well as to capture them precisely. Our systematic graph based approach to find the teleconnections in climate data is an attempt in that direction. We highlight the main contributions of this chapter as follows:

1. We show the importance of treating negative correlations differently than positive correlations unlike [69, 61, 62, 19]. We present a novel graph based approach to find dipoles in the climate data that overcomes the shortcomings of the previous approaches [61].

2. An important utility of the dynamic dipoles defined using our approach is that they have a much higher correlation with temperature anomalies as compared to the static indices used by the climate scientists, for e.g., the ones available at the Climate Prediction Center (CPC) [1].

3. Our approach allows us to have a single snapshot of all the dipoles on the globe. This ability enables us to discover new dipoles and comprehensively study the behavior, their interaction, and the movement of the various dipoles in a more precise manner.

4. Our approach of dipole discovery is able to find a pair of regions (dipoles) that highly correlate with the dipoles defined earlier using EOF analysis and have a number of well known limitations [18]. One of the major limitations of the EOF

analysis is the orthogonality constraint of the principal components because of which it is difficult to physically interpret the discovered patterns. Further the centers of action using the EOF analysis do not need to correspond to the actual physical process. Also the EOF analysis is very sensitive to the choice of spatial and time domain.

3.2 Network Construction

We first preprocess the raw data to remove the seasonality and trends as mentioned in Chapter 2. Once we get the detrended anomaly series from the raw data, we construct a complete graph out of the data using the approach used earlier by [69, 19, 62, 61, 49] by taking the pairwise correlation between the anomaly time series of all pairs of locations on the Earth. However, as compared to these approaches, we do not threshold the networks by taking the absolute value of correlation and then using a single threshold. The nodes in the graph represent locations on the Earth and the edges represent the correlation between the anomaly time series of the two locations on the Earth. Thus, after we get the anomaly values for every node, the networks are constructed by looking at the similarity values between the anomaly time series 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:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the mean of the two series X and Y.

The complete climate graph tends to have a lot of edges (for e.g. there are about 100 million edges in the NCEP data and most of them are uninteresting). A threshold is often used to eliminate spurious correlations in climate graphs [69, 19, 63]. For, e.g., Tsonis et al. [69] constructed networks using nodes on the globe and the edges of the

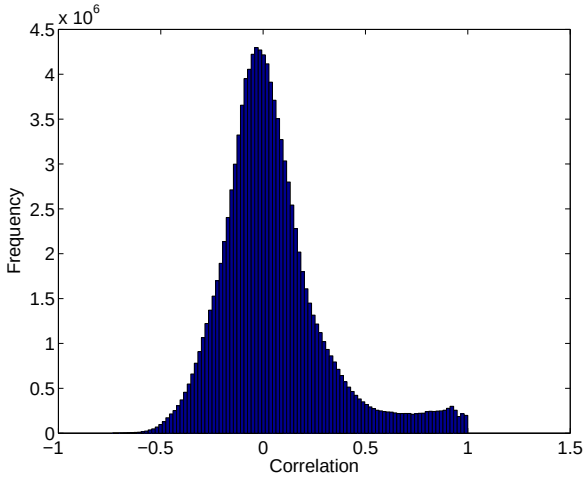


Figure 3.1: Distribution of positive and negative correlations in sea level pressure data spanning 1948-1967.

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. 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. Fig 3.1 shows the distribution of the correlations in the sea level pressure data. Due to autocorrelation in space, the positive correlations go 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 and that allows many spurious values to pass through. Fig. 3.2 shows the distribution of edges which are more 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. 3.1 and 3.2. This gives credence to their 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.

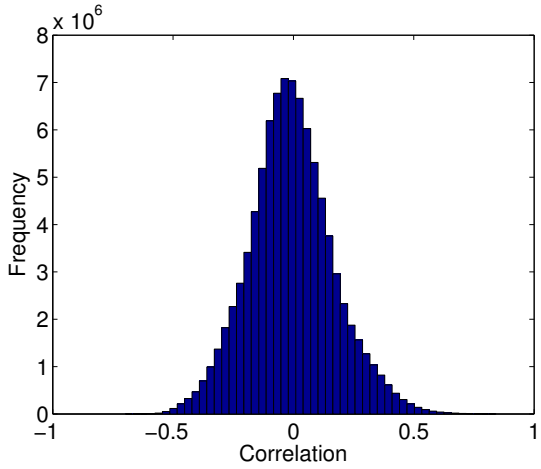


Figure 3.2: Distribution of correlation after filtering edges < 5000 km

Further, some dipoles are inherently weaker in nature as compared to the others. For example, the SO dipole is much weaker than the AO dipole and most of the negative edges spanning SO regions have correlation much weaker than the -0.4 threshold used to limit the number of edges. Fig. 3.3 shows the heat map of the distribution of negative edges for all locations on the Earth. The heat map shows is constructed by computing the area weighted sum of degrees of all the negative edges on the Earth as described in [69]. The map shows that the northern hemisphere has a much higher density of negative correlations as compared to other locations on the Earth.

3.3 Our Approach

We begin with a formal definition of a dipole.

Definition 1. *A dipole is defined as a pair of regions whose locations have strong negative correlation with locations in the other region and strong positive correlation*

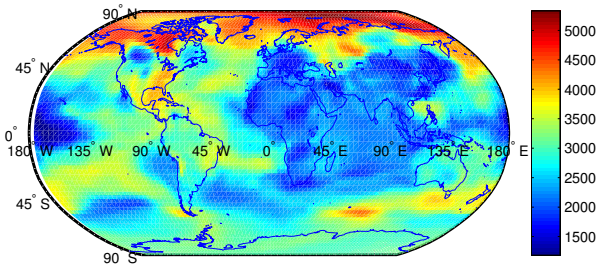


Figure 3.3: Heat map showing the area weighted distribution of negative edges around the Earth in the sea level pressure data.

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. Thus a careful balance needs to be struck. To find dipoles we use a clustering approach that groups together locations on the globe that are similar in terms of i) the locations to which they are most strongly negatively correlated and ii) locations to which they are positively correlated. This first requirement is motivated by the centrality of negative correlation in the definition of dipole, while the second helps to produce spatially contiguous clusters since nearby locations tend to have positive correlations. These clusters can serve as the ends of dipoles and the set of all possible pairs are further evaluated to yield candidate dipoles. Since regions involving dipoles can be of different size, shape and strength, we use a clustering scheme based on the shared nearest neighbor concept that is particularly effective in addressing such requirements [25]. In this section, we define our approach based upon shared reciprocal neighbors to find the dipoles in climate data.

We model the climate data as an undirected weighted graph $G^C = (V^C, E^C)$, where V^C is the set of nodes representing grid locations on the Earth and E^C is the set of undirected edges between these locations. As described earlier, the edge weight represents the correlation between the anomaly time series of the locations, such that,

positive edge weight between two locations indicates that they are subject to a similar climatic phenomenon and negative edge weight indicates that they exhibit an out of phase climatic phenomenon.

We represent the undirected weighted graph as $G = (V, E)$, where $V = \{V_1, V_2, ..., V_N\}$ represent the N $(= |V|)$ vertices in the graph and E is a N X N matrix in which each entry $E_{i,j}$, $1 \leq i, j \leq N$, indicates the edge weight between vertices V_i and V_j. For every vertex V_i the set $S_i = \{V_{i_1}, V_{i_2}, ..., V_{i_{N-1}}\}$, where $i_1, i_2, ..., i_{N-1}$ is a permutation of the set $\{1, 2, ..., N\} \setminus i$, such that, $E_{i,i_1} \geq E_{i,i_2} \geq ... \geq E_{i,i_{N-1}}$. Let $KNN_i^+ = \{V_{i_1}, V_{i_2}, ..., V_{i_K}\}$ and $KNN_i^- = \{V_{i_{N-K}}, V_{i_{N-K+1}}, ..., V_{i_{N-1}}\}$. The edges from V_i to nodes in KNN_i are referred to as extremal edges.

Our algorithm to compute dipoles consists of four major steps, which can be summarized as follows:

- Step 1: Construction of reciprocal graph G^R from the climate data graph G^C. This involves forming the list of k nearest positive and negative neighbors of each object using the original similarity measure, where k is a parameter chosen by the user and considering only the edges that are reciprocal, i.e. which lie on each other's nearest neighbor lists.

- Step 2: Construction of the shared nearest neighbor graph for the positive and negative reciprocal edges (G^{SNN_-} and G^{SNN_+}). This is done by redefining the similarity of each pair of objects in terms of the number of their common (shared) nearest reciprocal neighbors.

- Step 3: Merging G^{SNN_-} and G^{SNN_+} to construct the G^{SRNN} graph. This step involves multiplying the edge weights in the two graphs G^{SNN_-} and G^{SNN_+}.

- Step 4: Finding dipoles using density based clustering on G^{SRNN}. This step enables us to find clusters in the G^{SRNN} graph. Finally, cluster pairs with negative correlation are declared as dipoles.

The further details of the algorithm are mentioned as follows.

3.3.1 STEP 1: Construction of the Reciprocal Graph

We begin by considering the original graph $G^C = (V^C, E^C)$ as described earlier. We construct the reciprocal graph $G^R = (V^C, E^R)$, where $E^R \subseteq E^C$ as follows:

$$E^R_{i,j} = \begin{cases} 1 & \text{if } V^C_i \in KNN^+_j \wedge V^C_j \in KNN^+_i \\ -1 & \text{if } V^C_i \in KNN^-_j \wedge V^C_j \in KNN^-_i \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The main idea behind reciprocal graph is to pick the K highest positively and negatively correlated locations (*extremal* set) corresponding to a given location and then consider an edge between two locations if they appear in each other's extremal set. From the definition of dipoles, we know that any two regions that actually form dipoles are in each other's negative extremal set and the nodes within a region are in their positive extremal set. The benefits of computing the reciprocal graph are several: First, it reduces the size of the original graph drastically (asymptotic upper bound of reduction is $\theta(N/K)$ but in practice it is much more); Second, it filters noise (such as anomalous locations or regions, weakly correlated locations). Note that building the reciprocal graph is essential to eliminate spurious inter-connections between the locations. *If we set $G^R = G^C$ (by setting $K = N$) then the overall algorithm yields a large number of spurious dipoles.* The concept of reciprocity holds more importance in negative correlations than in positive correlations because nearby objects are very similar and hence reciprocity exists by default for positive correlations, but for negative correlations, reciprocity is much more meaningful and helps in weeding out spurious negative correlations.

To illustrate the importance of reciprocity in finding dipoles, we consider an example of the SO dipole. As we described earlier, the SOI is computed from examining the sea level pressure anomalies at the two locations, Tahiti and Darwin. Consider, the location

Tahiti. The KNN^- and the reciprocal edges coming out from Tahiti for the time period 1948-1967 are shown in the Figure 3.4. From the figure, we see that Tahiti has many edges going to the North pole in KNN^-, however only the ones going to Darwin in Australia survive in the reciprocal graph. Further, if we examine the correlation of all points on the Earth with Tahiti, the single point correlation map is as shown in the Figure 3.5. Recall, that a single point correlation map used earlier by climate scientists to find dipoles is constructed by plotting the correlation of all the other grid points on the Earth with respect to the single point. Thus we can see that there are many negative edges coming out from Tahiti going towards the Northern hemisphere and having a correlation < -0.2.

Similarly, consider the location, Darwin. The KNN^- and the reciprocal edges coming out from Darwin for the time period 1948-1967 are shown in the Figure 3.6. We see that from the top 50 negative edges coming out of Darwin some go to the South pole. However, only the edges near Tahiti are reciprocal. Further, if we examine the single point correlation map of Darwin as shown in the Figure 3.7 we see that the negative correlations < -0.2 are spread throughout the world. Similarly, if we examine other locations on the Earth, we see that reciprocity is helpful in reducing arbitrarily correlated regions connected by negative correlations.

3.3.2 STEP 2: Construction of Shared Nearest Neighbor graph (G^{SNN_-} and G^{SNN_+})

The reciprocal graph G^R retains the edges which are mutually extreme (highly positive or negative) between all the location pairs. G^R essentially captures the dipole regions and their interconnections yet the extraction of these regions requires us to cluster the graph nodes into a set of regions. Additionally, the clustering helps us to identify spurious regions that result due to a small number of spurious extremal edges (making G^R robust to any choice of $K << N$). We propose a variant of the SNN algorithm [61]

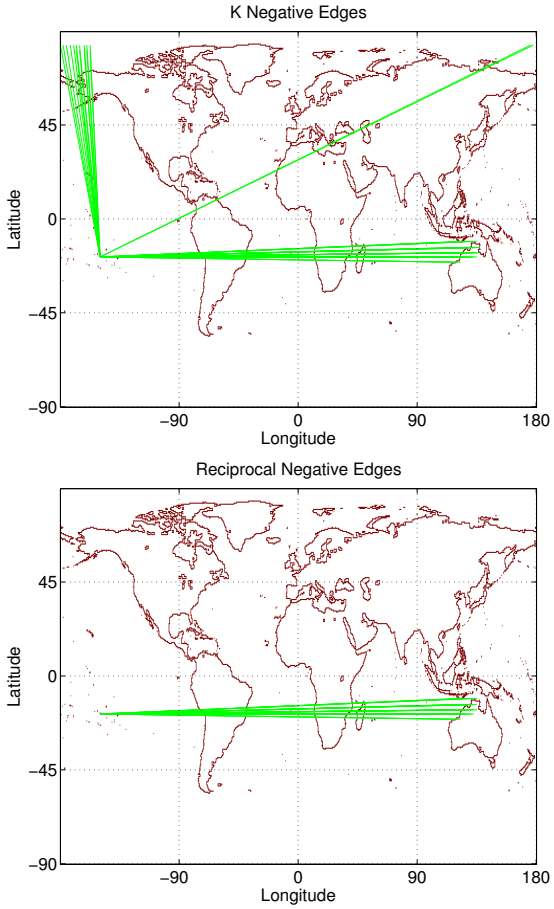


Figure 3.4: KNN negative and Reciprocal negative edges from Tahiti using K=50 in the time period 1948-1967

for clustering the reciprocal graph. The main idea of SNN algorithm is to form groups based on how many shared neighbors two nodes have in the graph. It is important to note that the SNN algorithm alone cannot extract the most precise dipoles, as suggested by prior work [49]. This motivates us to propose the following variant of SNN algorithm. We construct two graphs $G^{SNN_+} = (V^C, E^{SNN_+})$ and $G^{SNN_-} = (V^C, E^{SNN_-})$ by running SNN algorithm on positive and negative edges of G^R, respectively. More formally, the edge weights of the two graphs are estimated as follows:

$$E_{i,j}^{SNN_+} = |\{k : V_k^C \in V^C \wedge E_{i,k}^R = 1\} \cap \{k' : V_{k'}^C \in V^C \wedge E_{i,k'}^R = 1\}| \qquad (3.2)$$

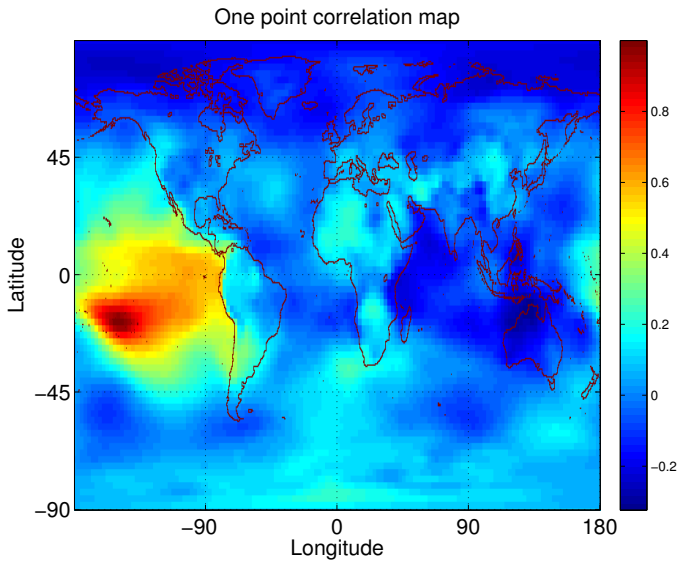


Figure 3.5: Correlation map showing the correlation of all the grid points on Earth with Tahiti for the time period 1948-1967

$$E_{i,j}^{SNN_-} = |\{k : V_k^C \in V^C \wedge E_{i,k}^R = -1\} \cap \{k' : V_{k'}^C \in V^C \wedge E_{i,k'}^R = -1\}| \tag{3.3}$$

Equations 3.2 and 3.3 estimate the number of shared neighbors two nodes have and considers the edge weight as a function of the number of shared neighbors. The motivation behind two separate graphs is that a node can have two types of neighbors in G^R; those with $+1$ edge weights and others with -1 edge weight. As a result, these neighbors need to be counted separately. It is crucial to treat the two types of edges separately because otherwise a single application of SNN algorithm using only positive correlations [61] would only look for regions that share similar climatic behavior even when they do not participate in the dipole phenomenon. On the other hand a single application of SNN using only negative correlation would result in disconnected regions as the edge weight only takes into account the number of negative neighbors shared and a region could have negative correlations going to two distant locations (perhaps a part of two different dipoles). The next step in which we combine the above two graphs makes it clear why a single application of SNN would not yield qualitatively and quantitatively

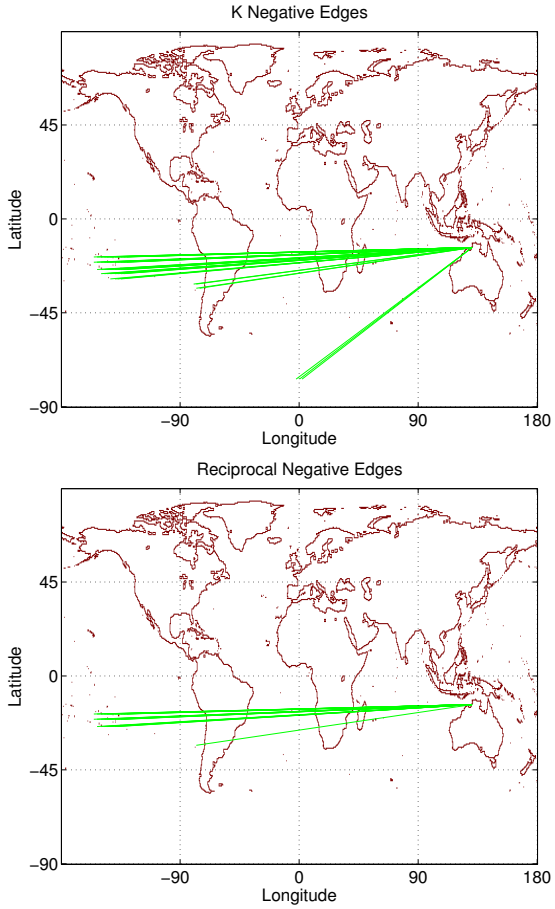


Figure 3.6: KNN negative and Reciprocal negative edges from Darwin using K=50 in the time period 1948-1967

reliable dipoles. In equation 3.2 and 3.3, we simply count the number of shared neighbors between all location pairs. Instead, we can also compute a weighted sum, where the weights take into account the ranks of the shared nearest neighbors from the two lists (see [41]). This idea allows us to compute the edge weight as a weighted sum of the reciprocal links shared between the nearest neighbor list of the two nodes. The weight is computed by taking the mean of the ranked order of the reciprocal links in the two neighbor lists. The weighted version performs slightly better than the counting version and we use it throughout the chapter to compute our results.

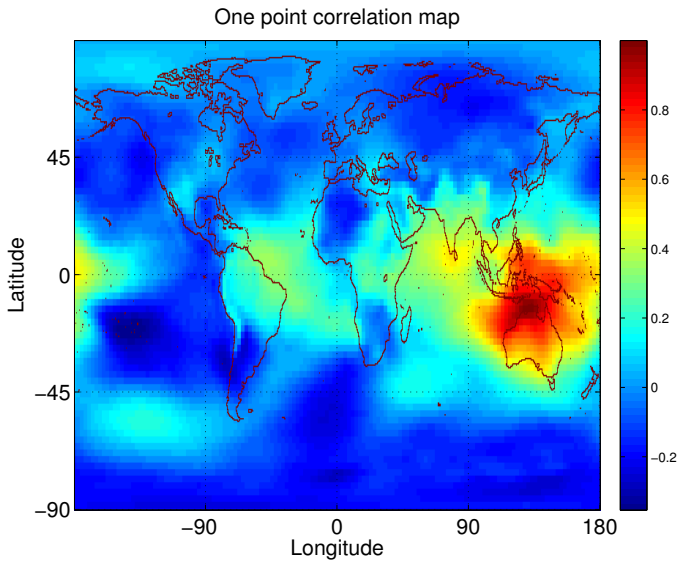


Figure 3.7: KNN negative and Reciprocal negative edges from Darwin using K=50 in the time period 1948-1967

Overall, nodes with high edge weights in G^{SNN_+} provide two types of information. First, the two locations, corresponding to the nodes, share positive correlation in their climate and this correlation is high for both the nodes (guaranteed by the construction of G^R). Second, these nodes are part of a cluster where this positive climate phenomenon is maximal (counting of positive neighbors, equation 3.2). In practice, this cluster corresponds to spatially co-located places on Earth. Similarly, G^{SNN_-} gives us a sense of which negative regions these nodes associate with. It is possible for two nodes to have high edge weight in one graph and yet have a low or 0 edge weight in other graph; forming the basis of the next step of our algorithm.

3.3.3 STEP 3: Merging of G^{SNN_-} and G^{SNN_+} to construct G^{SRNN} graph

The two graphs G^{SNN_+} and G^{SNN_-} form graph components or cliques with a higher inter clustering coefficient than intra clustering coefficient. It is possible for two nodes to have high edge weight in one graph yet a very low edge weight in another. To illustrate

this consider two geographically close points; one inside one end of a dipole (say x) and other outside it (say y). Indeed, x and y would share high positive correlation on climate variables (such as air pressure, temperature) due to spatial autocorrelation. As a result it is possible for the two nodes to have moderate to high $E_{x,y}^{SNN_+}$. On the other hand, the point y would not have very high negative correlation with the other end of the dipole region corresponding to x, as it is not a part of the dipole. It is also possible that two regions have a high edge weight $E_{x,y}^{SNN_-}$ and a low edge weight $E_{x,y}^{SNN_+}$ which indicates that the two locations are spatially distant and cannot be a part of the same end of the dipole. Hence a single application of SNN in step 2 does not yield good results (because then both point x and y are claimed to be part of the same dipole region even though they are disconnected). The example presents an intuitive justification of our merging criteria: multiply the edge weight of G^{SNN_-} and G^{SNN_+} to form $G^{SRNN} = (V^C, E^{SRNN})$. More formally,

$$E_{i,j}^{SRNN} = E_{i,j}^{SNN_-} * E_{i,j}^{SNN_+} \tag{3.4}$$

Note that the only parameter that our algorithm uses is K (defined in Section 3.3.1) to compute the extremal edges. A large choice of K would result in many spurious connections, while a small choice of K would reveal only the most significant regions within the dipoles. The merging criteria chosen above makes the dipole discovery less sensitive to the choice of K (and robust for moderate values of K). The robustness can be seen in Figure 3.8 which shows that increasing the value of K only increases the size of dipole regions but does not produce any spurious region that is a dipole. Steps 1, 2 and 3 of the algorithm are illustrated by the example presented in Figure 3.9.

3.3.4 STEP 4: Finding dipoles using density based clustering on G^{SRNN}

Figure 3.8 shows the density plot of locations, where density for a location is defined as the weighted degree (sum of edge weights) of that location. From the visual inspection

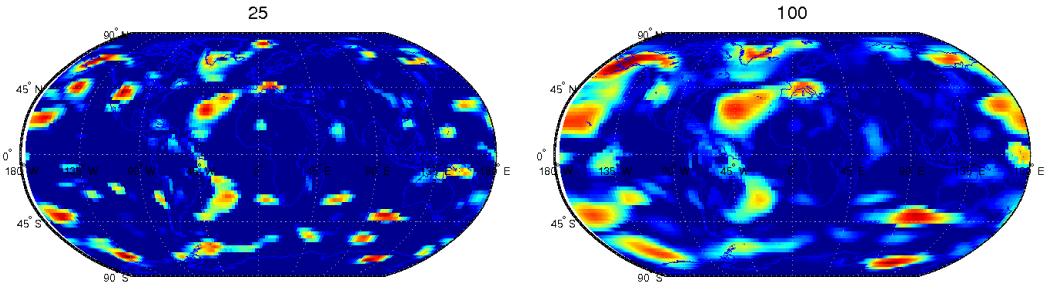


Figure 3.8: Dipoles discovered using our algorithm for $K = 25, 100$ (density plot of sum edge weight of nodes in G^{SRNN}).

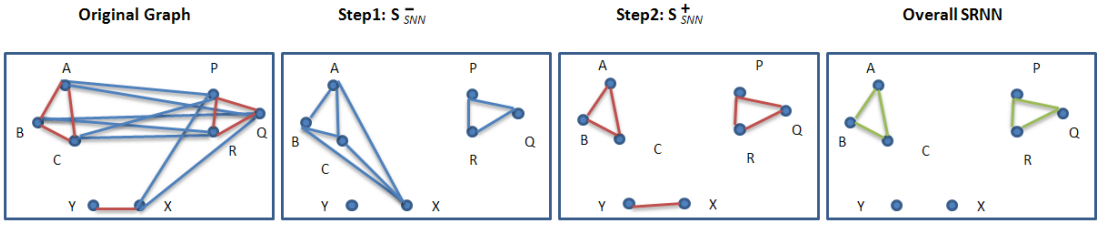


Figure 3.9: Illustration of Steps 1,2 and 3: Blue edges are reciprocal negative and red edges are reciprocal positive. First box on the left shows the original reciprocal graph. Second box shows G^{SNN-}. Note that edges A, B and C get connected as they share negative neighbors P, Q & R. Also the node X is connected to A, B and C since X shares P and Q with them. Third box has G^{SNN+} and all the nearby nodes get connected. Fourth box shows G^{SRNN} with overall similarity defined as the product of the two. This helps in separating the node X from nodes A, B & C as it does not have an edge in G^{SNN+} even though it has an edge in G^{SNN-}.

of Figure 3.8, it is clear that the spatially contiguous red regions form a single dipole region. These regions can be extracted using a spatial clustering algorithm over the latitude, longitude and the intensity of the locations. We considered several clustering algorithms such as KMeans and Gaussian Mixture Modeling (GMM). There are several limitations with these algorithms: 1) The number of clusters are not known, 2) These algorithms are sensitive to initial choice of cluster centroids, 3) These algorithms are computationally expensive, GMM in particular, 4) The detected clusters are circular in shape (KMeans and GMM). In order to eliminate these issues, we propose a spatial

Algorithm 1 Local attractor based clustering.

Let, $DS_{i,j}$ be geographical distance between locations i and j.
Let, $CORR_{i,j}$ be anomaly correlation between locations i and j.
Let, $D_i = \sum_{j=1}^{N} E_{i,j}^{SRNN},\ \forall i \in \{1, 2, ..., N\}$ (location density).
Let, $A = \{1, 2, ..., N\}$ (local attractor set - initially set to all locations on Earth).
Let $LA_i = i$ (local attractor of all nodes are set to themselves initially).
repeat
  for $i \in A$ **do**
    $j = \arg\min_k (DS_{i,k} : k \in A \wedge k \neq i)$
    if $DS_{i,j} <$ Distance-Thresh AND $CORR_{i,j} >$ Correlation-Thresh **then**
      if $D_i \geq D_j$ **then**
        $A = A \setminus j$ {Eliminate j from attractor set as i is the attractor of j}
        $LA_z = i, \forall z \in \{1, 2, ..., N\} \wedge LA_z = j$
      else
        $A = A \setminus i$ {Eliminate i from attractor set as j is the attractor of i}
        $LA_z = j, \forall z \in \{1, 2, ..., N\} \wedge LA_z = i$
      end if
    end if
  end for
until convergence {If A doesn't change in two successive iterations, then algorithm converges}

clustering algorithm which is similar to the Denclue algorithm [39] to find clusters in data based upon local density attractors. Specifically, we use latitude and longitude to determine the local attractor (point with the highest density) in the neighborhood of locations. The algorithm proceeds by attaching every node in the graph to its local attractor by moving in the direction of increase in density. In the next step, we hierarchically merge attractors that are very close and have a positive correlation in order to remove extraneous attractors. The details of the algorithm are presented in Algorithm 3.

The locations that remain in A form the cluster centers and they become the attractor of all the points in their neighborhood (as assigned in LA in algorithm 3). The points that are attracted to a given cluster center are a part of the same cluster. Next we compute the correlation of every cluster pair to find the dipoles from the clusters. After this we label all the cluster pairs having a *sufficient* negative correlation as a dipole, where by sufficient we mean a user provided correlation threshold. However, we can

ignore the correlation threshold and label all the cluster pairs having a correlation < 0 as dipoles. The significance of these cluster pairs can later on be ranked on the basis of their strength or impact on temperature/pressure anomalies as we see later in the Section. 3.4.1.

3.3.5 Algorithm Features

The proposed algorithm runs in $O(N^2)$ space and time. Moreover, our approach can be implemented quite efficiently. Our preliminary approach[49] takes more than 1 day to run but the proposed approach runs in less than 20 minutes for the NCEP/NCAR Reanalysis SLP dataset at a 2.5° resolution. The approach has only one parameter K (and not very sensitive to its choice as well). Additionally, we can find weaker dipoles as well.

3.4 Experiments and Results

In this section, we present the results of using our approach for discovering dipoles on the reanalysis datasets. The following subsections include the details of our experimental evaluation.

3.4.1 Evaluation of Dipoles

The goal of our experimental evaluation of the dipoles is to assess the ability of our approach to reproduce known dipole indices. The Earth System Research Laboratory [3] has a comprehensive listing of climate indices used in climate science which is also available at the NOAA Climate Prediction Center (CPC) [1]. We test the reproducibility of dipoles generated using our algorithm by testing the correlation between the static index generated from the CPC and our dynamic index as described further. Strong correlation

with known indices indicates that the generated dipoles are a good representative of the known climate indices and represent the same phenomenon.

We construct networks from the NCEP/NCAR data using anomaly time series for a period of 20 years with a sliding window of 5 years so as to study the gradual change in the climate networks. Thus, for the 60 years of NCEP/NCAR data (1948 - 2007) we had 9 networks spanning 20 years each. We ran the dipole detection algorithm for each of the 9 periods. To test whether the right dipoles are being found using our methodology, we compute the correlation of our dynamic dipole indices with the static indices from the CPC website [1]. For every dipole belonging to a time period, we took the cluster pair constituting the dipole and computed the centroid at each end by taking the mean of the anomaly at those locations during that time period. We computed the difference between the two cluster centroids to create a time series which we call as the *dynamic index*. The dynamic index is then compared with all the known climate indices over that period using linear correlation. We kept track of the best correlation of the dynamic indices to the climate indices during the period and marked the dipole cluster that best matched each climate index as the best representative of the known index.

Table 3.2: Correlation of our dynamic indices with known climate indices ($K = 25$)

Start year	**SRNN**					
	SOI	NAO	AO	PNA	WP	AAO[1]
1948	0.8822	0.7899	0.87901	0.64396	0.68713	0
1953	0.86997	0.78128	0.84894	0.68794	0.66561	0
1958	0.91571	0.76131	0.86291	0.70227	0.67141	0
1963	0.89091	0.73662	0.86316	0.71546	0.67874	0
1968	0.92129	0.73022	0.83395	0.70765	0.7277	0
1973	0.92637	0.72998	0.87512	0.69905	0.69036	0
1978	0.93071	0.75813	0.84287	0.62118	0.71245	0
1983	0.9471	0.69094	0.85245	0.65527	0.71651	0.91348
1988	0.94116	0.7191	0.85224	0.63878	0.69695	0.92053

Table 3.2 shows the best correlation between the static and dynamic climate indices using $K = 25$ nearest neighbors. From the table, we see that we are able to match the existing indices with a very high precision. The SO is one of the most important dipole

and using our algorithm, we are able to capture it with an average precision of 0.88. We are also able to capture the other dipole indices NAO, AO, PNA, WP and AAO with a very high precision.

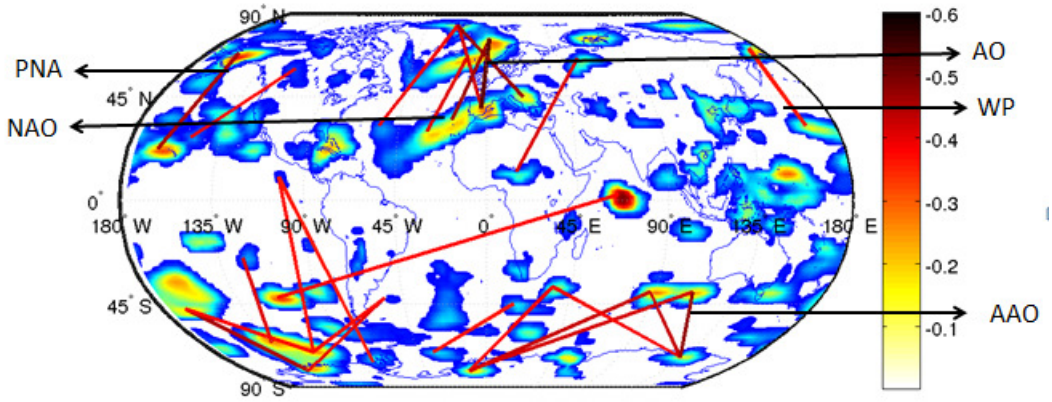
The biggest advantage of our approach is that it allows us to have a comprehensive view of the dipoles and their interactions. Figure 3.10 illustrates the dipole connections in the last network, which represents the time period from 1988-2007. Using the SRNN algorithm, at first we create a density map in which the density of a node is the sum of its edge weights in the G^{SRNN} graph. Using the local attractor algorithm as mentioned in Section 3.3, we find clusters in the graph G^{SRNN} by attaching every point to its local attractor. Finally, dipoles are cluster pairs having negative correlation between them. The figure is generated by connecting the local attractors of all the cluster pairs labeled as dipoles. The color of the edges represents the strength of the dipole. The color of the regions represents the density in the G^{SRNN} graph. Regions with red color represent the regions with the maximum density in the graph whereas regions with color blue represent the regions of low density. Regions not having any density in the G^{SRNN} graph are represented by the color white.

Figure 3.10(a) shows the dipole edges thresholded at -0.35. From the figure we see dipoles AO, NAO, PNA, WP and AAO. Most of the edges in the southern hemisphere are correlated with the AAO. Most of edges in the northern hemisphere near the NAO/AO region are correlated with the NAO/AO. The remaining unknown edges could either represent some unknown connections or some transient phenomenon and do not all appear in all the 9 network periods. When we reduce the threshold on the dipole edges to -0.3 as shown in the figure 3.10(b), we see the SO dipole. Upon reducing the threshold further to -0.25 (figure 3.10(c)), we see other dipole connections as well. Figure 3.10 shows that the NCEP/NCAR data reproduces the known climate patterns and indices during the last 20-year time range: the Northern Hemisphere pattern from west to east, the Pacific/North-America Pattern (PNA; which is actually a tripole) in

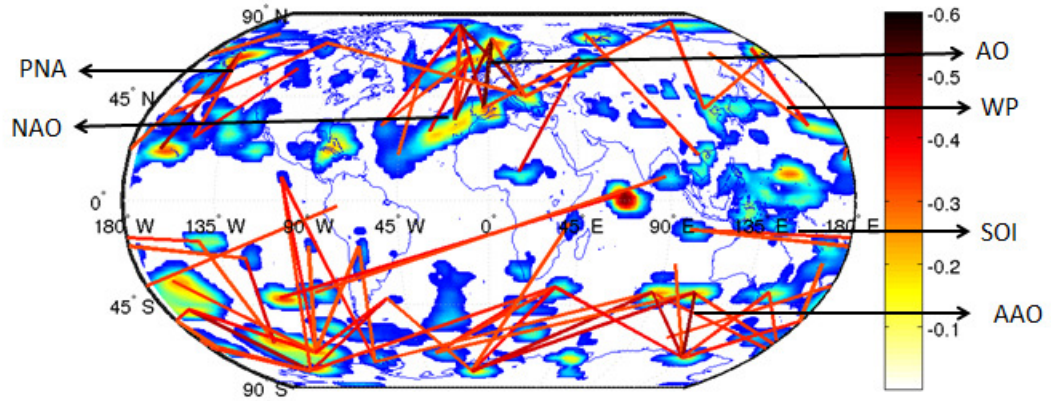
the top left corner, the NAO and AO in the central top, and the West Pacific oscillation (WP) on the top right. In the Southern Hemisphere and equatorial region, there are SOI connecting the west Pacific warm pool and eastern Pacific with a line from the central right eastward to the right end of the plot and showing up again in the far left to connect to the eastern Pacific. Also highlighted by the dense regions is the South Pacific Convergence Zone to the East of Australia crossing the map to the right and showing up on the left end in the southern Pacific.The dense regions over the Southern Ocean are a part of the AAO.

3.4.2 Impact on Global Temperature: Static vs Dynamic

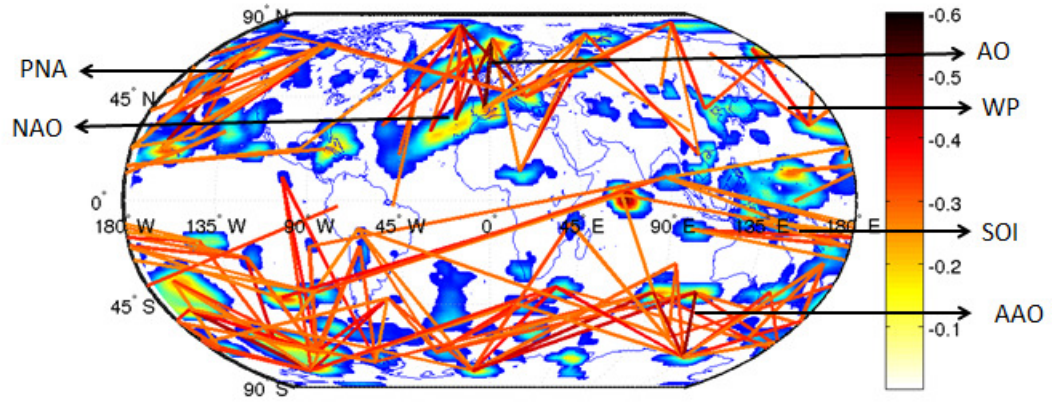
The real utility of data driven dipoles lies in the fact that they are able to capture temperature and precipitation anomalies related to these dipoles better than the static indices that are used in climate science(for e.g. those used at CPC). In order to show the utility of our dipoles, we take an area weighted correlation of temperature anomalies at every location on the globe with the static and dynamic indices. Only locations having an absolute correlation > 0.2 are considered to compute the aggregate area weighted impact in order to ignore noisy correlations. Further, the aggregate impact is divided by the total land area to generate a single number. We also varied the threshold to disregard noisy correlations by 0, 0.1, 0.3, 0.4, etc and saw a similar difference in between the static and dynamic indices.



(a) Dipoles thresholded at -0.35



(b) Dipoles thresholded at -0.3



(c) Dipoles thresholded at -0.25

Figure 3.10: Dipoles in SLP NCEP data from 1988-2007. The color background shows the SRNN density identifying the regions of high activity. The edges represent the dipole connection between two regions.

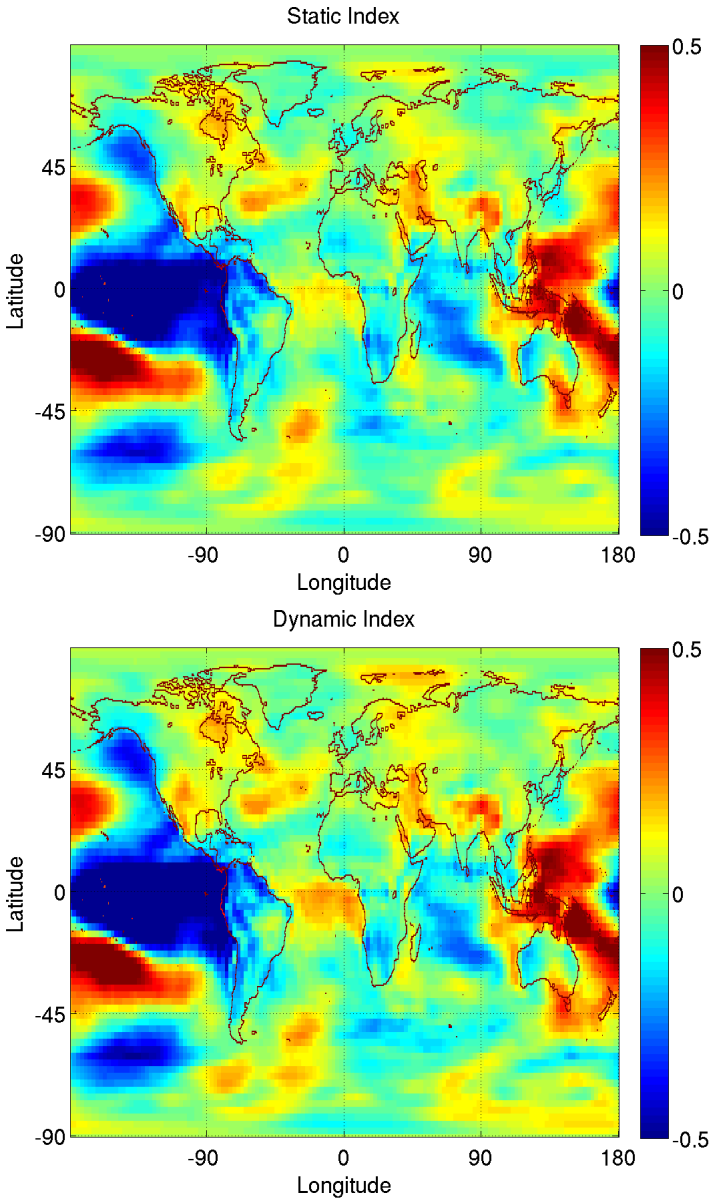


Figure 3.11: Impact on temperature for SO using static and dynamic index for the time period 1988-2008. The figure shows that both the static and the dynamic index generate a similar impact on temperature, however the dynamic index shows higher correlation than the static.

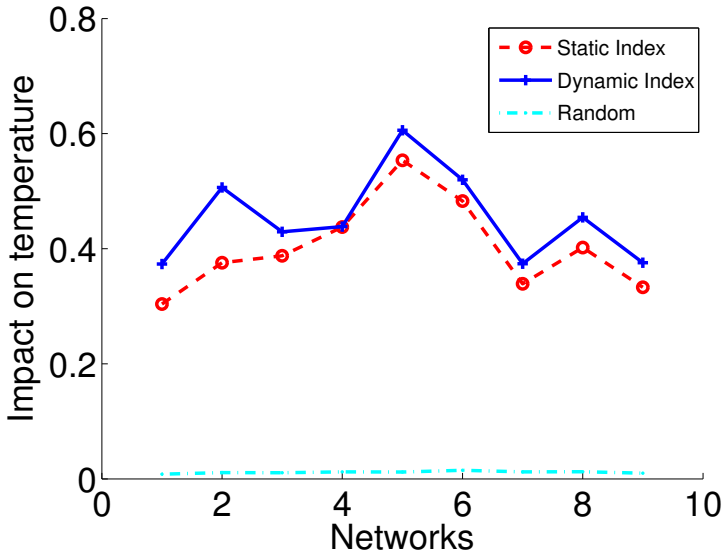


Figure 3.12: Aggregate area weighted impact on global temperature using static and dynamic SO. The figure shows that the dynamic index performs better than the static for all the 9 network periods.

Figure 3.11 shows the impact of both static and dynamic SO for the time period 1988-2008. From the figure, we see that the impact patterns in the two figures are very similar. This gives us confidence that the dynamic index the static index are representing the same phenomenon. Further, the dynamic index shows a somewhat better correlation with temperature anomalies as compared to the static index. Figure 3.12 shows the aggregate area weighted correlation of temperature anomalies using the static and dynamic SO dipoles for all the 9 time periods. The aggregate area weighted impact is higher using the dynamic index as compared to the static index for all the network time periods.

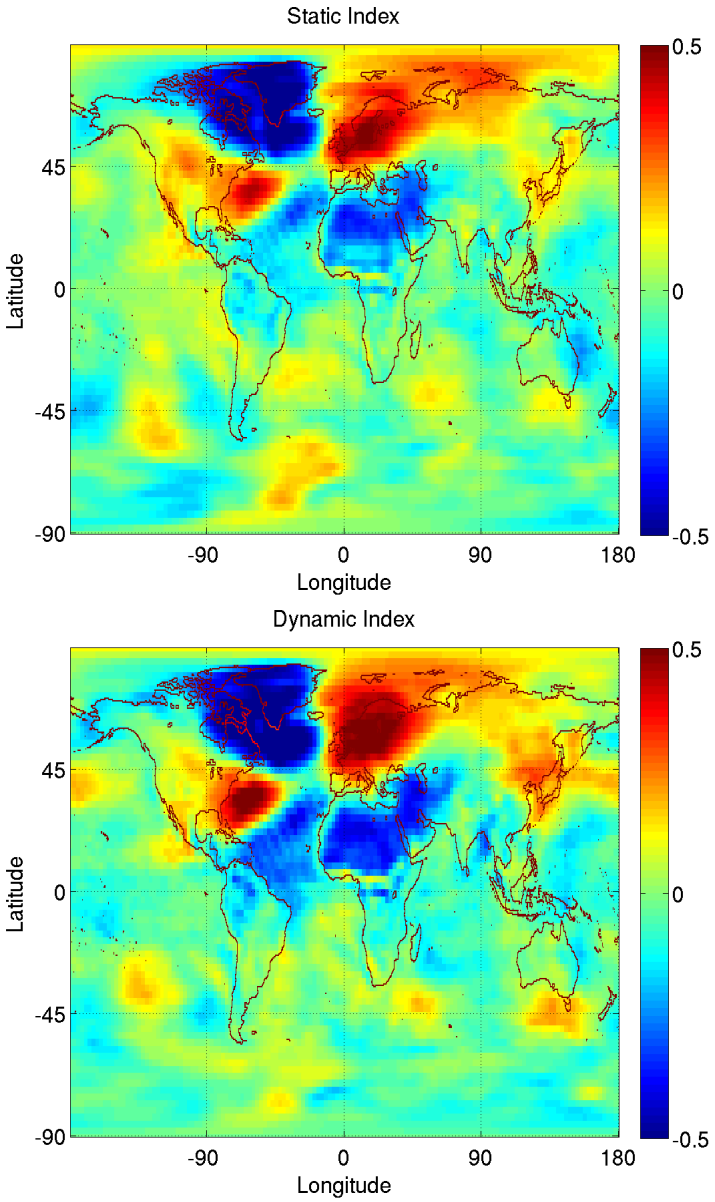


Figure 3.13: Impact on temperature for NAO using static and dynamic index for the time period 1988-2008. The figure shows that both the static and the dynamic index generate a similar impact on temperature, however the dynamic index shows higher correlation than the static.

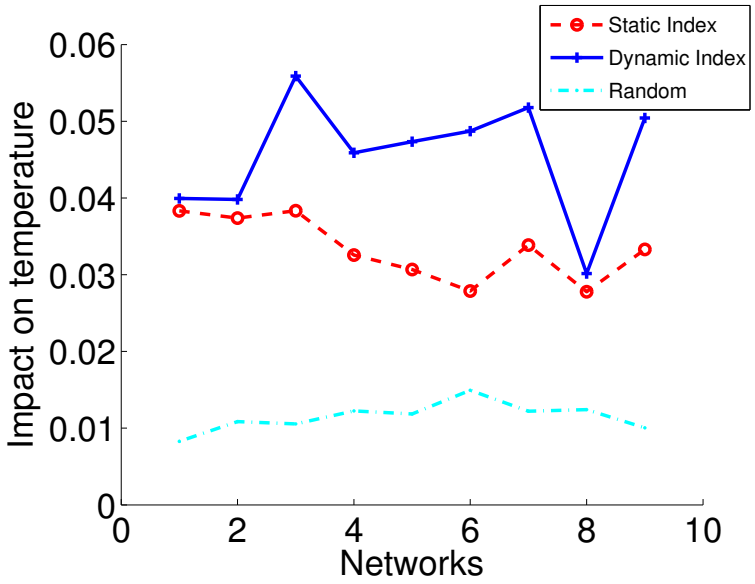


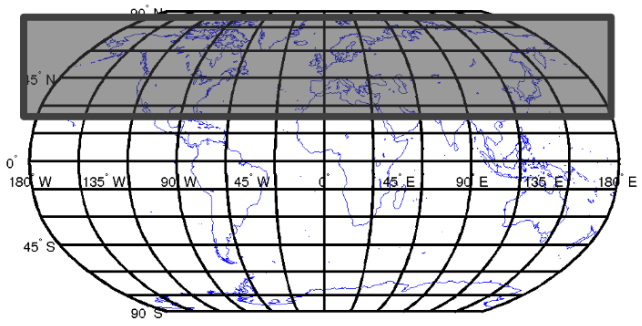
Figure 3.14: Aggregate area weighted impact on global temperature using static and dynamic NAO. The figure shows that the dynamic index performs better than the static for all the 9 network periods.

We also examined the impact on temperature anomalies using the static and dynamic NAO. Figure 3.13 shows the impact of both static and dynamic NAO for the time period 1988-2008. From the figure, we see that both static and dynamic NAO have a similar pattern but the dynamic index shows a stronger correlation with temperature anomalies. Fig 3.14 shows the aggregate correlation of global temperature anomalies using the static and dynamic NAO index for all the 9 time periods. From the figure, we see that the dynamic index shows significantly better aggregate area weighted correlation for most of the time periods.

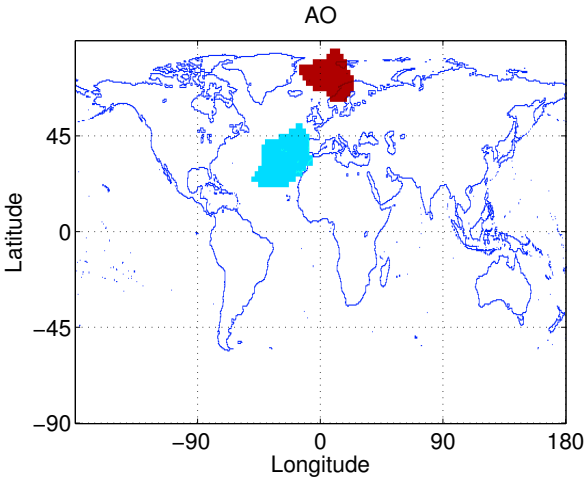
To validate that the land impact generated by our identified dipoles is not spurious, we perform a small randomization test. We randomly select 100 positively correlated time series from locations on Earth that are most likely not a part of any dipole. We compute their impact on global temperature anomalies. The blue line in Figure 3.12 and 3.14 shows the mean impact using the these 100 random locations. Note that static

and dynamic indices have a much stronger impact as compared to the random baseline. The dynamic index always generates a stronger impact than the static one. We are also able to show a better impact on precipitation anomalies using CRU observational data [2] but do not report the numbers to conserve space.

3.4.3 Region Based Definition of Dipoles



(a) $20 - 70°$ N region used to define AO



(b) Region based dipole for AO for the time period 1988-2008.

Figure 3.15: Region based definitions for the AO dipoles defined using EOF analysis.

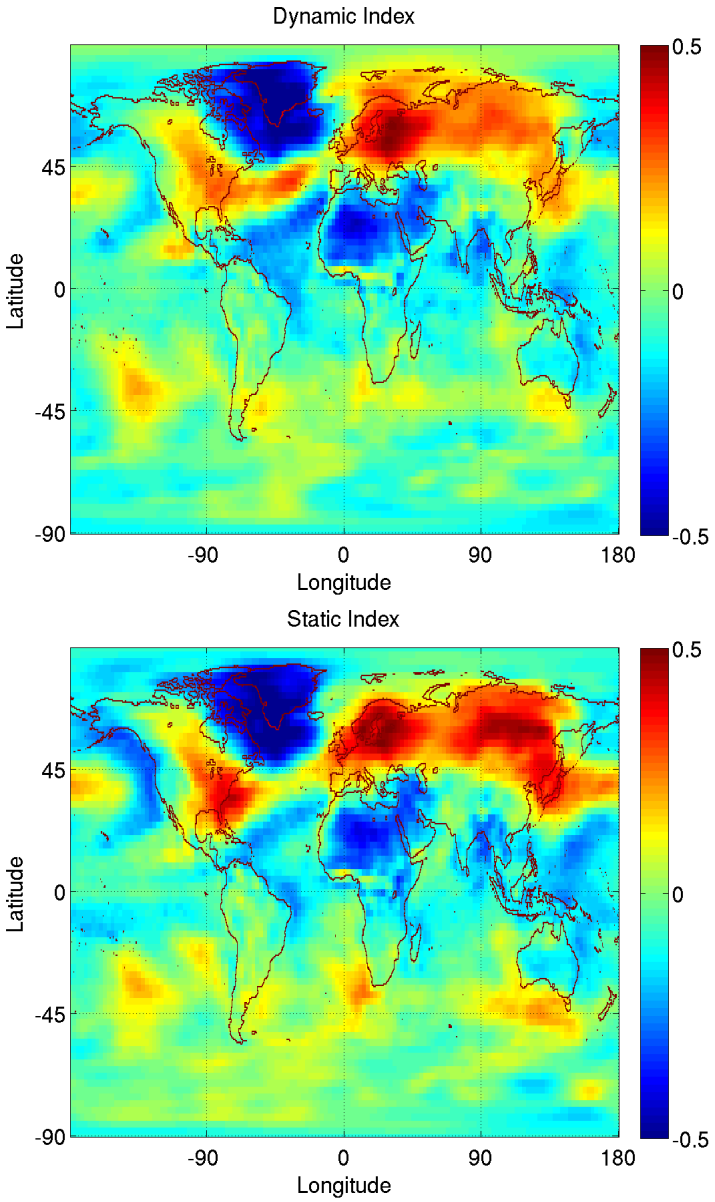
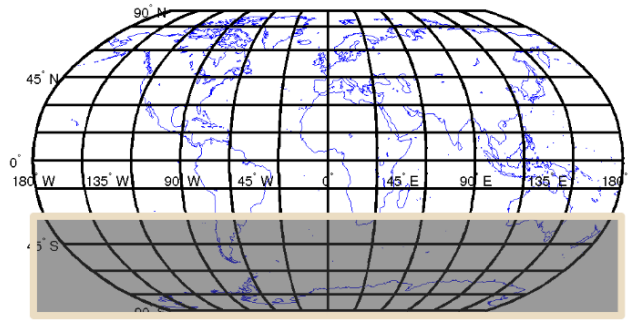
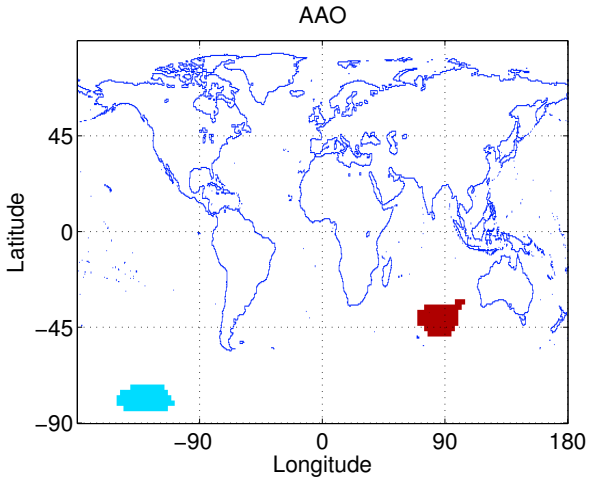


Figure 3.16: Impact on temperature for AO using static and dynamic index for the time period 1988-2008. The figure shows that both the static and the dynamic index generate a similar impact on temperature.



(a) $20 - 70^{\circ}$ N region used to define AAO



(b) Region based dipole for AAO for the time period 1988-2008.

Figure 3.17: Region based definitions for dipole AAO defined using EOF analysis.

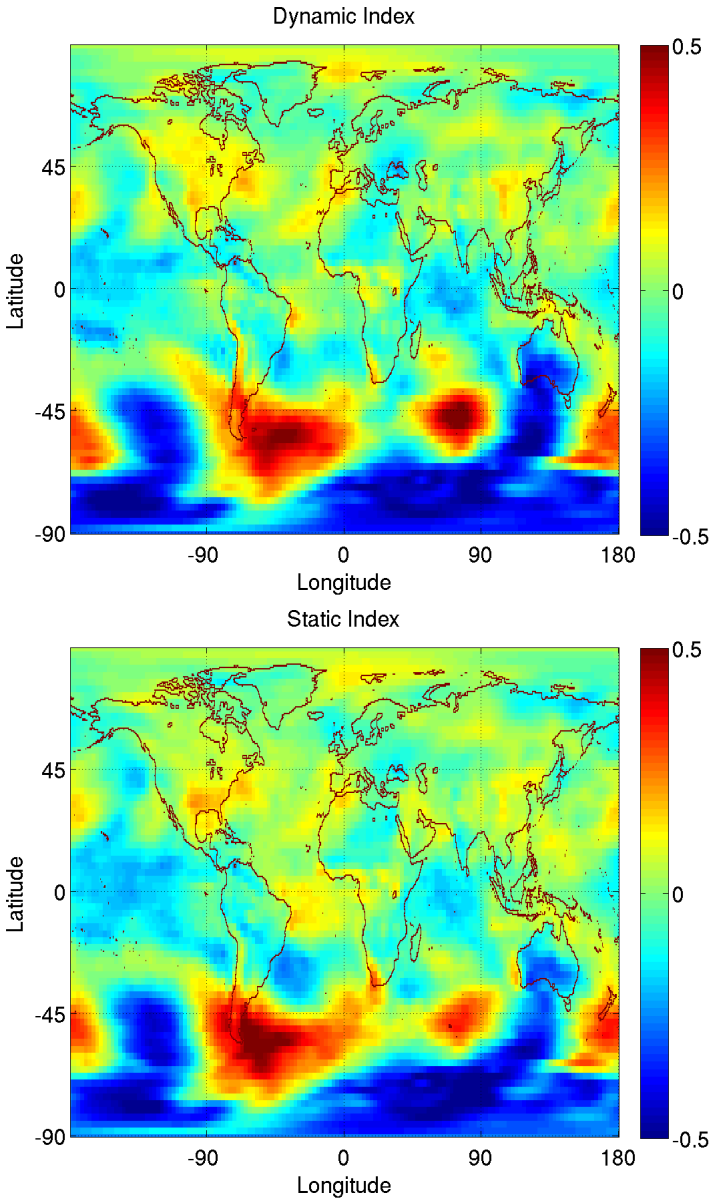


Figure 3.18: Impact on temperature for AAO using static and dynamic index for the time period 1988-2008. The figure shows that both the static and the dynamic index generate a similar impact on temperature.

Recall that dipoles like the Arctic Oscillation (AO), Antarctic Oscillation (AAO), etc are defined by climate scientists using EOF analysis [1]. In order to compute the AO, the EOF analysis takes an entire region from 20°N-90°N and finds the leading eigenvector which corresponds to the AO. Similarly, for AAO, the EOF analysis takes an entire region from 20°S-90°S and finds the leading eigenvector corresponding to the AAO. EOF analysis involves decomposition of the given data into orthogonal basis functions that characterize the covariability of the time series. EOF analysis is similar to Principal Component Analysis (PCA) and the basis functions are typically found by computing the eigenvectors of the covariance matrix of the given data. The *ith* basis function is orthogonal to all the basis functions from *1 to i-1* in order to minimize the residual variance.

The biggest advantage of using EOF analysis is its simplicity and convenience for characterizing dominant spatial pattern of variability. However, there are several disadvantages of EOF analysis [18] namely -

1. **Orthogonality constraint:** The major constraint in EOF analysis is that the principal components need to be orthogonal to each other. An important disadvantage of this constraint is that the teleconnections found using it are not necessarily associated with an underlying physical process. [18] show an artificial example in which the EOF analysis shows the presence of a dipole pattern that does not exist in the original data and is created as an artifact of the orthogonality constraint. Thus, EOF analysis can create patterns from "noise". Further, due to the orthogonality constraint on the PCs, it is uncertain if the PCs of 3rd order or more have any physical significance attached to them.

2. **Discovered dipoles not a good representation of reality:** The centers of action using EOF analysis do not need to correspond to the centers of actions of an actual physical process. Further, due to the way the PCs are constructed, the PCs of the dominant pattern are formed by a superposition of several uncorrelated

patterns often from remote regions with no influence on the center of action.

3. **Sensitive to input data:** The method also fails when there are prominent oscillations that are not spatially orthogonal to each other but are nearly statistically independent in the time domain. The EOF analysis can be sensitive to the choice of spatial and time domain. This is a critical point as often, a small region and a subset of time is used to find local teleconnection patterns.

Using our approach for dipole detection, we are able to find two regions based corresponding to the two dipoles AO and AAO and having a correlation of > 0.8 and > 0.9 respectively. Figure 3.15(b) shows the region based dipole corresponding to the AO and figure 3.17(b) shows the region based dipole corresponding to the AAO for the time period 1988-2008.

Further, if we compare the impact of global temperature using both the region based dipole using our algorithm and the EOF based dipole, we see that they both look very similar. Figure 3.16 shows the impact on temperature using the AO index from the Climate Prediction Center (CPC) which is defined by using EOF analysis and the one using our algorithm. From the figure, we see that the two indices show a very similar impact. Thus our region based index is a good surrogate for the index defined using EOF analysis. Further, figure 3.18 shows the impact on global temperature using the AAO index from the Climate Prediction Center (CPC) which is defined by using EOF analysis and the one using our algorithm. From the figure, we see that the impact on temperature using both the indices appear similar and our index is very good representative of the CPC index.

3.4.4 Dipoles in Other Reanalysis Datasets

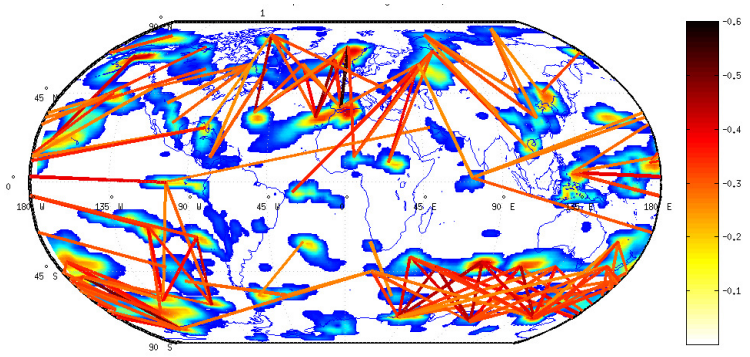
As mentioned in Section 2.1, there are many reanalysis products that serve as a proxy for global observation. To assess the stability of our algorithm for different reanalysis

datasets, we apply it to three reanalysis datasets- ERA [72], JRA [57] and NCEP [43] for the 21 yr time period 1979-2000 that is common to all the three datasets[1] . The ERA and JRA datasets have different grid resolutions but are interpolated to the NCEP grid size for a fair comparison. These reanalysis datasets are similar in nature but vary in some important aspects, such as the temporal extent and the details of the numerical model used for assimilation. For instance, the NCEP Reanalysis data is available from 1948 until the present and the European 40-Year Reanalysis (ERA-40) from 1957 to 2002. As a result, some artifacts may be introduced and examples have been reported in the literature, e.g., Hines et al. [38] noted artificial trends in the NCEP Reanalysis pressure fields. In contrast, the Japanese 25-Year Reanalysis (JRA-25) starts at 1979 and falls entirely in the satellite era, meaning that satellite-based measurements are available for the full duration of the project; it is considered to be one of the most reliable reanalysis datasets presently available.

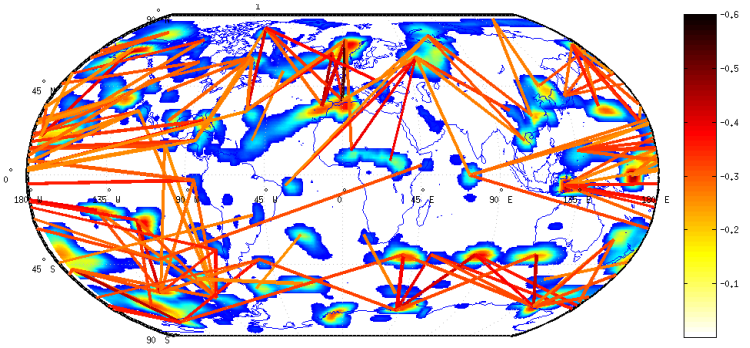
Figure 3.19 shows the dipoles detected in the reanalysis datasets - ERA, JRA and NCEP. From a visual inspection of the figure, we see in all the three reanalysis datasets our algorithm is able to capture all major dipoles (SO, NAO/AO, AAO, WP, PNA). As we noted earlier, most of the edges in the Southern hemisphere correspond to the AAO. The AAO appears to be well represented in all the three datasets and is more densely connected in the ERA and the NCEP dataset as compared to the JRA dataset. As mentioned earlier, the SO dipole is centered around Darwin in the north Australia and Tahiti. It is well represented in all the three datasets and more strongly represented in the NCEP and the JRA dataset as compared to the ERA dataset. The NAO/AO dipoles in the north Atlantic also appear to be well represented in the three datasets. The similarities in the results from the three reanalysis datasets gives us confidence that our approach is able to find dipoles in a consistent manner across the three datasets. We also see some unknown connections starting near India to South America particularly

[1] This period is common to all three datasets and also matches with thee 20th century simulations of the GCMs used by the IPCC

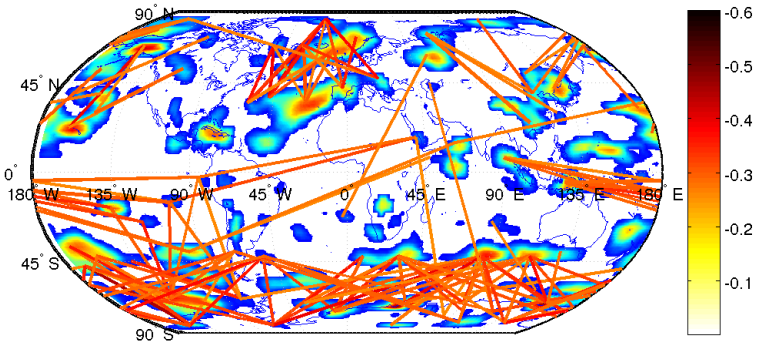
in the JRA and the NCEP dataset. These need to be investigated further by climate scientists.



(a) Dipoles in the ERA-40 dataset



(b) Dipoles in the JRA-25 dataset



(c) Dipoles in the NCEP dataset

Figure 3.19: Dipoles in the three reanalysis datasets for the time period 1979-2000 thresholded at -0.25.

3.5 Applications

3.5.1 Understanding dipole changes over time

Our approach's ability to detect and visualize all the dipoles on the globe as in Figure 3.10 empowers our understanding of climate data in many ways. For example, it allows us to study the changes in the dipole behavior. Understanding the dipole movements throughout the different time periods enables us to have a better prediction. For example, figure 3.20 shows the movement of the local attractor of the NAO dipole over the 9 network periods. The figure shows that the center of activity of the NAO is not fixed but moves throughout the different time periods.

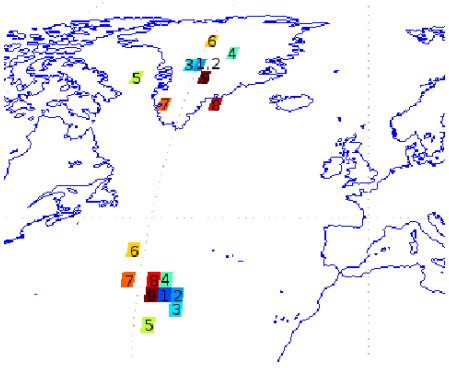


Figure 3.20: NAO.

3.5.2 Understanding Dipole Interactions

Understanding the interplay between the dipoles is crucial to understand the variability of the global climate system. Although the linkage of individual teleconnection patterns to weather changes in various parts of the globe is well known and documented, there are not many investigations examining the relationship between the various dipoles which is also critical as the synergy between different dipoles can have great influence on climate. E.g., while the cold winter over Europe in 2010 could be largely explained by

NAO and other local indices [16], the cold winter over North America at the same time is largely due to a combination of NAO and ENSO [31]. Thus knowledge of interactions between multiple dipoles can be useful and the potential societal payoff to mitigate and plan for the impacts of the interplay between various teleconnections can be significant. Figure 3.21 illustrates dynamic changes in the interactions of the NAO/AO dipoles in the NCEP data when compared for different time periods, 1948-1967, 1968-1987, and 1988-2007.

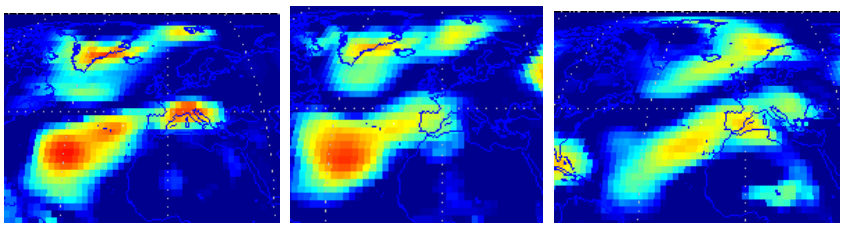


Figure 3.21: NAO/AO interactions in the three periods, 1948-1967, 1968-1987, and 1988-2007.

3.5.3 Understanding IPCC AR4 Models

Another important application of global dipole analysis is its use in the understanding of the skill of various General Circulation Models (GCMs) used for climate prediction. Various GCMs exhibit variability in their predictions of various climate variables, as they use different representations of physical interactions in the climate system. Hence they often diverge in their predictions and sometimes even offer contradicting projections of changes in various regions in response to different greenhouse gas emission scenarios. Our current approach provides a comprehensive view of the dipoles on Earth and, hence, a power to test various models in terms of their ability to capture the dipoles. Despite the prevalence and importance of teleconnections in climate science and climate related impacts, an adequate study quantifying the teleconnections in the climate models is still lacking. Similarities or differences in dynamic dipole structure can offer valuable insights to climate scientists on model performance, which further aids in assessing reliability of

climate prediction simulations.

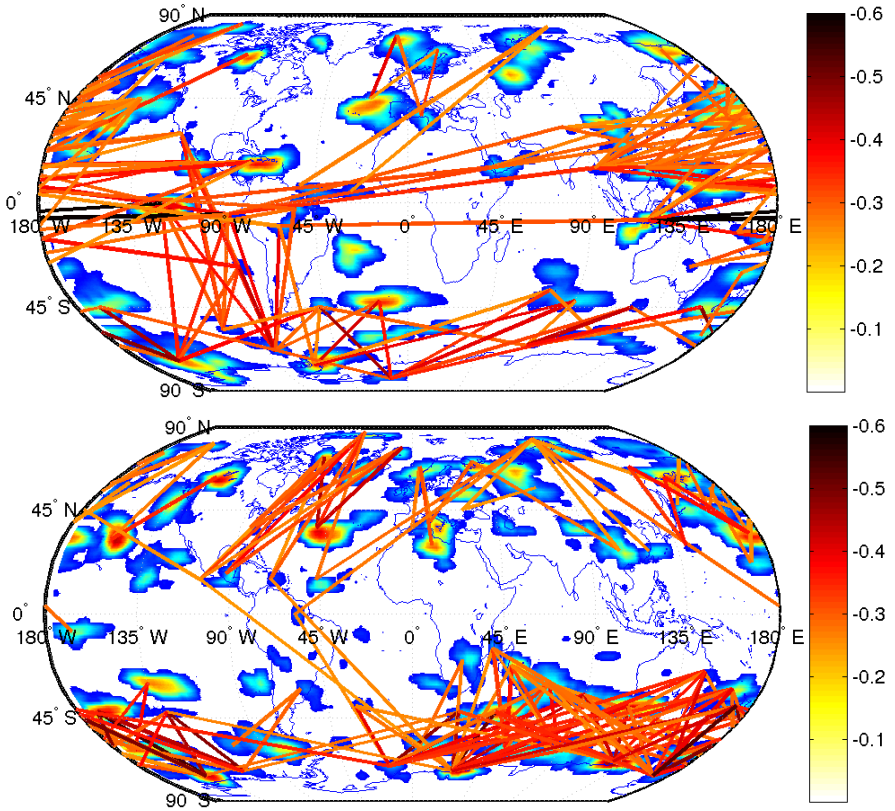


Figure 3.22: GFDL 2.1(above) and GISS E-R(bottom) hindcast data from 1975-2000. The figure shows that SO is very strongly present in GFDL but is missing in GISS.

As an illustrative example, we use simulation output data from two of the more than 20 general circulation models (GCMs) from the Fourth Assessment Report (AR4) of the IPCC [56], namely GFDL 2.1 and GISS E-R. Our SRNN based dipole detection algorithm allows us the ability to compare the performance of the different models by looking at their dipole networks. We detected dipoles in the data from the two IPCC climate models using backward (hindcast) data from the year 1975-2000. Figure 3.22 shows the dipole structure in the two models. From the figure, we see that using our approach the GISS model does not show SO but the GFDL model has a very strong representation of it which is much stronger than the same dipole in the reanalysis datasets (see Figure 3.19).

3.5.4 Discovery of New Teleconnections

Another important application of our approach is the ability to discover previously unknown teleconnections. Our approach allows us the unique ability to generate a single snapshot picture of all the dipole teleconnections on the Earth. This was not possible earlier. Thus it provides new avenues to discover some previously unknown relationships. As a part of our future work, we will work on examining individual edges that do not have a high correlation with any known phenomenon but still have a significant negative correlation and impacts regional temperature and precipitation.

3.6 Discussion

In this chapter, we presented a novel systematic shared nearest neighbor based approach to find dipoles in global climate data. We show the utility of our approach by effectively capturing the known dipole connections with a high correlation in the reanalysis datasets. Further, we show that we are able to get region based definitions for dipoles defined earlier using EOF analysis. The dynamic dipoles captured using our systematic

graph based approach have a better impact on temperature anomalies as compared to the static indices used by climate scientists. Furthermore, our approach generates a single snapshot picture of all the dipoles on the Earth and thus using it we can study the interconnections between dipoles and show possible interactions between atmospheric oscillations. Knowledge of these interactions is particularly important for predicting climate extreme events. For example, while the cold winter over Europe in 2010 could be largely explained by NAO and other local indices [16], the cold winter over North America at the same time is largely due to a combination of NAO and ENSO [31], thus knowledge of patterns that span multiple dipoles can be useful. Using this approach we can study the changes in their dynamics and structure in a much more systematic manner.

Moreover, our approach can be used as an alternative method to measure the climate model performance. Since the dipoles or teleconnections define the heartbeat of a climate system, we can measure how well the dipoles are represented in the different model simulations. From our preliminary investigation, we see that different models vary in their ability to capture dipoles. Indeed, some models seem to miss some dipoles completely. Climate predictions so far mostly rely on taking averages of the models and since dipoles are prevalent and important in climate data as they are known to be linked to climate variability across the globe, this result is very important in assessing the goodness of a climate model and the value of making regional predictions from the model. Further, this can provide insights into the creation of ensembles of the various models for further climate predictions.

Finally, we present a number of challenges and issues that form a part of the future work. The first important goal is to assess the statistical significance of the detected dipoles. There are many challenges in spatio-temporal context for randomization testing due to the presence of non i.i.d. data, seasonality and trends, etc. Another challenge is to address multivariate relationships in the data. So far, in this paper, we have mainly

looked at dipole representation in a single climate variable, i.e. sea level pressure. However, it would be very useful to examine dipole relationships in other variables like temperature, precipitation, etc. Further, the SRNN technique could be modified to handle a multivariate graph. The construction of a multivariate graph from the climate data itself would require several innovations. Also, in this paper, we have constructed networks using the linear correlation. However, most of the phenomenon in climate are characterized by a lagged correlation. Handling lagged correlation adds predictive power to the dipoles. However, there are no guidelines available on handling lagged correlations and assigning the direction of lags. Lagged dipole relationships are important and give predictive power to the dipoles and we discuss it further .

Chapter 4

Mining lagged relationships in spatio-temporal datasets [1]

4.1 Introduction

So far we have examined algorithms to capture instantaneous relationships among two locations. However, time series data in climate are often characterized by a delayed relationship between two variables, for example precipitation and temperature anomalies occurring at a place might also occur at another place after some time. Time-lagged relationships are generally discovered by examining the correlation between the time series of the two locations relatively shifted in time with respect to one another. In climate data, time-lagged relationships are crucial towards understanding the linkages and influence of the change in the climate at one region of the Earth on another region. These relationship are lagged in time because a climatic phenomenon that affects a specific location does not affect a different location at the same time but only at a

[1] **This chapter is based on the work [46] published in the proceedings of the IEEE Conference on Intelligent Data Understanding, CIDU 2012.**

later time. Lagged relationships can be a key indicator for weather prediction and understanding them can help assist prediction, monitoring and planning activities.

Identifying lagged relationships in climate data is challenging due to the various complex dependencies present in the data like spatial and temporal auto-correlation, seasonality, trends and long distance teleconnections. In this chapter, we present a general framework for finding all pairs of lagged positive and negative relations that can exist in a given spatio-temporal dataset. One such important time-lagged pattern in climate is the propagation of the Madden Julian Oscillation (MJO [84]). The MJO is one of the most dominant component of intra-seasonal variability in the tropics and has been established by climate scientists as a large scale coupled atmospheric pattern propagating eastwards around the tropics. Fig. 4.1 shows the spatial movement of the oscillation which also represents the movement of the convective centre and rainfall. Owing to its connection to the weather system across different parts of the globe, its understanding has the potential to aid forecasts of precipitation, hurricanes, tropical storms and weather variability across a number of regions on the globe [84].

Despite the prevalence and importance of time lagged relationships in climate data, there are not many approaches available to systematically model and extract all the lagged relations in a given spatio-temporal dataset. Earlier, these relationships in climate science were typically discovered by the targeted investigations of highly trained scientists of the phenomena of interest. Although, such manual approaches may produce quite noteworthy results (for example, in the discovery of most of the known teleconnections), they may also miss important relationships as we saw in the previous chapter. With the availability of large amount of satellite data, there are some data driven techniques like the Empirical Orthogonal Function (EOF) analysis [75] used to discover these relationships in climate science. However, a problem with the EOF analysis is that it imposes orthogonality constraint on the principal components and hence the physical interpretation of the signals is difficult [18].

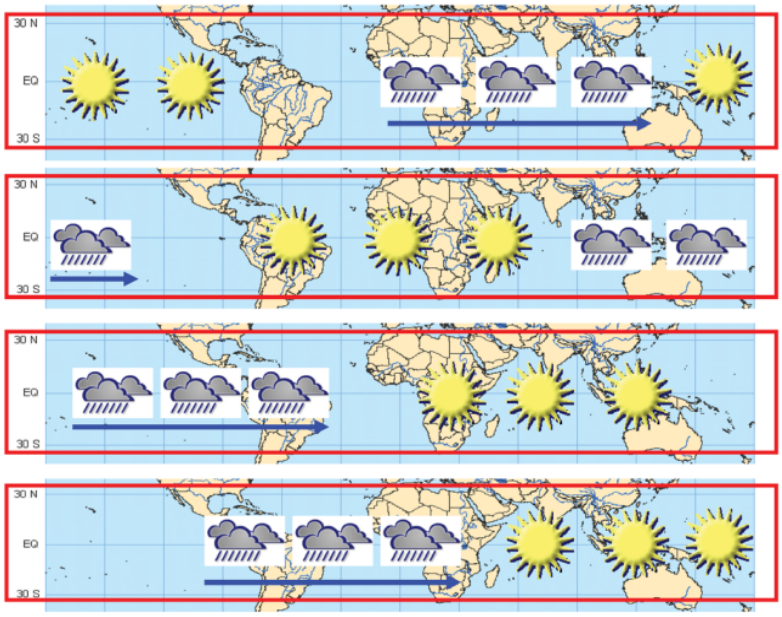


Figure 4.1: MJO spatial structure and evolution: Large scale pattern shifting eastwards over time. The cloud (Sun) icons represent the enhanced (suppressed) phase of the MJO respectively and the blue arrows indicate the eastward movement. Fig taken from [8]

We extend our previous graph based approach [47] based upon the concept of *shared reciprocal nearest neighbor* to generate cluster pairs of locations sharing similar or opposing behavior for every time lag. Our framework can be generalized to extract multivariate lagged relationships across different variables thus can be used to understand the lagged response of one variable on another. We show the utility of our approach by extracting some of the known delayed relationships like the Madden Julian Oscillation (MJO) and the Pacific North American (PNA) pattern at different lags using the sea level pressure dataset provided by the NCEP/NCAR. Our approach can be broadly applied to other problems in spatio-temporal domain to extract lagged relationships.

4.1.1 Challenges

Discovering lagged spatio-temporal relationships among climate variables involved is extremely challenging due to the nature and massive size of the data. We enlist some of the challenges encountered as follows:

Nature of spatio-temporal data

Mining lagged relationships in spatio-temporal data such as climate requires a deeper understanding of the complex dependence structures inherent in the data, for example spatial and temporal auto-correlation, seasonality, non-linear associations, trends or long-range spatial dependencies (teleconnections) [49, 47].

Difficulty in adapting current data mining algorithms

Traditional methods in data mining generally handle spatial and temporal aspects separately and hence are not suitable to capture the spatio-temporal lagged relationships in climate. For example, association analysis is a widely established approach and is conceptually well-suited to find relationships in large datasets [10, 11]. However, the use of association analysis for spatio-temporal data poses several challenges since it was originally developed for non-spatial data that consisted of objects described by categorical or binary variables and thus the notion of transactions is ill-defined in spatio-temporal datasets. Other data mining techniques like clustering, etc have also been widely studied for static datasets and non-lagged behavior [64].

Grouping Locations

Looking for lagged relationships among large numbers of locations poses many conceptual and computational problems. In particular, because of an explosion in space/time

complexity and/or generation of too many patterns many of which are redundant. Finding the right time lag for a phenomenon at two given places is challenging because the correlation may be visible at different time lags and impacting neighborhood regions with different intensity at different lags. For example, two places may appear to be correlated to a third location at different time lags.

4.2 Our Approach

We propose a graph based approach to find the time-lagged relationships in the climate dataset. Graph based approaches to analyze climate data have been used extensively by prior research work (such as [69, 61, 62, 49, 47]), however the aspect of *time lags* has not been explored previously. Time based lags leads to a directed graph which earlier approaches cannot handle, hence discovering lagged relationships is a *non-trivial extension* of our prior work [47].

The main idea of our approach is to cluster locations such that two locations within the same cluster have similar lagged relationship with other clusters. Our goal is to find such cluster pairs at different lags and we define the lagged relationship based on both positive (and negative) correlations. In order to find the clusters, we first construct the complete directed lag graph based on the lagged correlations, and then perform a sequence of graph operations, K-nearest neighbor and reciprocity filtering followed by shared reciprocal nearest neighbor algorithm to compute the clusters. We then find cluster pairs that share similar or opposing behavior at a particular lag. These cluster pairs are primarily regions on the Earth that influence one another over the specified lag. The following subsections provide step by step details of our algorithm.

4.2.1 Complete Directed Lag Graph

Let $G^l = (V, E^l)$ represent the directed weighted graph with lag l, where $V = \{v_1, \ldots, v_n\}$ represents the n graph vertices and in our case correspond to the locations on the Earth and E^l is the set of directed edges in the graph where each entry E^l_{ij}, $1 \leq i, j \leq n$, indicate the edge weight of an edge starting at vertex v_i and ending at v_j taken at lag l. Note that the edge E^l_{ij} is directed, such that, $E^l_{ij} \neq E^l_{ji}$. The edge weight E^l_{ij} is computed by using the cross-correlation function which involves shifting forward the time series (say x_j) at v_j by l, as follows:

$$E^l_{ij} = \frac{\sum_{k=1}^{m-l}(x_i^k - \bar{x}_i) \cdot (x_j^{k+l} - \bar{x}_j)}{[(x_i - \bar{x}_i)^T(x_i - \bar{x}_i) \cdot (x_j - \bar{x}_j)^T(x_j - \bar{x}_j)]^{\frac{1}{2}}} \tag{4.1}$$

where x_i is the time series at location i, $m = |x_i| = |x_j|$ is the length of the time series, l is the lag and x_i^k represent the k^{th} component of the time series. Note that x_j^{k+l} represents the time series at location j shifted by time lag l. Thus E^l_{ij} captures the traveling of the phenomenon from location i to j. We use eq. 4.1 to instantiate the complete directed graph at a specified lag between all the graph nodes.

4.2.2 Positive and Negative Lag Graphs

The next step is to consider separately the positive and the negative edge weights E^l_{ij}. This is because the polarity of the lag correlation conveys a different meaning: positive lag correlation indicates the traveling of a phenomenon from one location to another whereas negative lag correlation indicates opposite influence of a phenomenon and specifies a lagged dipole. Additionally, considering them separately aids in the clustering algorithm. We construct two graphs $G^{l+} = (V, E^{l+})$ and $G^{l-} = (V, E^{l-})$ from G^l as follows:

$$E_{ij}^{l+} = \begin{cases} E_{ij}^{l} & \text{if } E_{ij}^{l} > 0 \text{ and } E_{ij}^{l} > E_{ji}^{l} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

$$E_{ij}^{l-} = \begin{cases} E_{ij}^{l} & \text{if } E_{ij}^{l} < 0 \text{ and } E_{ij}^{l} < E_{ji}^{l} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

where, setting edge weight to 0 means that the edge is of no consequence and it can be deleted. So for every pair of node v_i, v_j, we keep either edge $v_i \to v_j$ or $v_j \to v_i$ based on whether the edge maximizes positive correlation or negative correlation. Note that if E_{ij}^{l} and E_{ji}^{l} have the same sign then only one of the edge $v_i \to v_j$ or $v_j \to v_i$ appears in either of G^{l+} or G^{l-}; otherwise both edges appear in one of the positive or negative lag graphs in a mutually exclusive fashion.

4.2.3 K-Nearest Neighbor Lag Lists

Next, we construct the K-nearest neighbor (KNN) lists based on G^{l+} and G^{l-}. KNN list for a node A is constructed by sorting all edges of A based on edge weights and then picking the top K nodes with the edge weights from A. Now since in our case the graphs are directed, we can construct these lists based on out edges as well as in edges. This leads to four such lists S^{lo+}, S^{li+}, S^{lo-}, and S^{li-}. For e.g., S_i^{lo+} is constructed by picking all outward edges from v_i and then sorting the edges from highest to lowest based on the edge weights in E^{l+} and then picking top K edges. Similarly, we can construct the other three lists.

4.2.4 Reciprocal Lag Graph

Once we get the outgoing and incoming KNN lists for both positive and negative correlations, we generate the reciprocal graphs. The concept of reciprocity as shown by

Kawale et al. [47] helps in removing noisy correlations. In case of non-lagged edges, two graph nodes a and b are said to be connected by a reciprocal edge $a \leftrightarrow b$ if a lies in b's KNN graph and vice-versa. However, in the case of lagged correlations the edges are directed. As a result, we redefine the concept of reciprocity for directed edges. We define an edge $a \rightarrow b$ to be directed reciprocal iff node a appears in $in-$KNN list of b and b appears in $out-$KNN list of a. Using this definition of reciprocity, we compute the reciprocal graph $G^r = (V, E^r)$ as follows:

$$E^r_{ij} = \begin{cases} 1 & \text{if } v_i \in S^{li+}_j \wedge v_j \in S^{lo+}_i \\ -1 & \text{if } v_i \in S^{li-}_j \wedge v_j \in S^{lo-}_i \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

Note that only one of the three conditions can be satisfied in the eq. 4.4. It is not possible for two nodes to have reciprocity in both negative and positive lists.

4.2.5 Shared Reciprocal Nearest Neighbors

After constructing the reciprocal graph, our next step involves clustering of the graph nodes. We perform this clustering based on the shared reciprocal nearest neighbor (SRNN) idea presented in [47]. Shared nearest neighbor (SNN) approach [24, 61] to find clusters in the climate data has been shown to be useful in finding clusters with arbitrary shapes and density. SNN graph is constructed by computing the edge weights between two nodes based on how many edges they have in common. SRNN uses the concept of SNN to find clusters in the reciprocal graph. Since the reciprocal graph is directed, SRNN algorithm cannot be directly applied. Instead, we construct 4 SRNN graphs G^{si+}, G^{so+}, G^{si-}, and G^{so-}. The graph $G^{so+} = (V, E^{so+})$ is constructed by considering the positive (with edge weight 1) out going edges in E^r, thus we get

$$E^{so+}_{ij} = \left|\left\{v_k : \forall k, E^r_{ik} = 1 \wedge E^r_{jk} = 1\right\}\right| \tag{4.5}$$

Similarly, we can compute $G^{si-} = (V, E^{si-})$ as follows:

$$E^{si-}_{ij} = \left|\left\{v_k : \forall k, E^r_{ki} = -1 \wedge E^r_{kj} = -1\right\}\right| \tag{4.6}$$

We can also use the rank of the neighbors in K-Nearest Neighbor lag lists to compute the edge weight in the SRNN graphs as also mentioned in [47].

Algorithm 2 Finding Clusters from Lagged Densities.

Let, $DS_{i,j}$ be geographical distance between locations i and j.
Let, $CORR^0_{i,j}$ be anomaly correlation between locations i and j at lag 0.
Let, $D^{so}_i = \sum_{j=1}^{N} E^{so}_{i,j},\ \forall i \in \{1, 2, ..., N\}$ (outgoing location density).
Let, $D^{si}_i = \sum_{j=1}^{N} E^{so}_{i,j},\ \forall i \in \{1, 2, ..., N\}$ (incoming location density).
for $i \in V$ **do**
  {Initialize local attractor with neighbor of maximum degree. Initialize it to self if this node has the maximum degree in the neighborhood.}
  LA^{so}_i, LA^{si}_i = FindMaxNeighbor(i)
end for
Let, $A^{so}, A^{si} = \{1, 2, ..., N\}$ (local attractor set - initially set to all locations on Earth).
repeat
  {Keep moving in the direction of the real local attractor for both incoming and outgoing densities respectively.}
  for $i \in A$ **do**
    $j = \arg\min_k(DS_{i,k} : k \in A \wedge k \neq i)$
    if $DS_{i,j} <$ Distance-Thresh AND $CORR_{i,j} >$ Correlation-Thresh **then**
      if $D_i \geq D_j$ **then**
        $A = A \setminus j$ {Eliminate j from attractor set as i is the attractor of j}
        $LA_z = i, \forall z \in \{1, 2, ..., N\} \wedge LA_z = j$
      else
        $A = A \setminus i$ {Eliminate i from attractor set as j is the attractor of i}
        $LA_z = j, \forall z \in \{1, 2, ..., N\} \wedge LA_z = i$
      end if
    end if
  end for
until convergence
Final Step: Link cluster pairs from D^{so}_i and D^{si}_i such that they have reciprocal edges across them.

4.2.6 Finding relationship clusters from SRNN graph

The final step of the algorithm involves extracting the relationship cluster from the four SRNN graphs. In the previous chapter 3 we proposed a local attractor based algorithm

(inspired from Denclue [39]) for undirected SRNN graph of negative correlations. For this work, we modify the local attractor algorithm to handle the directed SRNN graph with negative as well as positive correlations. Since our goal is to find cluster pairs with lagged relationships, we need to find cluster pairs A and B such that the locations in A have a lot of outgoing edges to locations in B in the graph G^{so} and locations in B should have a lot of incoming edges from A in the graph G^{si}.

We do this by first constructing the edge density for the graphs G^{so} and G^{si} for the positive and negative correlations each. For example, the edge density for the positive correlations is computed by taking the sum of the edge weights at each node in the graph as follows.

$$\forall_{i \in V} \; D_i^{so+} = \forall_{j \in V} \; \sum_{j=1}^{N} E_{ij}^{so+} \tag{4.7}$$

$$\forall_{i \in V} \; D_i^{si+} = \forall_{j \in V} \; \sum_{j=1}^{N} E_{ij}^{si+} \tag{4.8}$$

Once we compute the two densities for the two graphs, we find clusters in the densities using modified local attractor based algorithm mentioned in Algorithm 3. The main idea behind the algorithm is to first find the clusters in each of the two graphs G^{so} and G^{si} using their densities and then find the links across the clusters.

4.3 Results

We use the daily resolution data from the NCEP/NCAR Reanalysis project and construct the anomalies by aggregating the daily resolution data into pentads as mentioned in Chapter 2. We use the variable sea level pressure for most of our analysis. For the multivariate results, we examine the relationship between outgoing longwave radiation (OLR) and sea level pressure.

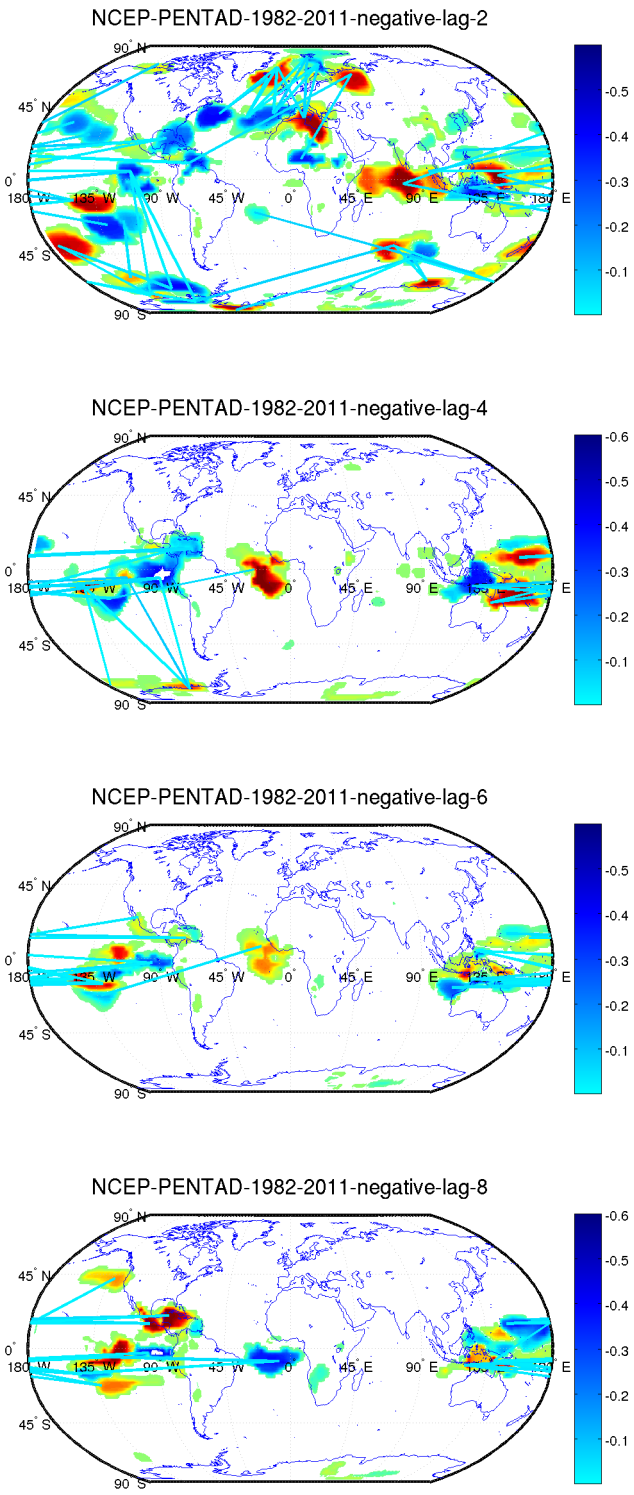


Figure 4.2: Lagged dipoles at different lags - 2, 4 and 8

The goal of our experimental evaluation is to qualitatively evaluate the patterns generated using our approach and see if they match some of the known patterns from climate science. We ran our algorithm using the sea level pressure dataset to generate the candidate cluster relations. We use lag values starting from 1 and go until 8 as lag values beyond that are unlikely from the climate perspective. We generated lagged cluster pairs for both positive and negative correlations. We examine the lagged cluster pairs generated at different lag values to verify if they revealed known patterns from climate science. In order to distinguish the regions which have predominantly outgoing edges from the ones which had predominantly incoming edges in the figures, we followed this simple scheme: Our approach generated the Shared Reciprocal Nearest Neighbor densities for the two graphs G^{so} and G^{si}. The density values for a node represents the number of shared outgoing or incoming connections respectively. We take a subtraction of the two densities and use red color on the map to represent regions which have more outgoing edges as compared to the incoming ones and the blue color to show regions which have more incoming edges as compared to the outgoing ones. We have used this coloring scheme throughout to show results in this chapter.

4.3.1 Negatively Correlated Clusters

Our first experiment involves evaluating the negatively correlated cluster pairs. The negatively correlated long distance connections in sea level pressure anomalies are known as *dipoles* [49, 47] and are important in understanding the variability of the climate. With our methodology, we are able to extract these negative relationships at different lags and not just at lag 0 which represents stand-alone oscillation patterns as discovered by the earlier data guided approaches [49, 47]. Lagged dipoles are important as they signify moving oscillation patterns. For example, the MJO is an intra seasonal fluctuation or atmospheric wave occurring in the tropics and is crucial for understanding intra-seasonal variability in the tropics.

Fig. 4.2 shows the negatively correlated teleconnections at lags $2, 4$ and 8 pentads respectively. We see that at lag 2 (which corresponds to 10 days) the negative teleconnections are not quite different from the ones seen at lag 0 (refer [47]). However, as we increase the length of the lags, the phenomena at the tropics remains pronounced and is less visible beyond it, which also conforms with knowledge from climate science that there is more predictability around the tropics. Further, the tropical teleconnections seen between lag 2 and lag 8 show the characteristics of the MJO [37]. Fig. 4.2 shows the movement of the dipole activity around the tropics from lag 2 to 8. In particular, the figure shows how the red region around west coast of Africa which was predominantly an outgoing region (red) at lag 4 turns into a predominantly incoming region (blue) at lag 8 hinting at the movement of the dipole phenomenon. However, we also need to look into other kinds of analysis like spectral analysis, etc., to confirm the MJO activity.

Apart from the MJO, which is the dominant pattern around the tropics at lags ≥ 2 and lags ≤ 8, there are other lagged dipole connections shown by our results which are known to climate scientists. One such connection as seen in the fig. 4.2 at lag 4 is connecting the region in Antarctica to the ENSO region. This connection was recently studied [60] as a slow linkage between AAO and the tropical El Niño. Using our method, we are able to find this connection at lag durations 2 to 5 in the sea level pressure dataset.

4.3.2 Positively Correlated Clusters

Apart from the dipoles which are formed by negative correlation between the anomalies at two locations, our algorithm can be used to find positively correlated cluster relationship pairs. Fig. 4.3 shows the positively correlated clusters at lags 1, 2 and 4 respectively. Due to spatial auto-correlation in climate data, for positive correlations we expect nearby locations to be highly correlated with each other and hence to be connected. At lag 1, as expected we see a lot of connections to nearby locations. As with the negative connections, the positive ones at lag vales ≥ 2 are also predominantly present

around the tropics. However, unlike the negative connections, the positive cluster connections mainly connect nearby locations. The positive cluster connections specify the movement of the same climate phenomenon unlike the lagged dipoles which represent the movement of the two ends of the oscillating pattern.

The MJO pattern which is seen as a moving dipole around the tropics in the negative correlations is also visible in the positive correlations. Fig. 4.3 shows the lagged cluster relationships at lag 4. We see that in the positive correlations, the clusters link nearby locations and reflect the movement of the oscillation. Apart from the MJO, there are other known connections also seen in the positive correlations. One such connection starts in the Atlantic near the coast of Brazil to Africa. This connection has been studied in climate science [58] to understand the impact of SST in Atlantic to the precipitation in Africa. This connection is important as it can help understand the predicability of precipitation in Africa. We also see this connection prominently when we only look at the data from Oct-May season and the connection appears more pronounced.

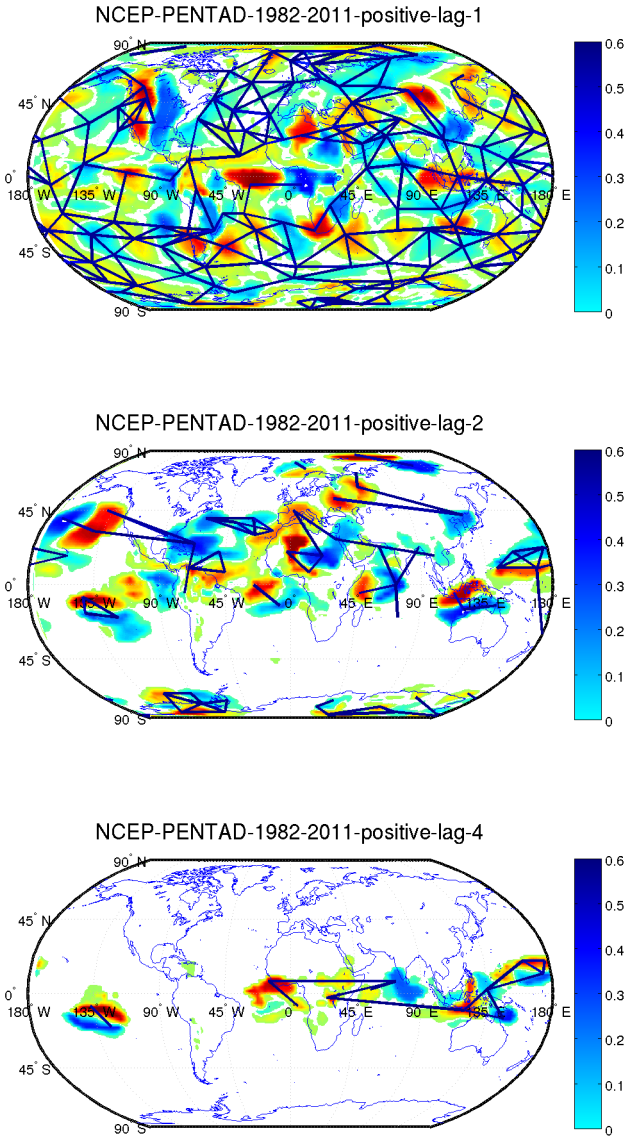


Figure 4.3: Positively correlated connections at different lag 1, 2 and 4.

Another established connection known to climate scientists is the Pacific North American (PNA) pattern. The PNA pattern is an important teleconnection pattern around North America has been linked with temperature and precipitation anomalies in the US. The PNA pattern is mainly represented by three poles namely - the vicinity of Hawaii and over the inter-mountain region of North America, south of the Aleutian Islands and over the southeastern United States. Data guided approaches to find dipoles only capture two of the three poles on the PNA and hence do not match the existing index with a very high correlation [47]. Using lagged connections, we see that at lag 2, we are able to find a connection linking the west coast of the US to the east coast and it corresponds to the PNA pattern.

4.3.3 Multivariate Clusters

So far, we saw that using our methodology we are able to extract both positively and negatively correlated cluster pairs at different lags. Our framework can also be used to study relationships across multiple variables. Using this framework, we can study the lagged impact on a variable using another variable as the leading variable. This feature is useful to examine and understand the relationships across different variables in climate data. As an example, we study the relationship across outgoing longwave radiation (OLR) and sea level pressure (SLP) dataset. This relationship is important to study for understanding the genesis of the MJO. Fig. 4.4 shows the plot showing the negatively correlated clusters of OLR with SLP at a lag of 4 pentads with OLR as the leading variable and SLP as the lagging. We see the dipole activity around the tropics is very prominent as also seen earlier.

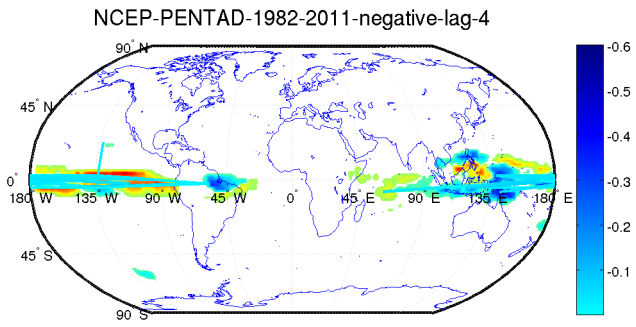


Figure 4.4: Relationship across OLR and SLP can be examined with the framework.

4.4 Discussion

In this chapter, we present a novel graph based approach to find lagged relationships in spatio-temporal climate dataset. Our approach is built upon the concept of reciprocity and shared reciprocal nearest neighbors to find lagged positive and negatively correlated cluster pairs. We demonstrate the usefulness of our method by extracting some of the known climate patterns using the sea level pressure dataset provided by NCEP/NCAR. Using our algorithm, we are able to identify some known connections such as the Madden Julian Oscillation and the Pacific North American pattern. Our approach can help to discover and examine all the lagged relationships in a given dataset. Further, our framework can be used to study the relationships across different variables and understand the inter-relationships. Our future work also involves examining the patterns extracted using our framework and assessing their impact on other variables for prediction, e.g. precipitation and temperature. The future work also involves examining some of the lesser known patterns to assess their significance both from the domain and statistical standpoint.

Chapter 5

Testing the significance of spatio-temporal teleconnection patterns[1]

5.1 Introduction

As we saw earlier, pressure dipoles are important long distance climate phenomena (*teleconnection*) characterized by anomalies (where anomalies are computed from raw data by subtracting the long term monthly means and are widely used in climate studies to take care of the seasonality in the data) of opposite polarity appearing at two different locations at the same time. These dipoles are important for understanding and explaining the variability in climate in many regions of the world, for e.g., the El Niño climate phenomenon is known to be responsible for precipitation and temperature anomalies over large parts of the world. Historically, these dipoles have been discovered by direct

[1] **This chapter is based on the work [45] published in the proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012**

observation of some climate phenomenon on land and have been defined using single point locations [78]. Later on, pattern analysis techniques such as the EOF [75] have been used to identify individual dipoles over a limited region, such as Arctic Oscillation (AO). However, there are several limitations associated with EOF and other types of eigenvector analysis; namely, it only finds a few of the strongest signals and the physical interpretation of such signals can be difficult due to the orthogonality of EOFs, whereas signals in climate are not necessarily orthogonal to each other. Systematic approaches for dipole discovery have been proposed in [47, 50, 61]. Kawale et al. [47, 50] present a graph based approach to find dipoles in the climate data and are able to match the existing dipole indices used by climate scientists with a very high precision and are able to provide region based definitions for dipoles defined earlier using EOF analysis. An important utility of the dynamic dipoles defined using this approach is that they are able to capture the dynamics of the climate phenomenon unlike the existing approaches that are based on pre-specified regions. Hence these dynamic dipoles tend to capture greater amount of climate variability at the global level [47, 50]. Further, they have been shown to be important in understanding the structure of the various General Circulation Models (GCMs) which are used to understand global climate change [47]. It is imperative to have a significance testing to rule out spuriously connected regions, correlated by random chance. This can help in discovering a new dipole phenomenon, previously not known to climate scientists. Given the importance of the teleconnections in influencing extreme weather events like tropical cyclones, droughts, hurricanes, etc., a previously unknown connection provides a critical missing link to the climate scientists.

Systematic approaches for dipole discovery generate a large number of *candidate* dipoles, i.e. two regions that are connected by negative correlation in their anomalies, that might possibly represent a physical phenomenon. Fig. 5.1 shows the dipoles generated by the algorithm given in [47]. The edges represent a connection between the two opposing ends of the dipoles. The figure captures most of the dipoles known to climate

scientists, however, it also shows a large number of edges that do not correspond to any known dipole phenomenon. Some of these might represent mechanisms unknown to climate scientists, but it is likely that most of them are spurious patterns. Indeed, because there are thousands of locations and hence tens of millions of possible pairs; thus the chances of finding strong negative correlations among pairs or even regions is quite high. To differentiate interesting dipoles (some of which may be unknown) from spurious ones, a method to evaluate their statistical significance is required. However, to our knowledge there are no such approaches in the literature that can incorporate all the nuances of climate dipoles. In this chapter, we present a novel method for testing the statistical significance of the dipoles. One of the most important challenges in addressing significance testing in a spatio-temporal context is how to address the spatial and temporal dependencies that show up as high autocorrelation. We present a novel approach that uses the wild bootstrap to capture the spatio-temporal dependencies, in the special use case of teleconnections in climate data. Our approach to find the statistical significance takes into account the autocorrelation, the seasonality and the trend in the time series over a period of time. This framework is applicable to other problems in spatio-temporal data mining to assess the significance of the patterns.

5.1.1 Challenges in Significance Testing

Statistical significance testing determines whether a given result is likely to occur by random chance and thus implies whether a result is of statistical importance, and therefore would generalize to other contexts. Historically, significance testing has been widely studied in statistics and there are several classical analytical hypothesis testing methods available. Analytical methods of hypothesis testing such as the *t-test* generally involve computing a test statistic from the observed data and computing a probability value to test if the observed data was derived from a *null hypothesis*. The null hypothesis is rejected in favor of the alternate one if the probability value is below the *significance*

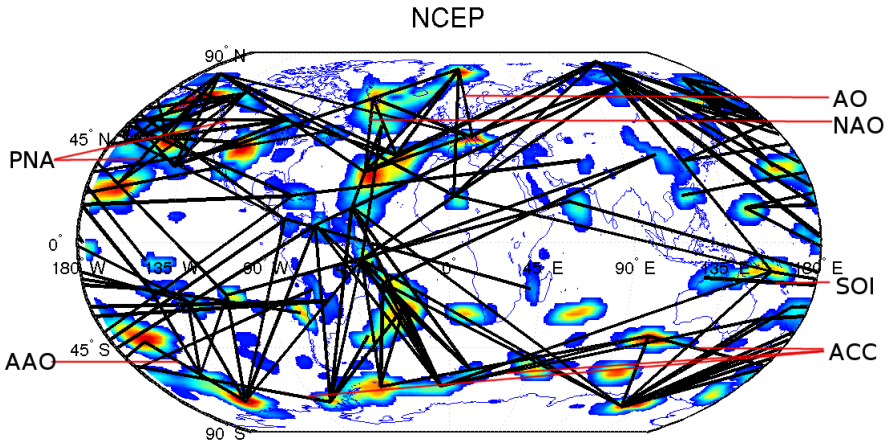


Figure 5.1: Dipole edges with correlation < -0.2 in the NCEP sea level pressure data taken from [47].

level. However, a main drawback of these approaches is that they impose a distribution structure on the data. Technically, t-tests are valid only for i.i.d. normally distributed data and are very sensitive to outliers.

An alternate method of significance testing widely used in data mining is empirical testing using randomization to determine the null model. Randomization tests proceed by following the sequence of steps: (i) rearrange or shuffle the observed value in each sample, (ii) compute the statistics for the randomized data, (iii) repeat it k times (e.g. 1000), and (iv) compare the test statistic generated from the original data and the random distribution to rule out patterns generated by random chance. The intuition behind generating a large sample of the datasets is to create a null model from the data. If the computed test statistics differ widely from the measurements on random datasets then we can reject the null hypothesis and declare the result to be significant.

Randomization tests [21] have been successfully used in many contexts in data mining to find interesting patterns in graphs [35], association rule mining [33], motif mining [15, 27, 55], etc. In ecology, significance testing has been used to study the analysis of species

nestedness patterns [71] and to study the diffusion of a spatial phenomenon [76] and spatial gradients [74]. Monte Carlo tests to test the significance of spatial patterns has been discussed in [13]. However, there are many challenges in using randomization tests for spatio-temporal patterns, some of which are listed below:

1. *Data independence*: One of the underlying assumptions in randomization testing is i.i.d. data. However, in the spatio-temporal context, generally there is a high spatial and temporal autocorrelation and homogeneity, thus violating the assumption of data independence.

2. *Heteroscedasticity*: Heteroscedasticity refers to the problem of different variances in a sub-population and the tests of randomization are sensitive to it. Heteroscedasticity exists in Earth science data in both space and time, i.e., not only the sub-population variances may be different for different locations but they can also vary over time for the same location [81].

3. *Seasonality and trends* In a spatio-temporal context, there are other influencing factors like seasonality, trends, etc. which greatly impact the values in a time series. This can make the tests of randomization either too liberal or too conservative (Type I vs Type II errors). A possible strategy to get rid of trends could be to de-trend the time series. However, de-trending of non-stationary time series data itself has several issues and may result in removing certain dipoles or adding spurious ones, which might require a detailed investigation [23, 34]. Results also depend upon the nature of trends, whether unit roots are present or not, and the nature of possible co-integrating relations, see Engle and Granger [23, 34] for further details. Seasonality is generally handled in climate data by creating an anomaly time series. However, even then there is annual cycle still left in the anomaly time series of some locations on the Earth which could result in the formation of spurious dipoles [79].

4. *Null model* We want the data generating process for drawing random samples to be as close as possible to the true data generating process which generated the observed values. While randomization tests are very often better than simple methods like the t-test, it is very hard to verify the assumption that (and is generally not true that) the multiple datasets created by randomization come from a null model representing the true data generating process.

5.1.2 Our Contribution

To the best of our knowledge, there are no existing approaches for testing the significance of spatio-temporal patterns that systematically model the spatio-temporal data and handle various aspects like auto-correlation, trends, etc. In this chapter, we provide a systematic approach to test the significance of the spatio-temporal teleconnection patterns that overcomes the challenges mentioned above. Our approach uses the general framework provided by the wild bootstrap procedure [82, 53] which is traditionally applied for heteroscedastic problems to present a technique that takes into account the various aspects of climate data like auto-correlation, trends, etc. One novel aspect of our approach is that we translate the space time problem to one where the errors can be modeled as independent but heteroscedastic. We capture the spatial dependence of each region of a dipole via a unified function and capture the temporal dependencies through a first order Markovian distribution. We show the utility of our approach by using it to test the significance of dipoles generated in the NCEP sea level pressure dataset. While we mainly use our algorithm to test the significance of teleconnection patterns, our approach can be instructive to other pattern mining algorithms in the spatio-temporal context to test the significance.

5.2 Approach

As we saw in the previous section, a significance testing based on randomizing time series would not be appropriate for climate data. Instead, it would be more desirable to compute the significance amongst those random series that preserve the same properties as the underlying climate data time series. Our approach for randomization is inspired from the wild bootstrap procedure [82, 53]. The wild bootstrap is a technique where random weights are multiplied to the residuals from the data after fitting a statistical model, then artificial datasets are created using these randomly weighted residuals, and inference is based on repeating the statistical model fitting exercise on these artificial datasets. The wild bootstrap has been mathematically proven to be consistent, and successfully applied to a variety of problems where the data may be heteroscedastic in nature, and the parameter dimension may be large compared to the sample size.

We present a novel approach that uses the wild bootstrap and capture the spatio-temporal dependencies, in the special use case of teleconnections in climate data. First, we develop a small area or state-space type decomposition of the spatio-temporal data to extract the underlying time series that governs teleconnection patterns, against the background of local noise variations. Our approach implicitly takes into account the space dependence of the data as we require each end of the dipole (consisting of many single point locations) to share the same global component. We account for the time dependencies by incorporating an auto-regressive term assuming a first order Markovian dependency in our time series decomposition. Once we extract out the properties (or dominant signals), we test the significance by examining the residual correlation at both the ends of the dipole and thus it helps us in identifying that the negative correlation between the two regions at the two ends is indeed coming from an underlying phenomenon or is just an artifact of the dominant properties. We assign a degree of confidence to our conclusions using a test of randomization. Further details of our approach are mentioned in the following subsections:

5.2.1 Notation

Let A and B represent the two ends of the dipole and let n_A and n_B represent the number of points at the two ends. Let $X_{it}\ t = 1, \ldots, T, i = 1, \ldots, n_A$ represent the time series for T time steps at the n_A points of region A. Similarly, let $Y_{it}\ t = 1, \ldots, T, i = 1, \ldots, n_B$ represent the time series for T time steps at the n_B points of region B.

5.2.2 Step 1: Time Series Decomposition

The first step in the significance testing of dipoles is a temporal decomposition that captures the spatial as well as the temporal bindings of the two ends of the dipoles. We begin by noting two key properties of the dipole anomaly time series.

1. **Trend**: Many locations on Earth experience a general linear trend in their anomalies over time. For some locations, the trend increases and for some it decreases over time and this pattern can vary with different magnitude at different locations.

2. **Seasonality**: Typically, Earth science data has seasonality in it. Apart from the annual seasonality which is accounted for by constructing the anomaly time series, the data typically has sinusoidal patterns of various periodicities and of varying strengths across regions. If we examine the periodicities of the anomaly time series using the power spectrum, we see that quite of few of them have a period of 12 months [79].

In order to model these two key characteristics of dipole locations, we propose a temporal function $f(t)$, defined as follows:

$$f(t) = \alpha + \beta t + \gamma \sin\left(\frac{2\pi(t+\delta)}{12}\right) \tag{5.1}$$

The function $f(t)$ captures the trend through the βt component and the seasonality through the $\gamma \sin(\cdot)$ component. The α component ensures that the constant effect due to altitude, latitude and other unknown phenomena is also captured. $f(\cdot)$ only captures

the temporal fluctuations at a given location independent of any spatial or temporal bindings.

Recall that a dipole consists of two regions, A and B, with opposite climate phenomenon. All the locations in a given region have a highly positive correlation in their anomalies and they are driven by the same underlying phenomenon. Let that underlying phenomenon for a specific end of dipole (say A) be indicated by U, where size of U is $T \times 1$. This results in the following linear heteroscedastic decomposition:

$$\forall_{i \in A}\ X_i = U + r_i \tag{5.2}$$

where r_i is the error term representing the local phenomenon at a location $i \in A$. Moreover, depending on how far a location i lies from the dipole center, its anomaly time series would be influenced accordingly. Let $w(i)$ indicate the weight or influence of U on X_i. The goal in this case is to reduce the residue of a given region.

$$SE_r = Tr\left[(X - U\mathbf{1})^T\ W\ (X - U\mathbf{1})\right] \tag{5.3}$$

where X is a $T \times N$ matrix with column i indicating anomaly time series of location $i \in A$, $\mathbf{1}$ is a matrix of size $1 \times N$ with all elements $= 1$, W is a diagonal matrix with $W_{ii} = w(i)$.

Equation 5.2 allows us to capture the spatial bindings of a dipole region and provides a unified anomaly time series U. It does not capture the temporal correlations of U. In order to do this, we consider the following auto-regressive formulation:

$$U_t = f(t) + \phi[U_{t-1} - f(t-1)] + \epsilon_t \tag{5.4}$$

Similar to equation 5.3, we aim to reduce the residue ϵ, such that the decomposition captures all the spatial and temporal properties of the dipole. We define the squared error of ϵ as

$$SE_\epsilon = \sum_t (V_t - \phi V_{t-1})^2 \tag{5.5}$$

where $V_t = U_t - f(t)$. The mathematical properties of the dipole detection algorithm is primarily governed by the bivariate time series

$$\mathbf{V}_t = \begin{pmatrix} V_{At} \\ V_{Bt} \end{pmatrix}, t = 1, 2, \ldots T.$$

This is a non-stationary time series, since the innovations for this time series are given by the independent bivariate random variables

$$\epsilon_t = \begin{pmatrix} \epsilon_{At} \\ \epsilon_{Bt} \end{pmatrix} \overset{ind}{\sim} \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{At}^2 & \rho_{AB}\sigma_{At}\sigma_{Bt} \\ \rho_{AB}\sigma_{At}\sigma_{Bt} & \sigma_{Bt}^2 \end{pmatrix} \right).$$

A variation of the Kolmogorov consistency theorem is used to establish the existence of the second order stochastic process $\{\mathbf{V}_t\}$. The properties of the dipole are dictated by the innovation correlation coefficient ρ_{AB}, which takes a high negative value for true dipoles.

We model $\mathbf{V}_t = \mathbf{\Phi}\mathbf{V}_{t-1} + \epsilon_t$ where we assume $\mathbf{\Phi}$ is a diagonal matrix with diagonal entries ϕ_A and ϕ_B. The deterministic trend functions $\{f_A(\cdot)\}$ and $\{f_B(\cdot)\}$ and the local noise perturbation terms $\{r_{Ai}(\cdot), i = 1, \ldots, n_A\}$ and $\{r_{Bi}(\cdot), i = 1, \ldots, n_B\}$ do not contribute towards the properties of a dipole, but are important nuisance factors in studying dipoles. Needless to say, we could adopt a more complicated model for the time series properties of $\mathbf{V}_t$, the deterministic trends or the local noise, and include co-integration and other complex features. However, in the context of the present application, such additional complexity seems unnecessary.

Our aim is to reduce the squared error SE_r and SE_ϵ and we do it by minimizing them in turn. The residue term ϵ_t represents error that is independent and heteroscedastic. Thus we are able to effectively translate the space time problem into one where we are able to model the errors as independent but heteroscedastic. We use a simplistic approach to obtain an approximate solution that minimizes Equation 5.3 and 5.5. The idea is to minimize SE_r independent of U_t's auto-regressive property and obtain estimates of

α, β, γ for a fixed choice of δ. After that, using U_t's auto-regressive properties estimate ϕ and compute ϵ. The attractive property of this approach is that it leads to a closed form solution for the parameters. We get,

$$\sum_t g_k(t) \cdot f(t) = \frac{\sum_{i,t} g_k(t) \cdot w(i) \cdot X_{it}}{Tr(W)}, k = 1, 2, 3 \tag{5.6}$$

where $g_1(t) = 1$, $g_2(t) = t$, $g_3(t) = \sin(\frac{2\pi(t+\delta)}{12})$. The three equations can be easily solved for a fixed δ using linear regression. Additionally, we get a closed form for ϕ as,

$$\phi = \frac{\sum_{t=2}^{T} V_t \cdot V_{t-1}}{\sum_{t=2}^{T} V_{t-1}^2} \tag{5.7}$$

In order to estimate the optimal δ, we begin with an estimate by varying it from $1, \ldots, 12$ and pick the one that minimizes $E[\epsilon_t]$.

5.2.3 Step 2: Residual correlation

After finding the residue at each end of the dipole, our next goal is to examine the residual correlation at the two ends of the dipole to check if the regions involved form a true dipole. The residue at the two ends represents the time series signal after extracting trend and the seasonality. We compute the pairwise correlation ρ_{ij} between all the nodes in ϵ_{it} and ϵ'_{jt}. We can use the raw correlation values to test the significance of the dipoles. However, we use a more stable transformation provided by Fisher to transform the correlation into Z_{ij} as described in the following subsection.

Fisher transformation

The Fisher transformation [28] is generally used in statistics to test the hypothesis about the correlation coefficient ρ between two variables. The transformation changes the probability density function (pdf) of any waveform so that the transform output

has an approximately Gaussian pdf. The transformation is defined as follows:

$$Z_{ij} = \frac{1}{2} log \frac{1+\rho_{ij}}{1-\rho_{ij}} \tag{5.8}$$

The Fisher transformation is a variance stabilizing transformation and converges to a normal distribution much faster.

5.2.4 Step 3: Assessing dipole statistical significance

In testing the significance of dipoles, the null hypothesis means that the dipole pattern is spurious or uninteresting. Our task is to generate the p-value to specify a confidence measure on whether the dipole is significant. Using our time series decomposition, we devise the following method of randomization inspired from the wild bootstrap algorithm [36] in which re-samples are generated by multiplying random noise to the residuals in order to preserve heteroscedasticity. The details of the steps are mentioned as follows:

1. Step 1: Compute the time series decomposition and the parameters, α, β, γ and δ. Compute the residue ϵ_A and ϵ_B at the two ends and the Fisher transformed correlation Z_{AB}.

2. Step 2: Generate random perturbations in the residual data such that the variance of the residual data is still σ_ϵ^2. This can be done by multiplying i.i.d. random noise $\mathcal{N}(0,1)$ to the original residue ϵ_A and ϵ_B.

$$\begin{aligned} E[(\psi\epsilon - E[\psi\epsilon])^2] &= E[(\psi\epsilon)^2] - (E[\psi]E[\epsilon])^2 \\ &= E[(\psi\epsilon)^2] = \sigma_\epsilon^2 \end{aligned}$$

   here we have used $E[\psi] = 0, E[\psi^2] = 1, E[\epsilon] = 0$.

3. Step 3: Recompute $X_{it}^{'}$ and $Y_{it}^{'}$ using α, β, γ and δ.

4. Step 4: Recompute the decomposition to generate $\alpha^{'}$, $\beta^{'}$, $\gamma^{'}$ and $\delta^{'}$. Compute the residue $\epsilon_A^{'}$ and $\epsilon_B^{'}$ at the two ends and the Fisher transformed correlation $Z_{AB}^{'}$.

5. Step 5: Repeat steps 2 to 5 $N = 10,000$ times and generate the p-value as follows:

$$p_{AB} = \frac{1}{N} \sum_{i=1}^{N} I_{(Z_{AB} \geq Z_{AB}^{'})} \tag{5.9}$$

Let $Z_{AB} = \frac{1}{2} log \frac{1+\rho_{AB}}{1-\rho_{AB}}$ and similarly define $\hat{Z}_{AB}$, and $T_n = T^{1/2}(\hat{Z}_{AB} - Z_{AB})$ where T is the observed length of the time-series. From the wild-bootstrap based generation, we obtain similar estimates from each resample, and let $\hat{Z}_{AB}^{*}$ be the equivalent of $\hat{Z}_{AB}$ from the resample. Define $T_n^{*} = T^{1/2}(\hat{Z}_{AB}^{*} - \hat{Z}_{AB})$. We have the following result as the theoretical counterpart of our algorithm:

Theorem 5.2.1. *In the framework presented above, the following hold:*

1. *The distribution of the statistic T_n converges weakly to the standard Normal distribution $N(0,1)$.*

2. *The distribution of the statistic T_n^{*}, conditionally on the observed data from regions A and B, converges weakly to the standard Normal distribution $N(0,1)$ almost surely.*

The second part of the above theorem states that for all possible data sets arising from regions A and B, the convergence of the wild bootstrap-based statistic T_n^{*} to the same distribution as that of the original statistic T_n is guaranteed with probability one.

5.2.5 Step 4: Multiple Hypotheses

Multiple comparisons is an important issue in dipole significance testing as there are a set of statistical inferences computed simultaneously. Multiplicity leads to false positives or the type I errors, i.e., the errors committed by incorrectly rejecting the null hypothesis. In order to control the false discovery rate (FDR), we use the standard procedure

by Benjamini-Hochberg-Yekutieli [12] which controls the false discovery when the m hypothesis tests are dependent, which is true in our case. The method refines the threshold of p-values to find the largest k such that:

$$P_{(k)} \leq \frac{k}{m \cdot c(m)} * \alpha \tag{5.10}$$

We compute c(m) by examining the correlation between the 10000 random values generated for each end of the dipole. As they are positively correlated, we set c(m) to 1. We discard all the dipoles having a p-value less that $P_{(k)}$.

5.2.6 Dataset

We use the data from the NCEP/NCAR Reanalysis project provided by the NOAA/ESRL [43]. We use the monthly resolution of data and it has a grid resolution of 2.5° longitude x 2.5° latitude on the globe. We use the sea level pressure (SLP) data to find the dipoles because most of the important climate indices are based upon pressure variability. For the analyses and results presented here, we use the 50 year of data starting from 1951 to 2000.

5.2.7 Results

We ran the dipole algorithm using the NCEP dataset and the algorithm mentioned in [47] and obtained all the dipoles at a correlation threshold of -0.25. We ran our approach on significance testing for this data to generate the residual correlation and obtained the p-values. Fig. 5.2 shows the scatter plot of original correlation versus the residual correlation amongst the dipoles found in the dataset. From the figure, we see that a high negative original correlation does not necessarily transform to a high negative residual correlation. Fig. 5.3(a) shows an example of a dipole having an original correlation of -0.32 but a residual correlation of 0.1359 (p-value = 1). If we examine

the time series at the two centres of the dipole (see Fig. 5.3(b), we see that there is a linear trend in the opposite direction which the model is able to effectively capture. Fig. 5.4 shows an example dipole that did not have significant trends but was discarded due to the seasonality component γ. The original correlation of the dipole is -0.24, whereas the residual correlation is 0.0219.

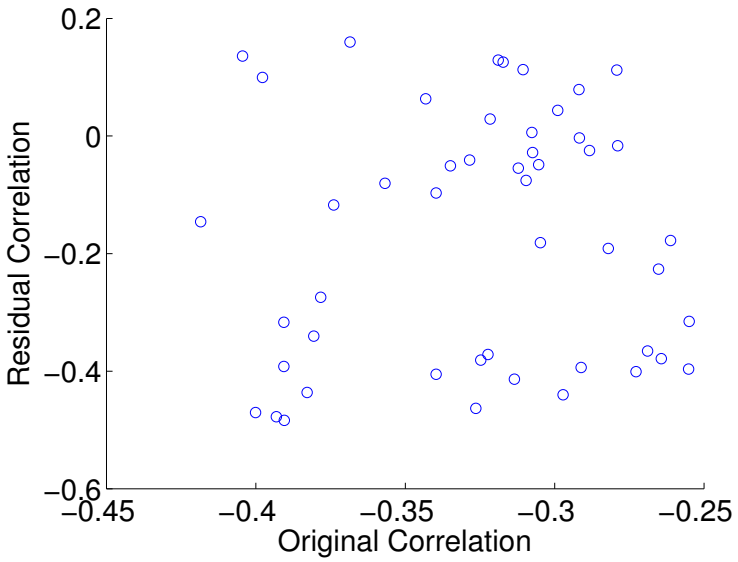
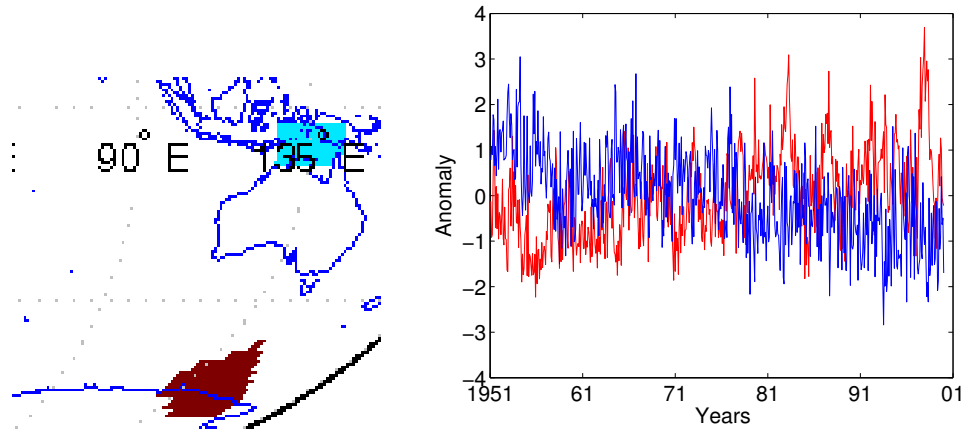


Figure 5.2: Scatter plot showing original vs residual correlation.

On the other hand, Fig. 5.5 shows an example of a dipole that had an original correlation of -0.25 but has a higher negative residual correlation of -0.39 (p-value $= 0$). This dipole represents one of the known connections AAO and has a correlation of 0.8 with the AAO index defined by the CPC [1]. We see that the approach effectively eliminates about 16 dipoles with a p-value ≥ 0.01. Further, it declares all the known connections as significant. However, we see that there are still a few dipoles (10) left that require post-processing which is described below.



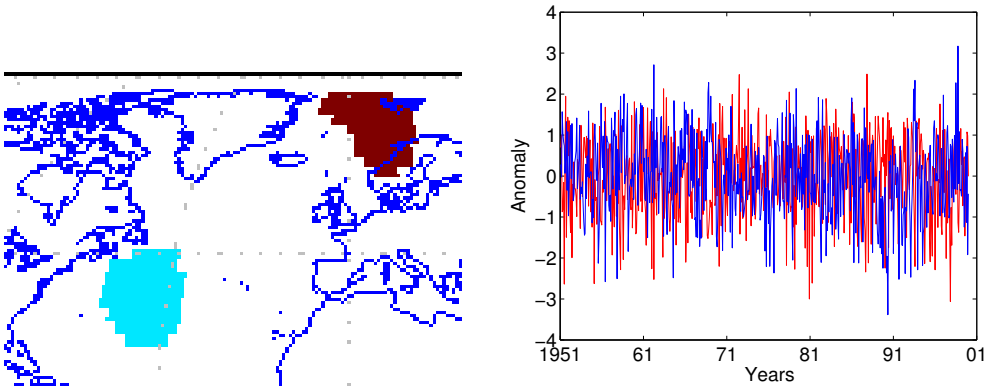
(a) Dipole having a correlation of −0.32 but a residual correlation 0.1359

(b) Time series at the two centres

Figure 5.3: Dipole rejected due to linear trend.

5.2.8 Post Processing Using Domain Knowledge

Our model for deterministic trend accommodates a linear function and a sinusoidal component at each end-point of a potential dipole. A careful analysis of some of these time series show that non-linear trends may occasionally exist. Fig. 5.6(a) shows an example of a dipole that had an original correlation of −0.39 but has high non-linear trends. Fig. 5.6(b) shows the time series at the two centres of the dipole. From the figure, we see that the trends in the two dipole ends are not linear, thus making the post processing necessary. One end of the dipole corresponds to the Sahel region in Africa which underwent an abrupt change a long period of drought around 1969 [40] which is also reflected in the time series as shown in the Fig. 5.6(b). Based on domain knowledge and prior experience, we know that this dipole does not make physical sense. De-trending the data before applying the dipole detection algorithm might appear to be a solution. However, as we discussed earlier, detrending of climate data has many challenges and can lead to adding spurious connections especially when the trends are non-linear. Using domain knowledge, we want to further eliminate these trend dipoles



(a) Dipole having an original correlation of −0.24 but a residual correlation of 0.0219

(b) Time series at the two centres

Figure 5.4: Dipoles discarded due to seasonality filtering.

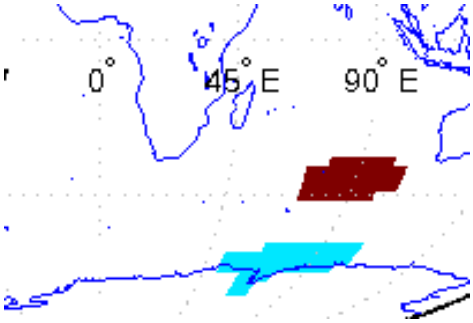


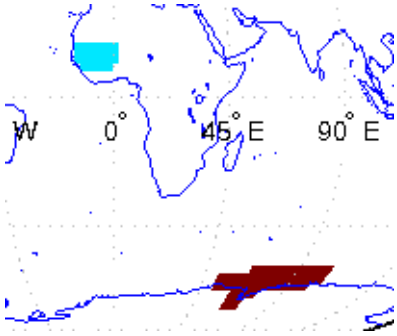
Figure 5.5: Dipole having an original correlation −0.25 but a residual correlation −0.39 corresponds to the known dipole AAO.

in order to identify the real dipole structure.

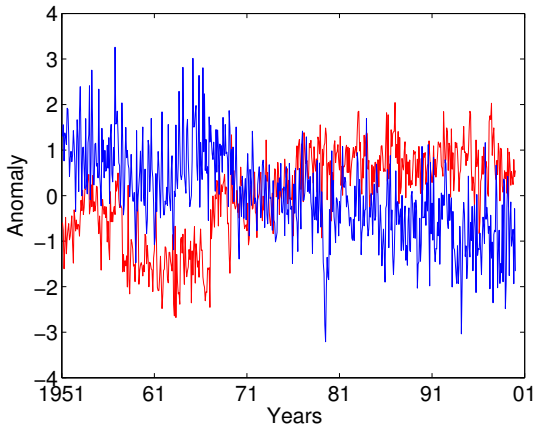
Our parametric form comes to rescue in this case as this allows us to put bound on the value β can take. We use a simple method to examine the β values at the two ends. If the difference in β values at the two ends of the dipole is greater than a threshold, we discard them.

$$Discard\ dipoles\ if\ |\beta_A - \beta_B| \geq \hat{\beta} \tag{5.11}$$

In order to compute $\hat{\beta}$, we considered the 6 well known dipoles and computed the absolute difference in their beta values and selected our threshold of $\hat{\beta}$ based upon that.



(a) Dipole having a correlation of −0.39



(b) Time series at the two centres of the dipole

Figure 5.6: Dipole showing non-linear trend corresponding to abrupt change due to the Sahel drought

With the use of this filtering, a number of trended dipoles mainly starting from the Sahel region in Africa that initially passed the significance threshold were eliminated. This intuitively makes sense because our parametric form removes trend as well as cyclic patterns from the data but not the small local oscillations (which are captured by ϵ). There is a possibility that one end of the spurious dipole is influenced by one end of a true dipole. In this case, those local oscillations as captured by epsilon could be of opposite polarity and hence manage to pass the significance test. The above filtering mechanism using $\hat{\beta}$ seems like a simple way to eliminate such cases.

5.2.9 Comprehensive Evaluation

Table 5.1 shows the summary of the number of dipoles declared as significant using a significance level $\alpha = 0.01$ and the post-processing that we described above. From the table, we see that 23 dipoles are declared as significant in the dataset having a correlation < -0.25. Figures 5.7 shows the dipoles declared significant in the NCEP dataset at a threshold of -0.25. From a quick visual inspection of the figures, we see that the

	NCEP -0.25		NCEP -0.2	
Total	49		85	
	$p < 0.01$	$p \geq 0.01$	$p < 0.01$	$p \geq 0.01$
No trends	**23**	4	**31**	13
Trends	10	12	23	18

Table 5.1: Number of dipoles declared as significant using our approach in the NCEP data.

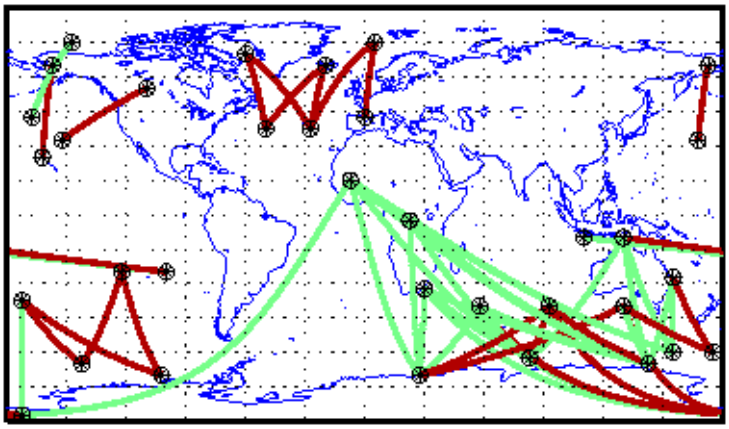


Figure 5.7: Dipoles declared significant in the NCEP dataset at a threshold of -0.25. Red denotes significant dipoles and green denotes insignificant dipoles.

well-known dipoles like North Atlantic Oscillation (NAO), Southern Oscillation (SO), Western Pacific (WP), Pacific North America pattern (PNA) and Antarctic Oscillation (AAO) are all identified as significant. Fig 5.8 shows the dipoles declared as significant at a lower threshold of -0.2. From the figure, we see that apart from the well known dipoles, other weaker connections start appearing as significant, for example the Scandinavia pattern starting around Russia and ending at the Atlantic.

Our next goal is to check whether our algorithm has a bias to declare dipoles having a higher negative correlation as significant. Fig. 5.9 shows the histogram of correlation values of dipoles declared as significant and insignificant in the NCEP data. The histogram shows that at times the algorithm even declares dipoles with higher negative correlation as insignificant. However, using our approach, we are still able to remove

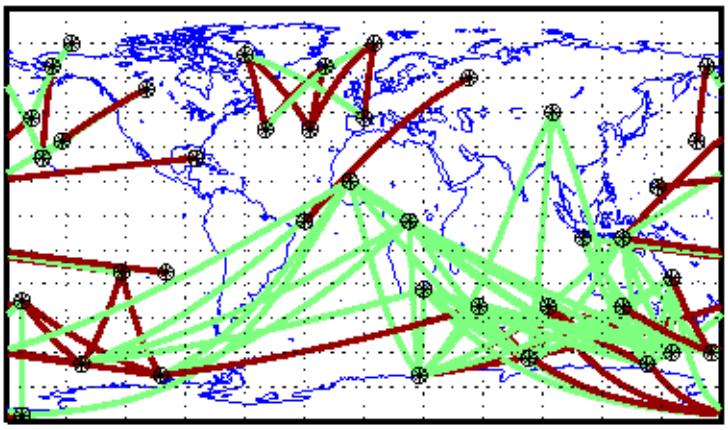


Figure 5.8: Dipoles declared significant in the NCEP dataset at a threshold of -0.2. Red denotes significant dipoles and green denotes insignificant dipoles.

about 1/2 of the dipoles from the NCEP data having a correlation < -0.25 as insignificant. Also the histogram of correlations of significant and insignificant correlations shows that the algorithm has no particular bias. Next, we examine closely the two reasons in our algorithm to label the dipoles as insignificant.

Recall, that the β values capture the linear trend present in the data. Spurious dipoles can be formed if the two regions involved in the dipole have significant trends in the opposite direction and the negative correlation between the two regions is accounted for by the negative trends and not a periodic oscillation. Fig. 5.10 shows a plot of β values for the NCEP dataset. From the figure, we see that there are quite a few dipoles with strikingly opposite trends in the NCEP data and most of them going to the southern hemisphere. This also conforms with the existing knowledge about the NCEP data from the climate science [38] about the presence of significant spurious trends in the southern hemisphere.

Table 5.1 shows that half of the rejected dipoles have significant trends in the opposite direction. Apart from the dipoles with trends, the other dipoles which are discarded using our algorithm are the ones with very little negative residual correlation left in

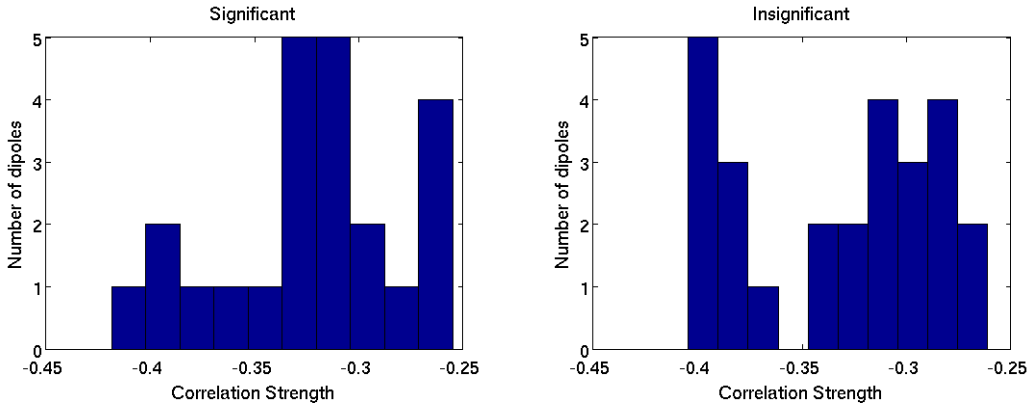


Figure 5.9: Histogram of correlation strengths for significant and insignificant dipoles.

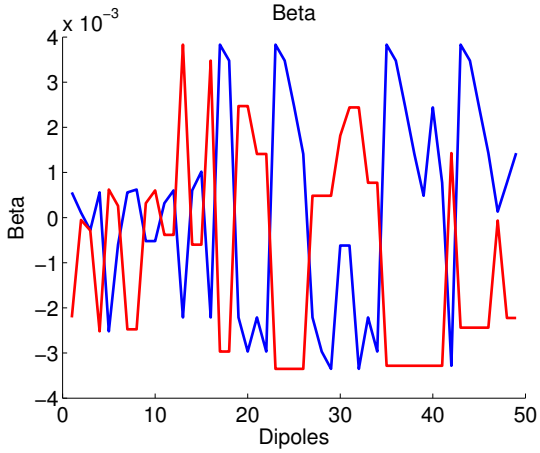


Figure 5.10: Beta values at the two ends of the dipoles for the NCEP dataset.

them (see Table 5.1). Seasonality in the dipoles could be one possible reason. Fig. 5.11 shows the gamma values of the dipoles. From the figure, we see that quite a few of them have significant value of gamma.

p-value for the known dipoles

A good measure of evaluation is to examine the p-values generated for the 6 of the most well known dipoles - SOI, NAO, AO, AAO, WP, PNA. The existence and the

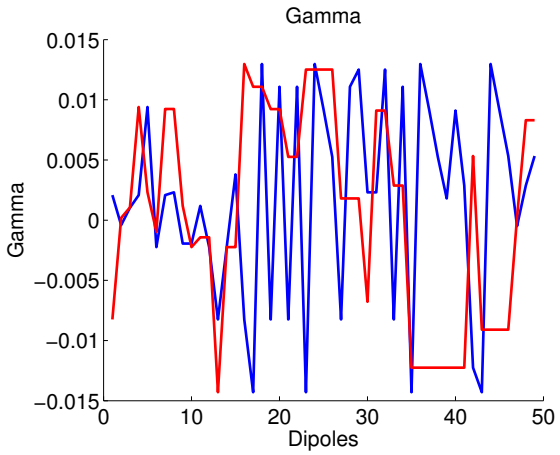


Figure 5.11: Gamma values at the two ends of the dipoles for the NCEP dataset.

	NCEP/NCAR		
	p-value	Residual Corr	Fisher transform
SOI	1.2897e-13	-0.1814	-10.3147
NAO	0	-0.4137	-54.2486
AO	0	-0.3092	-19.913
AAO	0	-0.39	–32.4990
WP	0	-0.1755	-9.2258
PNA	0	-0.0968	-8.4194

Table 5.2: p-values for the known dipoles using the random approximation along with residual correlation.

impact of these dipoles has been well established in literature from climate science. At first, we pick up a data driven dipole which represents the static index of the closest in correlation. After that, we examine the p-values generated for the data driven dipole closely matching the static index. Table 5.2 shows the p-values generated for the known dipoles using our approach. From the table, we see that all of the 6 well known dipoles are declared significant using our algorithm and have a p-value of 0 up to the order of machine precision. Further the residual correlation at the two ends of the dipole generated by removing $f(t)$ from the time series at the two ends is also very highly negative for the known dipoles. This provides empirical evidence that our approach to estimate the statistical significance works well in practice.

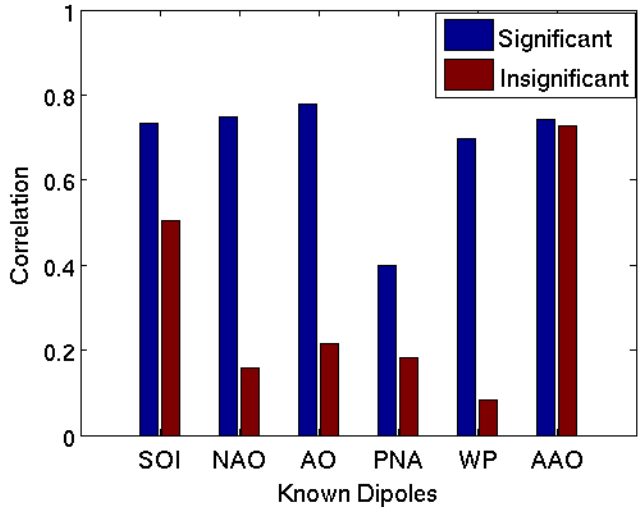


Figure 5.12: Maximum correlation with known indices in the two sets of dipoles.

Correlation with Static indices

In order to further assess the quality of the extracted dipoles, we did another experiment to understand the nature of the dipoles. Most of the candidate dipoles should be a representative of some known phenomenon. We considered 6 teleconnection patterns identified by the Climate Prediction Centre website [1]. From the NCEP data, we considered two sets of dipoles significant and insignificant. There were about 25 dipoles in each subset. We computed the correlation of each of these dipoles with the 6 known climate indices. Fig. 5.12 shows the maximum correlation of the two groups of dipoles with the known indices. From the figure, we see that all the surrogates of the known phenomenon are captured very well in the significant group as compared to the insignificant one. PNA is not captured with a very high correlation in both the groups as the actual phenomenon consists of three epicenters and is not a dipole. AAO has high correlation with significant as well as insignificant group. This might be due to trends in the insignificant group.

A new dipole ?

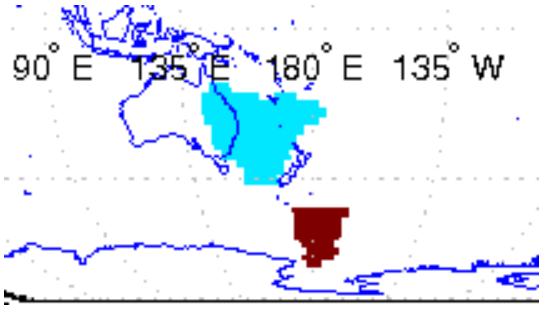


Figure 5.13: Dipole near Australia shows up as statistically significant.

A larger implication of our work on significance testing lies in identifying potentially new teleconnection patterns not known to climate scientists so far. A careful evaluation of all the dipoles from Fig. 5.7 shows that most of them have a very high correlation with the known climate indices and thus are some variant of the known phenomenon. However, there are some teleconnection patterns that are declared as significant and that do not have a high correlation with any known phenomenon. One such striking dipole is a dipole near Australia as shown in the Fig. 5.13. It appears as significant in the NCEP data and its correlation with the known indices is also very low (see Fig. 5.14). Further, it is not supported by the existing literature on teleconnections. This might represent a new dipole phenomenon not known to climate scientists so far. Our preliminary investigations show that this dipole also has a different impact on land temperature as compared to other known dipoles. A comprehensive evaluation of the physical significance of the phenomenon is a part of our future work.

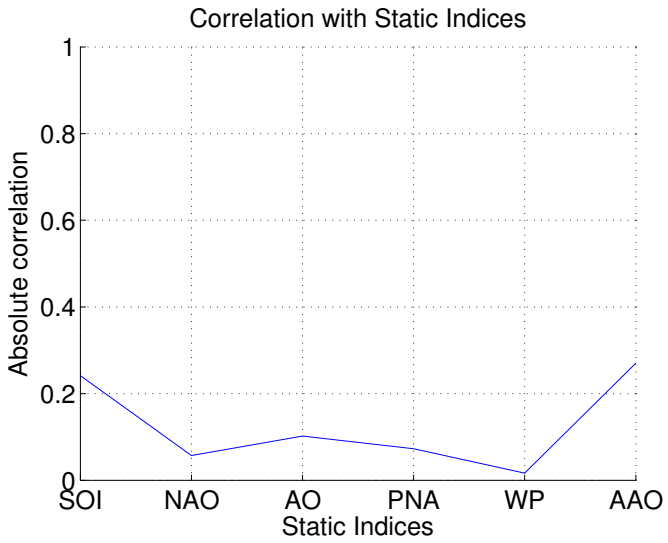


Figure 5.14: Correlation of the dipole near Australia with known indices

5.3 Conclusion

Significance testing in spatio-temporal data presents many challenges due to the inherent autocorrelation dependencies in time and space. However, significance testing of spatio-temporal patterns has received little attention. In this chapter, we present a systematic approach to detect the significance of spatio-temporal teleconnection patterns. We ran our algorithm on the NCEP sea level pressure data. From our results, we see that our algorithm is able to capture the known dipoles. We show the utility of using a simple model to extract out the characteristics of climate data time series. A larger implication of our work is that the algorithm can be instructive to other researchers in the spatio-temporal domain to test the significance of patterns. A part of the future work involves handling non-linear trends. Another limitation of the model is that the marginal analysis of the periodic component distort co-periodicity properties. We propose to address this in our future research work. In particular, we propose to simultaneously model the deterministic trends and periodic components at the two ends of a dipole, along with the stochastic components of the bivariate time series. Two-dimensional wavelets would

be used for this purpose, since evidence shows some erratic patterns and discontinuities. Also, as part of our future work, we would like to explore if some potential dipoles are governed by co-integrating relations. We also propose to explore the choice of resampling weights for which the wild bootstrap inference would be second order accurate. Another future direction is to integrate the significance testing into the algorithm for dipole detection and thus not allow spuriously connected regions to be declared as candidate dipoles.

Chapter 6

Anomaly Construction in Climate Data [1]

6.1 Introduction

In this chapter, we discuss some of the issues in handling climate datasets. We examine how different methods of preprocessing can impact the output of the detected the algorithm. An important first step in handling climate datasets is anomaly construction. Earth science data consists of a strong seasonality component as indicated by the cycles of repeated patterns in climate variables such as air pressure, temperature and precipitation. The seasonality forms the strongest signals in this data and in order to find other pattens, the seasonality is removed by subtracting the monthly mean values of the raw data for each month. However since the raw data like air temperature, pressure, etc are constantly being generated with the help of satellite observations, the climate scientists usually use a moving reference base interval of some years of raw data to calculate the

[1] **This chapter is based on the work [44] published in the proceedings of the Conference on Intelligent Data Understanding, CIDU 2011.**

mean in order to generate the anomaly time series and study the changes with respect to that. In this chapter, we evaluate different measures for base computation and show how an arbitrary choice of base can skew the results and lead to a favorable outcome which might not necessarily be true. We perform a detailed study of different base selection criterion and base periods to highlight that the outcome of data mining can be sensitive to choice of the base. We present a case study of the dipole in the Sahel region to highlight the bias creeping into the results due to the choice of the base. Finally, we propose a generalized model for base selection which uses Monte-Carlo based methods to minimize the expected variance in the anomaly time-series of the underlying datasets. Our research can be instructive for climate scientists and researchers in temporal domain to enable them to choose the right base which would not bias the outcome of the results.

A main component of Earth Science data is the seasonal variation in the time series. Seasons occur due to the revolution of the Earth around the Sun and the tilt of the Earth's axis. The change in seasons brings about annual changes in the climate of the Earth such as increase in temperature in the summer season and decrease in temperature in the winter season. The seasonality component is the most dominant component in the Earth science data. For example, consider the time series of monthly values of air temperature at Minneapolis from 1948-1968 as shown in Figure 6.1. From the figure, we see that there is a very strong annual cycle in the data. The peaks and valleys in the data correspond to the summer and winter season respectively and occur every year. The seasonal patterns even though important are generally known and hence uninteresting to study. Mostly, scientists are interested in finding non-seasonal patterns and long term variations in the data. As a result of the effect of seasonal patterns, other signals in the data like long term decadal oscillations, trends, etc. are suppressed and hence it is necessary to remove them. Climate scientists usually aim at studying deviations beyond the normal in the data.

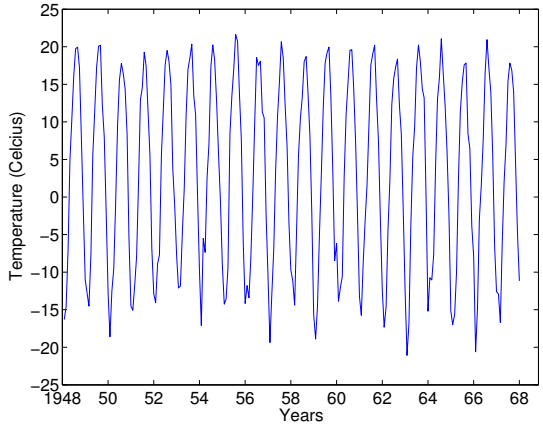


Figure 6.1: The figure shows the monthly mean air temperature at Minneapolis for a 20 year period. From the figure we can see that there is a very high annual cycle and the temperatures go up and down with the change of seasons.

In order to remove seasonality from the raw data, climate scientists generally remove the monthly mean value from the raw data. For example, although more than 100 years of data are available for the temperature anomaly time series at the National Climatic Data Center, only the 100 years 1901-2000 are used to calculate the annual cycle [9]. Often, climate scientists only take 30 years as a reference interval and construct anomalies with respect to that interval. There are several important results and implications derived from the anomalies constructed using a short reference base. In general, climate data has complex structures due to spatial and temporal autocorrelation. The choice of the base significantly impacts the patterns that can be discovered from it and some really important climate phenomenons are computed using a fixed base. For example, teleconnections or long distance connections between two regions on the globe are represented by time series called *climate indices*. Climate indices are time series that summarize the behavior of the selected regions and are used to characterize the factors impacting the global climate. These climate indices are computed by the Climate Prediction Center [1] using a moving 30-year base period and currently they use a base period of 1981-2010. Another important set of results computed using a fixed base are

incorporated in the International Panel on Climate Change (IPCC) Fourth Assessment Report on understanding climate change[56].

In this chapter, we show how an arbitrary choice of base can skew the results and lead to favorable outcome which might not necessarily be true. We examine four simple criterions for base selection and empirically evaluate the differences in them. Our empirical evaluation of the different measures reveals that the z-score measure is quite different from the other measures like mean, median and jackknife. We further study the impact of using different base period to highlight that the outcome of further analysis can be sensitive to the choice of the base. We present a case study of the Sahel region to show that the dipole in precipitation in the region moves around and even disappears with the choice of a different base. Finally, we propose a generalized model for base selection which uses Monte-Carlo based sampling methods to minimize the expected variance of the underlying datasets. Our research can be especially instructive to climate scientists in helping them construct a generalized anomaly that does not create a bias in their analysis. Further, other researchers in temporal domain can also benefit from our work and it will enable them to choose a bias-free base. The main contributions of this work are as follows:

- We present a systematic evaluation of four different measures of computing the base to construct the anomalies. Our evaluation shows that using the mean for anomaly computation might not be the right thing to do.
- We show that using a short base reference introduces a bias in the variance and show an alternative approach to take care of the bias.
- We present a case study of Sahel region to highlight that outcome of further analysis like dipole detection can be sensitive to the selection of the base.
- We propose an algorithm based on Monte Carlo sampling for automatic selection of the appropriate base which minimizes the variance across the time series. The

algorithm suggests using weighted base of 55 year time period rather than 30 years for our data spanning 62 years.

6.2 Related Work

Anomaly computation is a fundamental problem in climate science as most of the analysis of climate data relies upon computation of the anomalies as the first step. There have been some studies in the climate domain analyzing anomalies. Climate scientists mostly use a 30 year period to construct the anomalies and remove the annual cycle. Other ways to remove the annual cycle are 1) computing second moment statistics over each individual season by removing the first two harmonics of the respective time series; and 2) averaging the second moment statistics over all years. More techniques to remove the annual cycle include removing the first two or three harmonics (periods of 365.25, 182.625, and 121.75 days) e.g., [42] and [17]. Some of the less common practices involve looking at more sophisticated techniques like removing the cyclostationary empirical orthogonal function [51] or bandpass filtering, e.g. using a low-pass filter with 0.5 cycle/year [65]. More general methods are described in Wei et al. [80].

However, these procedures fail to take into account the natural interannual variability that should remain visible in the data. Therefore the procedures result in biased estimates of certain statistics [67]. In particular, lag-autocorrelations are systematically negatively biased, which indicates that uncertainty is added to climate data. Trenberth [67] shows for first order autoregressive time series that the autocorrelations computed after the annual cycle is removed become negative after just a few days lag. Consequently, the stochastic character of meteorological time series can result in less statistically significant analysis. Kumar *et al.* [52] state that the analysis of observed climate data often lacks separation of the total seasonal atmospheric variance into its

external and internal components, with external components being the influence of atmospheric initial conditions, the coupled air-sea interactions, and boundary conditions other than sea surface temperatures, whereas internal components are described by the atmospheric variability over time. Removing the annual cycle should provide insight into the internal variability while leaving the external forcing intact.

Tingley *et al.* [66] discuss the impact of using a short reference interval in anomaly construction. They show that using a short reference period, the variance of the records at the time interval is reduced and inflated elsewhere. They show that the choice of the reference interval has a significant impact on the second spatial moment of the time series in the temperature data set whereas the first moment of the time-series is largely unaffected . They further use two factor ANOVA model within a Bayesian inference framework.

Despite the importance of anomalies in the further impact on the results, there is no firm consensus on how to deal with the systematic construction of anomalies and their impact on the various results. Apart from this, the authors are not aware of any systematic study comparing the different aspects of anomaly construction in the climate data.

6.3 Different aspects of Anomaly Construction

We examine two aspects of anomaly construction: 1) the measure for anomaly construction and 2) the period used for anomaly construction in the following subsections.

6.3.1 Different measures for Anomaly Construction

The central idea behind anomaly construction is to split the data into two parts: (a) data with expected behavior, and (b) anomaly data that shows the variability from the expected, which is generally used for understanding climate change phenomenon. For a

given location i, its anomaly times series f_i' is constructed from the raw time series f_i by removing a base vector b_i from it as follows:

$$f_i' = f_i - b_i \tag{6.1}$$

A simple measure of computing the base b_i is by taking the mean of all data ($\overline{f}_i$) present for location i. However the sample mean would not be a good measure as the Earth science data is associated with a large amount of seasonality. In order to account for this the base b_i is computed by taking a monthly mean for each month separately. It is not yet clear whether the mean is the right way to compute the base or if there is a better measure to compute the base. We examine four simple measures of base computations as follows:

- **Mean**: In this measure, the monthly mean values of the raw data are considered as the base and subtracted from the data to get the anomaly series.
- **Z-score**: Another possible way to construct the anomalies is to remove the monthly z-score values from the raw data. The z-score also accounts for the standard deviation in the monthly values.
- **Median**: This is constructed by removing the monthly median values instead of the monthly means as median can be a more robust measure when the data is skewed.
- **Jackknife**: This approach involves considering all points apart from the point itself in the computation of the mean and variance measures and it produces an unbiased estimation of variance just like Maximum aposterior Estimate (MAP).

We elaborate these measures and how they are computed in the following sub-sections.

Mean

Monthly mean computation is the most widely used method to extract the anomalies from the raw data. The mean subtraction makes the anomaly time series to have a zero mean. More formally,

$$f'_i(t,m) = f_i(t,m) - \mu_m, \forall t \in \{total - start, \ldots, total - end\}, \forall m \in month \quad (6.2)$$

where total-start and total-end values represent the actual size of the data. In general it is known that taking mean would minimize the variance of the resulting series but it can also lead to over fitting and conclusions that might not be true. Further, instead of using the entire data for base computation, a short reference interval can be chosen. For example, if the data begins from 1900 to 2010, the base start and end years could be chosen as 1960-1990. We further discuss the issue of choosing a short base in Section.6.3.2.

Z-score

The z-score normalization ensures that the resulting anomaly series has mean = 0 and standard deviation = 1. As a result, z-score can be considered to be more robust than the mean but at the same time z-score based standardization can eliminate variations across different locations on Earth which might not be desirable. The z-score measure is computed as follows:

$$f'_i(t,m) = \frac{f_i(t,m) - \mu_m}{\sigma_m}, \forall t \in \{total - start, \ldots, total - end\}, \forall m \in month \quad (6.3)$$

Median

In scenarios where data is skewed, mean can be sensitive to outliers. In such settings, median is typically considered to be more robust to outliers. As a result, we consider

median as a method for base computation:

$$f'_i(t,m) = f_i(t,m) - median_m, \forall t \in \{total - start, \ldots, total - end\}, \forall m \in month \tag{6.4}$$

Jackknife estimate

The Quenouille Tukey jackknife approach [83] is a useful nonparametric estimate of mean and variance. The basic idea behind the jackknife estimator is to systematically compute the mean estimate by leaving out one observation at a time from the sample set. Let $f_1, f_2, \ldots, f_n$ be the n points in the time series of a location x. The jackknife mean estimate is computed at point f_i by taking the mean of all points except f_i as follows:

$$Mean(f_i) = \frac{f_1 + \ldots + f_{i-1} + f_{i+1} + \ldots + f_n}{n-1} \tag{6.5}$$

Thus the anomalies are constructed by excluding the value at each point f_x^i. We however still use all the monthly values only to compute the jackknife estimate at each point. The variance measure using the jackknife approach turns out to be:

$$Variance = \left(\frac{n}{n-1}\right)^2 \times Variance(f_1, \ldots, f_n) \tag{6.6}$$

In order to see this consider $f_1, \ldots, f_n$ to be variable during a given month. Then we

have to following:

$$\begin{aligned} Variance &= \frac{1}{n}\sum_i (f_i - Mean(f_i))^2 \\ &= \frac{1}{n}\sum_i (f_i - \frac{n}{n-1} \times Mean + \frac{1}{n-1} \times f_i)^2 \\ &= \frac{1}{n}\sum_i \frac{n^2}{(n-1)^2} \times (f_i - Mean)^2 \\ &= \left(\frac{n}{n-1}\right)^2 \times Variance(f_1, \ldots, f_n) \end{aligned}$$

The variance essentially turns out to be an unbiased estimate and is similar to the maximum aposterior probability (MAP) estimate of the model. MAP is similar maximum likelihood estimate (MLE) but also incorporates a prior distribution over the quantity one wants to estimate. MAP estimation can therefore be seen as a more robust form of MLE estimation. However the main problem with an approach based on jackknife is that it requires a lot of computation.

6.3.2 Different Time Periods for Anomaly Construction

As mentioned earlier, an anomaly series is constructed from the raw time series by removing a base value from it. The base value is generally considered to be the mean of the data. Since the true theoretical mean is not known, the base value is created by taking the sample mean of the data. However, most of the times a short reference interval is chosen to compute the base and changes with respect to that are studied. There is no absolute truth or guidelines available to choose the reference interval. Climate scientists generally choose the base as a moving 30 year period and study the changes with respect to that. However a moving short reference interval is problematic and can result in spurious results and conclusions. In order to highlight the problems associated with picking an arbitrary short base, we consider an example of teleconnections looking into the drought of the Sahel region in Africa in the Section 6.4.3.

6.4 Experiments and Results

We use the precipitation, air temperature and sea level pressure NCEP data (see Chapter 2) for our analysis as they represent the most important climate variables. In all, we have 62 years of data (corresponding to 744 monthly values) for 10512 grid locations on the globe.

6.4.1 Comparison of Different Measures of Anomaly Construction

Our first task is to empirically evaluate the differences in the four different measures described in Section 6.3.1. We use the precipitation data for our analysis. Using all the 62 years of data from the NCEP/NCAR website, we first construct an anomaly series for each location on the Earth using the four different measures. Further, we also construct complex networks by taking pairwise correlation between all locations on the Earth as used by several researchers like [61], [49], [20], [69] to find patterns in climate data. The nodes in the graph represent all the locations on the Earth and the edges represent pairwise correlation between the anomaly time series of all the nodes on the Earth. Our goal is to evaluate whether there are statistically significant differences between different measures to compute anomalies. In order to measure the statistically significant difference, we consider the following three criterion:

- *Mean based difference*: We compute the mean of anomaly time series using different measures and then compute the difference in mean for each pair of measure. The mean difference would be statistically significant if we can say with 95% confidence that the mean of difference is non-zero.

- *Correlation based difference*: Here we compute the correlation of every point with respect to other points on the globe using the four measures and check if the correlation values are impacted by using different measures for anomaly construction.

- *Monthly variance based difference*: Here we check if the monthly variance of the anomaly time series at each location is different for pairs of anomaly computation measure or not.

We use *t-test* to test if the difference between two measures follows a Gaussian distribution with $mean = 0$ and unknown variance. Thereby, our null hypothesis, $H0$ is that two measures lead to the same result and alternate hypothesis, Ha is that the two measures are different.

Tables 6.1, 6.2 and 6.3 show the number of locations where two measures lead to significant differences in the anomaly time series. Here we make an observation that z-score and median lead to significant differences from each other as well as the mean and the jackknife. Z-score based base computation yields the most significant difference as it leads to statistically significant changes in correlations and monthly variances at more than 9000 grid locations on the Earth. The z-score measure also stands out if we look at the monthly variance of each point. On the other hand, mean and jackknife seem to be similar. Median differs from the two over the mean difference based comparison. This is perhaps expected as the median and the mean values are not the same and all the other bases have zero monthly mean. Overall this result indicates that different measures used to compute base can lead to drastically different results. This result makes it intuitively clear that z-score might not be the best way to compute the base. In order to compare the mean and the median, we examine the skew in the anomaly time series after using the mean to construct the anomalies. To determine the skew, we check the *kurtosis* of the anomaly series at each location on the Earth. The kurtosis falls within the range of 2.6-3.5 for more than half of the locations on the Earth which is acceptable for a normal distribution. However, some locations have a very high skew and the kurtosis value is as high as 10. Fig 6.2 shows the histogram of kurtosis for all the locations and also shows the skew in the time series at a random location on the Earth. This suggests that the mean might not a good measure to compute the anomalies and median might

be a better choice. However, further investigations are still needed to understand the right measure for anomaly computation.

Method	Mean	Median	z-score	Jackknife
Mean	-	692	0	0
Median	-	-	1281	674
z-score	-	-	-	0
Jackknife	-	-	-	-

Table 6.1: Number of locations that rejected the null hypothesis at 95% confidence interval in the two sample t-test examining the anomalies at the different locations for precipitation.

Method	Mean	Median	z-score	Jackknife
Mean	-	0	5303	0
Median	-	-	5152	0
z-score	-	-	-	5303
Jackknife	-	-	-	-

Table 6.2: Number of locations that rejected the null hypothesis at 95% confidence interval in the two sample t-test examining the correlation of each location with the different locations for precipitation.

Method	Mean	Median	z-score	Jackknife
Mean	-	0	9152	0
Median	-	-	9152	0
z-score	-	-	-	8998
Jackknife	-	-	-	-

Table 6.3: Number of locations that rejected the null hypothesis at 95% confidence interval in the two sample t-test examining the monthly variance at the different locations for precipitation.

6.4.2 Comparison of different time periods for anomaly construction

The previous results show that for a fixed base period there exists different ways to compute the base which can lead to drastic differences in the anomaly time series. *Here we try to see if we can fix a measure (say mean) then check if varying the base period affects the anomaly time series.* In order to do this, we examine three base periods:

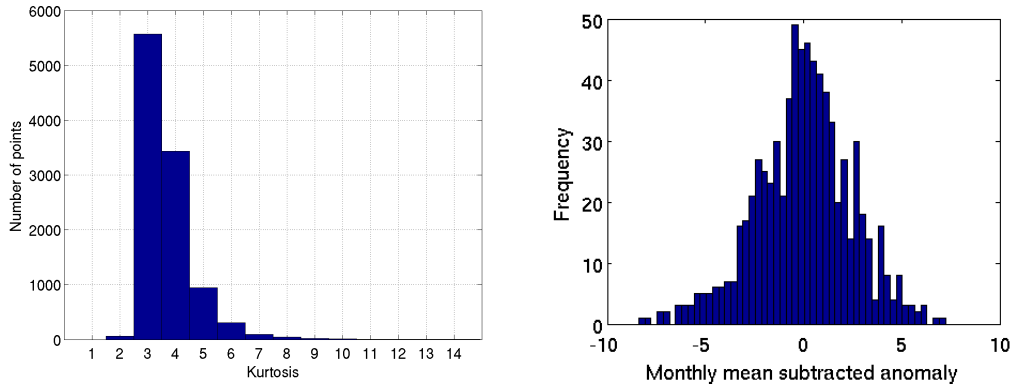


Figure 6.2: a) Kurtosis histogram b) The mean subtracted anomaly shows a skew in the data.

a) first 20 years b) entire 62 years and c) last 20 years. We also experiment with base period length of 30 years but that leads to similar results so for the sake of presenting the extremes, we present results choosing 20 years as a base.

We construct the anomaly series for each location corresponding to the given base periods. We selected mean as the measure of computing the base. In order to compare the time series, we used KL-divergence criteria to see if different base periods have different effects on anomaly time series. The KL-divergence is defined as follows:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{6.7}$$

KL-divergence of 0 means that the two series are exactly the same. A KL-divergence value indicates that the series are quite different. We plot the divergence value for each location on the globe in Figure 6.3. The white region shows that these locations are severely affected by our choice of base. In general last 20 years vs 62 years (second figure) has a light shade of gray indicating that all the locations on earth would be affected (in their anomaly series) if we make a choice between last 20 years vs all 62 years.

Also the variance in the anomaly time series changes when we move the 20 year base

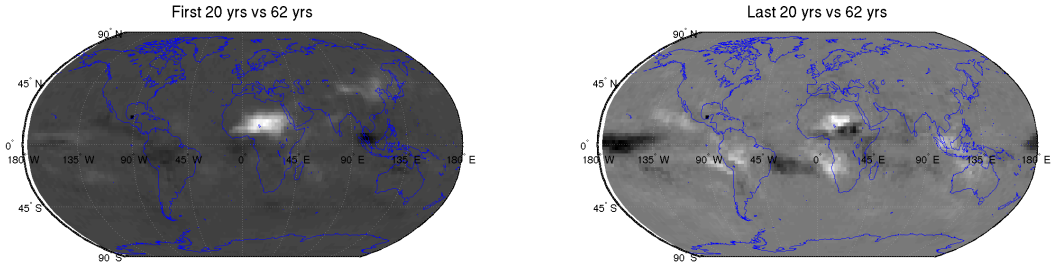


Figure 6.3: KL-divergence of the anomaly series the different bases a) first 20 years vs entire 62 years b) last 20 years vs entire 62 years and c) first 20 years vs last 20 years. The white shaded regions represent regions of maximum divergence.

period across the entire length of the time-series. Fig. 6.4 shows the change in variance at two random locations by picking up different 20 year base periods by varying the starting times from 1948-1988. From the figure, we see that average variance in the anomalies at different points varies using different start times for the base periods. This makes the problem complex as different regions show minimum variance in different windows of the time period.

In order to further analyze the impact of the choice of short reference base on the correlation of anomaly time-series, we consider two anomaly construction scenarios using the base as: a) first 20 years from 1948-1967 and b) using the entire 62 year time period from 1948-2009. We examine the changes in the mean correlations of locations with respect to every other location on the Earth using the two base period. Figure 6.5 shows the change in mean correlation of 100 random locations and the locations in Sahel using the two base periods to compute the anomalies. From the figures, we see that the locations in Sahel are much more impacted by the change in the base period as compared to the 100 random locations. We also find similar trends in other variables like pressure and temperature in Sahel but do not report it due to lack of space. These

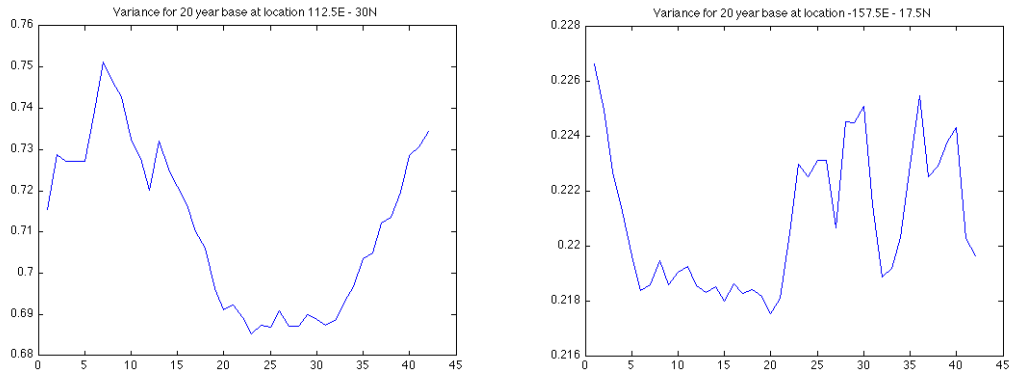


Figure 6.4: Change in variance of two random locations on the Earth choosing a 20 year reference period and moving the starting year from 1948-1988.

results underline the fact that a reference interval is crucial in the computation of the anomalies. In the next section, we show a case study on teleconnections where the actual analysis results and implications are impacted by the choice of the reference base.

6.4.3 Case study of the Sahel dipole

Teleconnections are long distance connections connecting the climate of two places on the Earth. One such class of teleconnections are the dipoles which consist of two regions having anomalies in the opposite direction and thus having negative correlation. The climate in Sahara and Sahel region of Africa has undergone some radical shift in the past century. The region received heavy rainfall till about 1969 until when it went into a period of severe drought for about 30 years which brought a *regime shift* in the region. The drought in the region and its environmental causes and consequences have been well studied in the past [29].

The precipitation in the region has recovered slightly but not enough to come back to the same levels as that before 1969. The severe loss of precipitation at Sahel was accompanied with a heavy increase in precipitation at the same time in the Gulf of

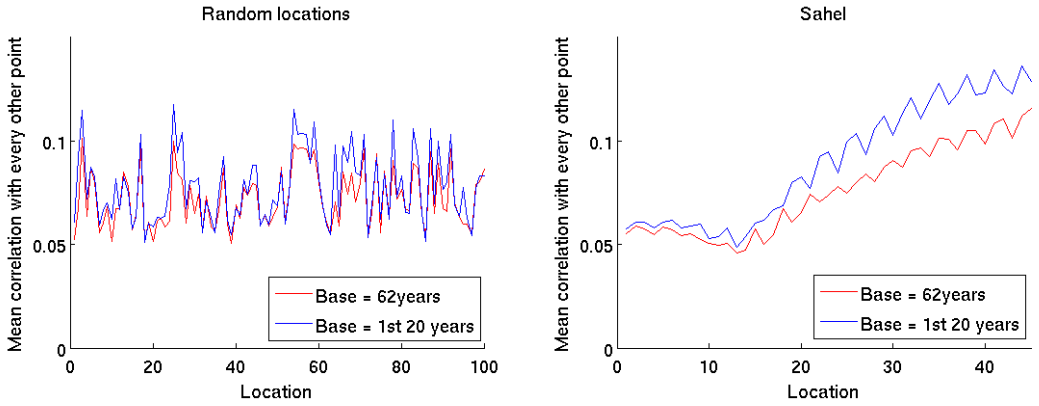


Figure 6.5: a) Mean correlation of 100 random points with all the points in the globe for precipitation using the entire 62 years as the base and only the first 20 years as the base. b) Mean correlation of 100 random points with all the points in the globe for precipitation using the entire 62 years as the base and only the first 20 years as the base. The difference in correlation (red and blue) is much more pronounced in Sahel as compared to the random locations.

Guinea around Africa, thus forming a dipole in precipitation [32]. The two regions Sahel and the Gulf of Guinea are marked in the Fig.6.6. The raw precipitation time series of the two locations in Sahel (7.5E, 20N) and the Gulf of Guinea (2.5E, 2.5S) are shown in the Fig. 6.7.

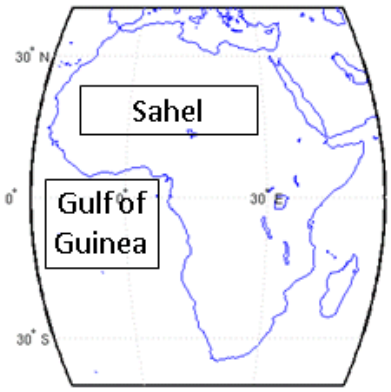


Figure 6.6: Sahel and the Gulf of Guinea in Africa.

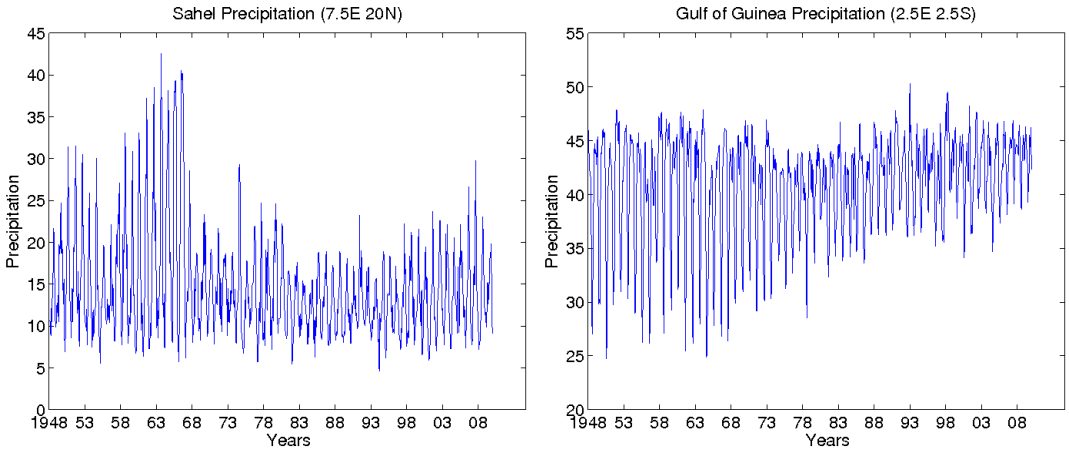


Figure 6.7: Raw precipitation time-series at Sahel and the Gulf of Guinea.

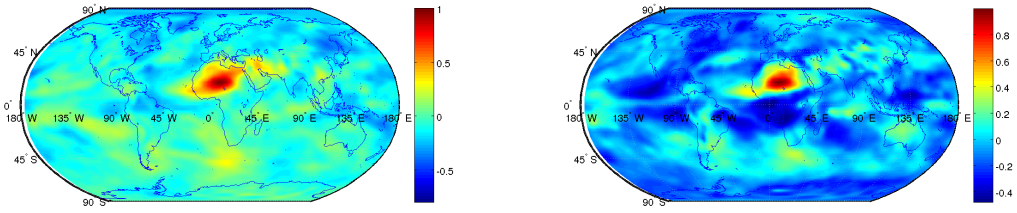


Figure 6.8: Correlation of precipitation time series for different places on the Earth with respect to a single point in Africa using the two base a)1948-1967 and b)1987-2008 for the 1st network (1948-1967).

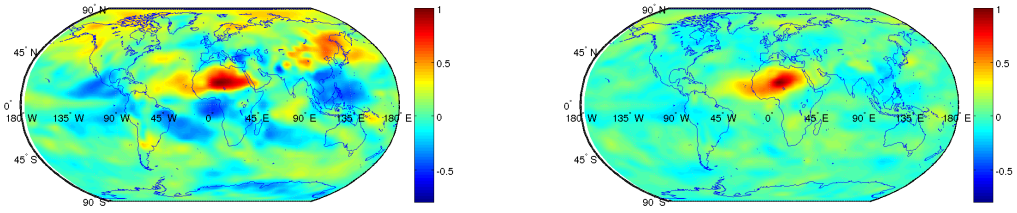


Figure 6.9: Correlation of precipitation time series for different places on the Earth with respect to a single point in Africa using the two base a)1948-1967 and b)1987-2008 for the 2nd network (1968-1987).

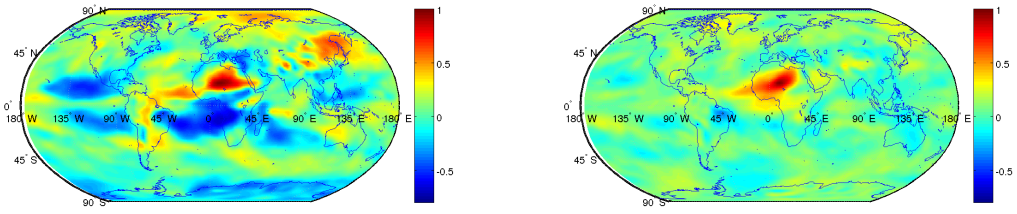


Figure 6.10: Correlation of precipitation time series for different places on the Earth with respect to a single point in Africa using the two base a)1948-1967 and b)1987-2008 for the 3rd network (1987-2007).

From the figure, we can see the dramatic decrease in precipitation in Sahel and an increase in precipitation in the Gulf of Guinea around the 1970s. Now using the 62 years of NCEP precipitation data, we choose two base years, the first 20 years (1948-1967) and the last 20 years(1988-2007). We further construct three networks by taking pairwise correlation between the anomaly time series of all locations on the Earth for a 20 year time period each (1948-1967, 1968-1987 and 1988-2007). Consider the point A(7.5E, 20N) in Sahel. Let us examine the correlations of this point with all the regions on the Earth. Fig. 6.8, 6.9 and 6.10 show the correlation of all the points on the Earth with respect to a single point in Sahel for the three time periods 1948-1967,1968-1987 and 1988-2007 respectively. From the figures, we see that the if we choose the base period to be the first 20 years, the Sahel dipole is clearly visible (positive and negative correlations as shown by red and blue regions in the figures) in the period 1988-2007, however if we choose the last 20 years as the base period the dipole is seen in the interval 1948-1967. Further we use the dipole detection algorithm on the complete network as given in [49]. The algorithm begins by picking up the most negative edge on the Earth and grows the two ends of the negative edge into two regions such that they are negatively correlated with each other and positively correlated within each other. Using the algorithm, we see that the Sahel dipole appears in different time periods and also in different regions as also shown in the Fig.6.11. Thus the choice of a base period severely impacts the results and subsequently the interpretations that can be drawn from the results. Hence extra caution needs to be exercised while constructing anomalies in order to avoid spurious conclusions to be drawn from the results.

6.5 A Generalized Approach for Anomaly Construction

In the previous section, we saw that there is a bias introduced in the results upon considering different measures of the base and different durations. So the primary

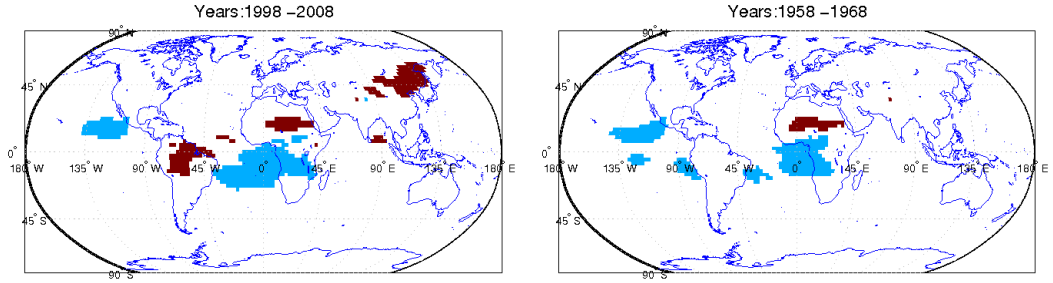


Figure 6.11: Different regions and different time periods are identified as dipoles in precipitation using the first 20 years as the base and the last 20 years as the base. (The red and blue regions represent two ends of the dipole and have negative correlation in their anomalies.)

question arises, *What is the right base to choose for anomaly construction?*. In this section, we discuss our approach to handle the problem of the anomaly construction. The intuition behind our approach is to have a weighted mean to construct the anomalies and use an objective criteria to pick up the right set of weights using Monte Carlo sampling. The weighted base for anomaly construction for a location i is created as follows:

$$b_i(t,w) = \sum_{t=t0}^{t0+k} w_t * f_i(t) \quad subject\ to \sum_{t=t0}^{t0+k} w_t = 1 \tag{6.8}$$

where $t0$ represents the starting time period, k represents the length of the time period. We further assume that the weights w_t are the same for each year and do not depend upon the month in the year. Further, the anomalies are constructed by removing the weighted base for each month as follows:

$$f_i'(t,w) = f_i(t) - b_i(t,w) \tag{6.9}$$

We run Monte Carlo simulations to get the right set of the weights w_t and define the objective function as minimizing the variance of the anomaly time series over time and space. By minimizing the variance, we are trying to enforce uniformity over the data. There can be some other objective functions like the median of the lowest 10% of the correlations. The intuition behind this objective function is that for computing dipoles,

Algorithm 3 A Generalized approach for Anomaly construction.

Let $f_i(t)$ be the monthly values of raw time series of location i
Let, N = Length of total time period.
Let, T_{base} = Shortest length of reference interval.
Let, $NumSimulations$ = Number of simulations to run.
Initialize $GlobalVariance$, $OptimalWeights$ to ∞
repeat
  for $k \in T_{base}, ...T_N$ **do**
    for $t \in T_0, ...T_N$ **do**
      for $i \in 1....NumSimulations$ **do**
        Compute weight vector $w_1, w_2, ..., w_t$
        subject to the constraint $w_1 + w_2 + ... + w_t = 1.$ using a Dirchlet prior.
        Compute the weighted base as $b_i(t, w) = \sum_{t=t0+1}^{t0+k} w_t * f_i(t)$
        Compute the Anomalies from the weighted base as $f_i'(t, w) = f_i(t) - b_i(t, w)$
        Compute the Variance of all the anomaly time-series across the globe.
        if $Variance < GlobalVaraince$ **then**
          Update the $GlobalVariance$ and $OptimalWeight$
          $GlobalMedian = Median$
          $OptimalWeight = w_1, w_2, ..., w_t$
        end if
      end for
    end for
  end for
until convergence

we need to examine the most negative correlations. Hence we want to find a weight and a base vector corresponding to our criterion for dipoles. However, we consider a general objective function that is not dependent upon the problem. The further details of the algorithm are present in Algorithm 3.

6.5.1 Results

We use the precipitation data and run our Monte Carlo based simulation algorithm to get the right reference base period. Figure 6.12 represents the final converged weights. The other parameters of the final convergence of the algorithm are as mentioned in Table 6.4. Using our new weighted anomaly, we re-construct the correlation plots around the Sahel to get a sense of the dipole in the Sahel. Figure 6.13 shows the new results using the

Table 6.4: Final algorithm convergence details.

Parameter	Value
Period	55
Starting year	1948

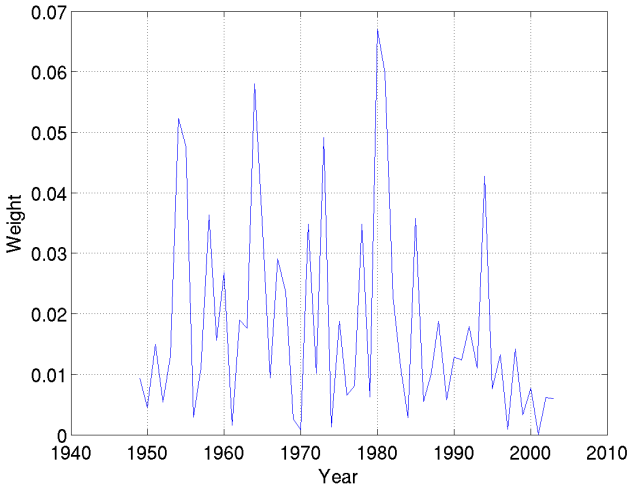


Figure 6.12: Final converged weight vector.

dipole in the Sahel using the weighted anomaly. Using a bias free base gives us confidence about the non-spuriousness of the discovered climate pattern or climactic phenomena such as a dipole. It implies that a dipole does exist in the region and that bases chosen which result in the dipole appear vanishing are not good bases. This objective function thus helps us in observing phenomena which would be more prominent if favorable bases are assumed but a bias-free base gives us a worst case scenario and more confidence in the results.

6.6 Discussion and Conclusions

The issue of anomaly construction is a fundamental problem in climate science as most of the analysis and results are derived after the raw data is transformed into an anomaly

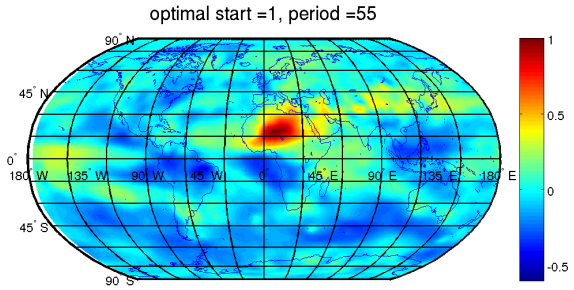


Figure 6.13: A correlation map as seen from the Sahel location A(7.5E, 20N).

series. However there are no current guidelines available on anomaly construction and climate scientists usually rely upon computing a moving reference base for anomaly construction. In this chapter, we examine the various issues pertaining to the construction of the anomalies. We assess the four methods of anomaly construction i.e. mean, median, z-score and jackknife. Our results show that if z-score is used as a measure for anomaly computation then the correlation values across different locations come out to be significantly different at 95% confidence interval. The mean, median and the jackknife measure do not show significant differences. However, due to the skewness in the data, the mean might not be a good measure and the median might be a good measure in such a case. However, further investigation is required to understand the right measure should be used.

We further show the bias in results introduced due to a choosing a short reference interval and show the difference in conclusions and results using a case study of the Sahel dipole. It is important to handle the bias introduced due to a short base as subsequent conclusions derived from it get affected. We further propose a generalized algorithm to handle the the issue of a bias-free base. Using our algorithm, we get the optimal base period to be 55 years. The algorithm can be modified to have different objective functions to handle different specific scenarios. The future work includes examining different approaches to learn the weight vector as opposed to using the Monte Carlo

simulations. We will also evaluate different objective measures and their impact on the base construction.

Chapter 7

Discussion and Future Work

In this thesis we presented novel algorithms and approaches to handle the spatio-temporal datasets in the climate domain. We summarize the main contributions, implications and limitations of our work as follows.

7.1 Contributions, Implications and Limitations

Discovering Dipoles in Climate Data

This thesis presented a novel graph based approach to find all the dipoles in a given dataset. One of the major benefits of our work is that our approach allows us to have a single snapshot picture of all the dipoles in a given spatio-temporal dataset. This enables us to comprehensively study the dipoles, understand the changes in their behavior and interactions (e.g. their movement) in a more precise manner. It also allows us to discover potentially new connections not known to climate scientists do far. We show that the dipoles that we discover have a high correlation with the static indices used by climate scientists and thus are a good representative of the existing phenomenon.

A larger significance of this work, which might impact how climate scientists perceive the climate indices, is that we show that some of the dynamic dipoles defined using our approach (e.g. NAO and SO) have a much correlation with global temperature anomalies as compared to the static indices. This result suggests that the climate indices are better explained as centroids of dynamic clusters as opposed to a pair of static locations (e.g. Tahiti and Darwin for SO) or an index formed with EOF over a limited region (e.g. NAO).

One of the main limitations of our analysis is that we rely upon the notion of reciprocity in a graph to prune noisy edges. While this assumption has been proven successful in finding all the known dipoles which are also present as a very strong phenomenon, for new weaker dipoles we might need to examine the condition of reciprocity at a more local level. Further, our analysis computes the negative correlation of the dipoles for the entire length of the time series. As a result, we do not capture dipoles which could be present at a small scale at an arbitrary point and not after that. Further, our analysis computes dipoles as spatially contiguous regions. This leads to missing the teleconnections that could be connecting more than two regions. For e.g., the PNA pattern consists of three poles and our dipole analysis is only able to capture two of the three regions. Another limitation of our analysis is that we have several thresholds and parameters in our algorithm. While we used empirical evaluation over a number to them to determine the best set of parameters for our analysis, a more systematic approach to determine the correct parameters for a general setting is a part of the future work.

Mining time lagged relationships in climate datasets

We presented a novel approach to find time lagged relationships in climate datasets. We presented a novel graph based approach based upon a modification to our previous approach to find teleconnections to capture both positively and negatively correlated

time lagged teleconnections at each time lag (in our case, we ran the experiments for a lag value of 1 to 10 pentads). In climate data, time-lagged relationships are crucial towards understanding the linkages and influence of the change in the climate at one region of the Earth on another region. These relationship are lagged in time because a climatic phenomenon that affects a specific location does not affect a different location at the same time but only at a later time. Using our approach, we are able to identify some of the known connections like the MJO and the PNA pattern. We also show that our framework can be used to obtain multivariate relationships across different variables and thus understand their inter-relationships.

One major significance of our framework is that it can be used to understand time-lagged relationships across important variables like flood occurrence, drought, etc. and thus can assist in weather prediction and monitoring. For example, Lu et al. [54] extend our framework to understand the dipole structure of global extreme flood events. They show the benefit of using the SRNN based framework as compared to a principal component analysis for aiding the prediction of precipitation. A limitation of our analysis is that we explicitly generate patterns for every lag value. Some of the cluster pairs could be present at 2 or more lag values. In order to judge which lag value is more appropriate for a relationship we need to have a significance based testing algorithm. Significance testing can also help in ruling out spuriously connected edges possible by random chance.

Testing the significance of teleconnection patterns

We presented a significance testing algorithm to test the significance of the generated teleconnection patterns that us based upon the wild bootstrap framework and overcomes the challenges arising in a spatio-temporal setting from non i.i.d. data, seasonality and trends. We showed the utility of our approach by showing the results on the NCEP dataset. We see that our approach declares the known dipoles as significant with a

pvalue close to zero. Also, dipoles declared as insignificant also do not make sense from the physical perspective.

An important implication of our approach is that it can help us remove a lot of spurious patterns generated by the previous approach and potentially help us identify new connections. Our framework can be instructive to other researchers in the spatio-temporal domain to test the significance of patterns. One limitation of our work is that we model only linear trends in the climate data. This might be a simplistic assumption given the heterogeneity of Earth science data.

Anomaly construction in climate datasets

Anomaly construction is an important first step in dealing with climate datasets and is used to take care of seasonality in the data. Typically, climate scientists use a moving reference base of 30 years and compute anomalies with respect to that interval. We examined how the choice of the base can greatly impact the outcome of an analysis. Our analysis showed that appearance and the disappearance of the dipole phenomenon in the Sahel region was linked with the particular choice of the base for the anomaly construction. Our algorithm based upon Monte-Carlo based simulations show that a we can reliably construct a weighted base that reduces the variance of the anomaly time series.

7.2 Future Work

Our work on dipole analysis can be employed to understand various GCM models that study study climate change. These GCMs exhibit variability in their predictions of various climate variables, as they use different representations of physical interactions in the climate system. Hence they often diverge in their predictions and sometimes

even offer contradicting projections of changes in various regions in response to different greenhouse gas emission scenarios. Our current approach provides a comprehensive view of the dipoles on Earth and, hence, a power to test various models in terms of their ability to capture the dipoles. One of the next steps of this thesis is to test the accuracy in the predictions of GCMs based on how well the capture the dipole phenomenon.

Another next step of the thesis is to study the interactions between the different dipoles. So far these dipoles are mainly studied in isolation, and the interactions between dipoles have been considered as weak. Understanding these dipoles and their interplay is crucial to understand the variability of the global climate system. Although the linkage of individual teleconnection patterns to weather changes in various parts of the globe is well known and documented, there are not many investigations examining the relationship between the various dipoles which is also critical as the synergy between different dipoles can have great influence on climate. E.g., while the cold winter over Europe in 2010 could be largely explained by NAO and other local indices [16], the cold winter over North America at the same time is largely due to a combination of NAO and ENSO [31].

Additionally, an important future work direction is to understand the composition of extreme events of temperature and precipitation at a location and attribute them to individual dipoles. Understanding this decomposition can be very helpful for predictions and potential societal payoff to mitigate and plan for the impacts of the interplay between various teleconnections.

References

[1] Climate prediction centre, http://www.cpc.ncep.noaa.gov/.

[2] Climatic research unit, http://www.cru.uea.ac.uk/.

[3] Earth system research laboratory, http://www.esrl.noaa.gov/psd/data/climateindices/.

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

[5] http://www.ecmwf.int/products/data/archive/descriptions/e4/index.html.

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

[7] http://www.jreap.org/indexe.html.

[8] Meet the MJO, Feature Article From Intermountain West Climate Summary, May 2008.

[9] Temperature anomaly time series, national climatic data center, noaa, http://www.ncdc.noaa.gov/ghcnm/time-series/index.php.

[10] Agrawal, R., Imieliński, T., and Swami, A. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (1993), vol. 22, ACM, pp. 207–216.

[11] Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB* (1994), vol. 1215, pp. 487–499.

[12] Benjamini, Y., and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* (2001), 1165–1188.

[13] Besag, J., and Diggle, P. Simple monte carlo tests for spatial pattern. *Applied Statistics* (1977), 327–333.

[14] Bridgman, H., and Oliver, J. *The global climate system: patterns, processes, and teleconnections*. Cambridge Univ Pr, 2006.

[15] Castro, N., and Azevedo, P. J. Time series motifs statistical significance. In *SDM* (2011), pp. 687–698.

[16] Cattiaux, J., Vautard, R., Cassou, C., Yiou, P., Masson-Delmotte, V., and Codron, F. Winter 2010 in Europe: A cold extreme in a warming climate. *Geophysical Research Letters 37*, 20 (2010), L20704.

[17] Compo, G., Kiladis, G., and Webster, P. The horizontal and vertical structure of east asian winter monsoon pressure surges. *Quarterly Journal of the Royal Meteorological Society 125*, 553 (1999), 29–54.

[18] Dommenget, D., and Latif, M. A cautionary note on the interpretation of eofs. *Journal of Climate 15*, 2 (2002), 216–225.

[19] Donges, J., Zou, Y., Marwan, N., and Kurths, J. Complex networks in climate dynamics. *The European Physical Journal-Special Topics 174*, 1 (2009), 157–179.

[20] Donges, J., Zou, Y., Marwan, N., and Kurths, J. Complex networks in climate dynamics. *The European Physical Journal-Special Topics 174*, 1 (2009), 157–179.

[21] Edgington, E. *Randomization tests*, vol. 147. CRC Press, 1995.

[22] Elsner, J., Jagger, T., and Fogarty, E. Visibility network of United States hurricanes. *Geophysical Research Letters 36*, 16 (2009), L16702.

[23] Engle, R., and Granger, C. Co-integration and error correction: representation, estimation, and testing. *Econometrica: Journal of the Econometric Society* (1987), 251–276.

[24] Ertoz, L., Steinbach, M., and Kumar, V. A new shared nearest neighbor clustering algorithm and its applications. In *Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining* (2002), pp. 105–115.

[25] Ertoz, L., Steinbach, M., and Kumar, V. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *SIAM international conference on data mining* (2003), vol. 47.

[26] et al., H. H. The enso effect on the tropical atlantic variability- a regionally coupled model study. *Geophysical research letters* (2002).

[27] Ferreira, P., Azevedo, P., Silva, C., and Brito, R. Mining approximate motifs in time series. In *Discovery Science* (2006), Springer, pp. 89–101.

[28] Fisher, R., et al. On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron 1* (1921), 3–32.

[29] Foley, J., Coe, M., Scheffer, M., and Wang, G. Regime shifts in the sahara and sahel: interactions between ecological and climatic systems in northern africa. *Ecosystems 6*, 6 (2003), 524–532.

[30] Gadgil, S., Vinayachandran, P., Francis, P., and Gadgil, S. Extremes of the indian summer monsoon rainfall, enso and equatorial indian ocean oscillation. *Geophysical Research Letters 31*, 12 (2004), L12213–1.

[31] García-Serrano, J., Rodríguez-Fonseca, B., Bladé, I., Zurita-Gotor, P., and de la Cámara, A. Rotational atmospheric circulation during north atlantic-european winter: the influence of enso. *Climate Dynamics* (2010), 1–17.

[32] Giannini, A., Saravanan, R., and Chang, P. Oceanic forcing of sahel rainfall on interannual to interdecadal time scales. *Science 302*, 5647 (2003), 1027.

[33] Gionis, A., Mannila, H., Mielikinen, T., and Tsaparas, P. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data (TKDD) 1*, 3 (2007), 14–es.

[34] Granger, C. Some properties of time series data and their use in econometric model specification. *Journal of econometrics 16*, 1 (1981), 121–130.

[35] Hanhijarvi, S., Garriga, G., and Puolamaki, K. Randomization techniques for graphs.

[36] Hardle, W., and Mammen, E. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics 21*, 4 (1993), 1926–1947.

[37] Hendon, H., and Salby, M. The life cycle of the Madden-Julian oscillation. *Journal of the Atmospheric Sciences 51*, 15 (1994), 2225–2225.

[38] Hines, K., Bromwich, D., and Marshall, G. Artificial surface pressure trends in the ncep–ncar reanalysis over the southern ocean and antarctica. *J. Climate 13*, 22 (2000), 3940–3952.

[39] Hinneburg, A., and Gabriel, H. Denclue 2.0: Fast clustering based on kernel density estimation. *Advances in Intelligent Data Analysis VII* (2007), 70–80.

[40] Janowiak, J. An investigation of interannual rainfall variability in africa. *Journal of Climate 1* (1988), 240–255.

[41] Jarvis, R., and Patrick, E. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers* (1973), 1025–1034.

[42] Jones, C., and Schemm, J. The influence of intraseasonal variations on medium- to extended-range weather forecasts over south america. *Monthly Weather Review 128*, 2 (2000), 486–494.

[43] Kalnay, E., and et al. The ncep/ncar 40-year reanalysis project. *Bull. Amer. Meteor. Soc. 77* (1996), 437–471.

[44] Kawale, J., Chatterjee, S., Kumar, A., Liess, S., Steinbach, M., and Kumar, V. Data guided discovery of dynamic dipoles. In *Conference on Intelligent Data Understanding CIDU* (2011).

[45] Kawale, J., Chatterjee, S., Ormsby, D., Steinhaeuser, K., Liess, S., and Kumar, V. Testing the significance of spatio-temporal teleconnection patterns. In *KDD* (2012), pp. 642–650.

[46] Kawale, J., Liess, S., Ganguly, A., Lall, U., and Kumar, V. Mining time-lagged relationships in spatio-temporal climate data. In *IEEE Conference on Intelligent Data Understanding CIDU* (2012).

[47] Kawale, J., Liess, S., Kumar, A., Steinbach, M., Ganguly, A., Samatova, N., Semazzi, F., Snyder, P., and Kumar, V. Data guided discovery of dynamic climate dipoles. In *CIDU* (2011), pp. 30–44.

[48] Kawale, J., Liess, S., Kumar, A., Steinbach, M., Ganguly, A., Samatova, N., Semazzi, F., Snyder, P., and Kumar, V. A graph based approach to find teleconnections in climate data. In *Statistical Analysis and Data Mining Journal* (2013).

[49] Kawale, J., Steinbach, M., and Kumar, V. Discovering dynamic dipoles in climate data. In *SIAM International Conference on Data mining, SDM* (2011),

SIAM.

[50] Kawale, J., Steinbach, M., and Kumar, V. Discovering dynamic dipoles in climate data. In *SIAM Conference on Data Mining, SDM* (2011), pp. 107–118.

[51] Kim, K.-Y., and Chung, C. On the evolution of the annual cycle in the tropical pacific. *Journal of Climate 14*, 5 (2001), 991–994.

[52] Kumar, A., Jha, B., Zhang, Q., and Bounoua, L. A new methodology for estimating the unpredictable component of seasonal atmospheric variability. *Journal of Climate 20*, 15 (2007), 3888–3901.

[53] Mammen, E. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics 21*, 1 (1993), 255–285.

[54] Mengqian, L., Lall, U., Kawale, J., Liess, S., and Kumar, V. Extreme global floods and their correlation networks with climate precursors.

[55] Oates, T. Peruse: An unsupervised algorithm for finding recurrig patterns in time series.

[56] on Climate Change, I. P. *Fourth Assessment Report: Climate Change 2007: The AR4 Synthesis Report*. Geneva: IPCC, 2007.

[57] Onogi, K., Tsutsui, J., Koide, H., et al. The jra-25 reanalysis. *J. Meteor. Soc. Japan 85* (2007), 369–432.

[58] Paeth, H., and Hense, A. On the linear response of tropical African climate to SST changes deduced from regional climate model simulations. *Theoretical and applied climatology 83*, 1 (2006), 1–19.

[59] Pekeris, C. L. Atmospheric oscillations. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 158*, 895 (1937).

[60] Peterson, R., and White, W. Slow oceanic teleconnections linking the Antarctic circumpolar wave with the tropical El Nino-Southern Oscillation. *Journal of Geophysical Research 103*, C11 (1998), 24573–24.

[61] Steinbach, M., Tan, P., Kumar, V., Klooster, S., and Potter, C. Discovery of climate indices using clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (2003), ACM, pp. 446–455.

[62] Steinhaeuser, K., Chawla, N., and Ganguly, A. An exploration of climate data using complex networks. *SIGKDD Explorations 12*, 1 (2010), 25–32.

[63] Steinhaeuser, K., Chawla, N., and Ganguly, A. Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. *Statistical Analysis and Data Mining 4*, 5 (2011), 497–511.

[64] Tan, P., Steinbach, M., Kumar, V., et al. *Introduction to data mining*. Pearson Addison Wesley Boston, 2006.

[65] Thomson, D. Dependence of global temperatures on atmospheric co2 and solar irradiance. *Proceedings of the National Academy of Sciences of the United States of America 94*, 16 (1997), 8370.

[66] Tingley, M. A bayesian anova scheme for calculating climate anomalies, submitted 2011.

[67] Trenberth, K. Some effects of finite sample size and persistence on meteorological statistics. part i: Autocorrelations. *Mon. Wea. Rev 112* (1984), 2359–2368.

[68] Tsonis, A., and Roebber, P. The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications 333* (2004), 497–504.

[69] Tsonis, A., Swanson, K., and Roebber, P. What do networks have to do with climate? *Bulletin of the American Meteorological Society 87*, 5 (2006), 585–595.

[70] Tsonis, A., Swanson, K., and Wang, G. On the role of atmospheric teleconnections in climate. *Journal of Climate 21*, 12 (2008), 2990–3001.

[71] Ulrich, W., and Gotelli, N. Null model analysis of species nestedness patterns. *Ecology 88*, 7 (2007), 1824–1831.

[72] Uppala et al, S. M. The era-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society 131*, 612 (2005), 2961–3012.

[73] Vecchi, G., and Wittenberg, A. El niño and our future climate: where do we stand? *Wiley Interdisciplinary Reviews: Climate Change 1*, 2 (2010), 260–270.

[74] Veech, J. A null model for detecting nonrandom patterns of species richness along spatial gradients. *Ecology 81*, 4 (2000), 1143–1149.

[75] Von Storch, H., and Zwiers, F. *Statistical analysis in climate research*. Cambridge Univ Pr, 2002.

[76] Waldron, A. Null models of geographic range size evolution reaffirm its heritability. *American Naturalist* (2007), 221–231.

[77] Walker, G. Correlation in seasonal variations of weather, viii. a preliminary study of world weather. *Memoirs of the India Meteorological Department 24*, 4 (1923), 75–131.

[78] Wallace, J., and Gutzler, D. Téléconnexions in the geopotential height field during the northern hemisphere winter. *Mon. Wea. Rev 109* (1981), 784–812.

[79] Weatherhead, E., Reinsel, G., Tiao, G., Meng, X., Choi, D., Cheang, W., Keller, T., DeLuisi, J., Wuebbles, D., Kerr, J., et al. Factors affecting the detection of trends: Statistical considerations and applications to environmental data. *Journal of Geophysical Research 103*, D14 (1998), 17–149.

[80] Wei, L., Kumar, N., Lolla, V., Keogh, E., Lonardi, S., and Ratanamahatana, C. Assumption-free anomaly detection in time series. In *Proceedings of the 17th international conference on Scientific and statistical database management* (2005), Lawrence Berkeley Laboratory, pp. 237–240.

[81] Wilks, D. *Statistical methods in the atmospheric sciences*, vol. 100. Academic press, 2011.

[82] Wu, C. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics 14*, 4 (1986), 1261–1295.

[83] Wu, C. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics 14*, 4 (1986), 1261–1295.

[84] Zhang, C. Madden-Julian Oscillation. *Rev. Geophys 43*, 2 (2005), 1–36.