Testing the Significance of Spatio-temporal Teleconnection Patterns

Jaya Kawale*, Snigdhansu Chatterjee‡, Dominick Ormsby*, Karsten Steinhaeuser*
Stefan Liess†, Vipin Kumar*

*Department of Computer Science ‡ School of Statistics † Department of Soil, Water & Climate
University of Minnesota
{kawale,ormsby,ksteinha,kumar}@cs.umn.edu chatterjee@stat.umn.edu liess@umn.edu

ABSTRACT

Dipoles represent long distance connections between the pressure anomalies of two distant regions that are negatively correlated with each other. Such dipoles have proven important for understanding and explaining the variability in climate in many regions of the world, e.g., the El Niño climate phenomenon is known to be responsible for precipitation and temperature anomalies over large parts of the world. 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. In this paper, we present a novel method for testing the statistical significance of the class of spatio-temporal teleconnection patterns called as 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.

Categories and Subject Descriptors

H.2.8 [**Databse Management**]: Database Applications—*Data mining*

Keywords

Significance testing

1. INTRODUCTION

Pressure dipoles are important long distance climate phenomena (*teleconnection*) characterized by anomalies[1] of opposite polarity appearing at two different locations at the same time. Dipoles are of great importance in understanding climate variability and are known to impact precipitation and temperature anomalies throughout the globe. Fig. 1 shows the pressure anomaly time series at Tahiti and Darwin representing one of the most well-known dipoles - the El Niño Southern Oscillation which is known to drive precipitation and temperature anomalies worldwide. The anomaly time series of the two regions are in the opposite direction representing an oscillation.

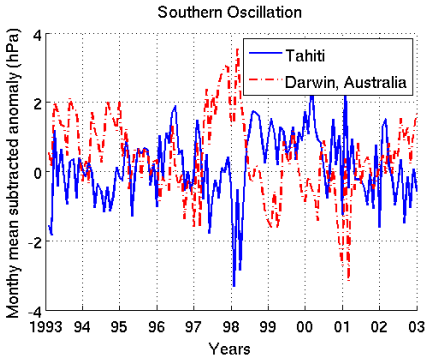


Figure 1: Pressure anomaly time series at the two ends of the Southern Oscillation.

Historically, these dipoles have been discovered by direct observation of some climate phenomenon on land and have been defined using single point locations [1]. Later on, pattern analysis techniques such as the EOF [2] 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 [3, 4, 5]. Kawale et al. [3, 4] present a graph based approach to find dipoles in the climate data and are able to match the existing dipole indices used by climate scien-

[1]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.

KDD'12, August 12–16, 2012, Beijing, China.

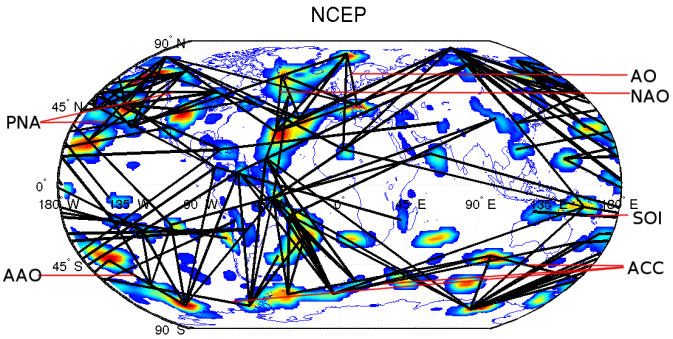


Figure 2: Dipole edges with correlation < -0.2 in the NCEP sea level pressure data taken from [3].

tists 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 [3, 4]. 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 [3]. 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. 2 shows the dipoles generated by the algorithm given in [3]. 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.

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 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 [6] have been successfully used in many contexts in data mining to find interesting patterns in graphs [7], association rule mining [8], motif mining [9, 10, 11], etc. In ecology, significance testing has been used to study the analysis of species nestedness patterns [12] and to study the diffusion of a spatial phenomenon [13] and spatial gradients [14]. Monte Carlo tests to test the significance of spatial patterns has been discussed in [15]. However, there are many challenges in using randomization tests for spatio-temporal patterns, some of which are listed below:

1.1.1 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.

1.1.2 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 [16].

1.1.3 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 [17, 18]. 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 [17, 18] 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 [19].

1.1.4 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.

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 paper, 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 [20, 21] 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.

2. PROPOSED 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 [20, 21]. 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:

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.

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 [19].

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

$$f(t) = \alpha + \beta t + \gamma \sin\left(\frac{2\pi(t+\delta)}{12}\right) \quad (1)$$

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:

$$\forall_{i \in A}\ X_i = U + r_i \tag{2}$$

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.

$$SE_r = Tr\left[(X - U\mathbf{1})^T\ W\ (X - U\mathbf{1})\right] \tag{3}$$

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 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:

$$U_t = f(t) + \phi[U_{t-1} - f(t-1)] + \epsilon_t \tag{4}$$

Similar to equation 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

$$SE_\epsilon = \sum_t (V_t - \phi V_{t-1})^2 \tag{5}$$

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 3 and 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,

$$\sum_t g_k(t) \cdot f(t) = \frac{\sum_{i,t} g_k(t) \cdot w(i) \cdot X_{it}}{Tr(W)}, k = 1, 2, 3 \tag{6}$$

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,

$$\phi = \frac{\sum_{t=2}^{T} V_t \cdot V_{t-1}}{\sum_{t=2}^{T} V_{t-1}^2} \tag{7}$$

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 $E[\epsilon_t]$.

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 [22] 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:

$$Z_{ij} = \frac{1}{2} log \frac{1 + \rho_{ij}}{1 - \rho_{ij}} \tag{8}$$

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

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 [23] 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:

$$p_{AB} = \frac{1}{N}\sum_{i=1}^{N} I_{(Z_{AB} \geq Z'_{AB})} \tag{9}$$

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 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. The proof of the above theorem is omitted here due to the lack of space.

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 [24] 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:

$$P_{(k)} \leq \frac{k}{m \cdot c(m)} * \alpha \tag{10}$$

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 $P_{(k)}$.

3. EXPERIMENTS AND RESULTS

3.1 Dataset

We use the data from the NCEP/NCAR Reanalysis project provided by the NOAA/ESRL [25]. The NCEP reanalysis product uses an assimilation scheme embedded in a physical model to interpolate global observations from 1948 onward into a gridded projection of the state of the atmosphere. The reanalysis datasets are created by assimilating remote and in situ sensor measurements using a numerical climate model to achieve physical consistency and interpolation to global coverage; they are considered the best available proxy for global observations. 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.

3.2 Results

We ran the dipole algorithm using the NCEP dataset and the algorithm mentioned in [3] 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. 3 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. 4(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. 4(b), we see that there is a linear trend in the opposite direction which the model is able to effectively capture. Fig. 5 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.

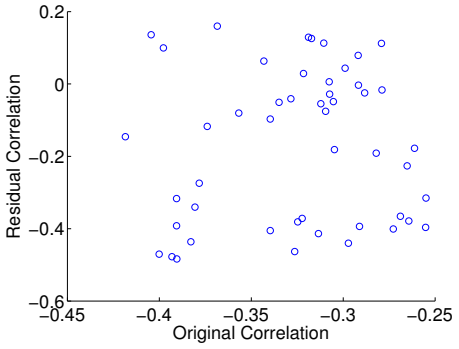
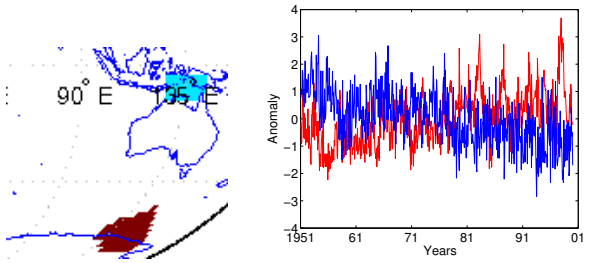


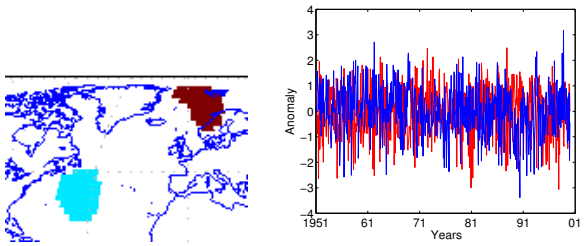
Figure 3: Scatter plot showing original vs residual correlation.



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

(b) Time series at the two centres

Figure 4: Dipole rejected due to linear trend.



(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: Dipoles discarded due to seasonality filtering.

On the other hand, Fig. 6 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 [26]. 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.

3.3 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. 7(a) shows an example of a dipole that had an original correlation of −0.39 but has high non-linear trends. Fig. 7(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 [27] which is also reflected in the time series as shown in the Fig. 7(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,

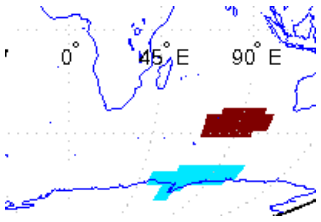
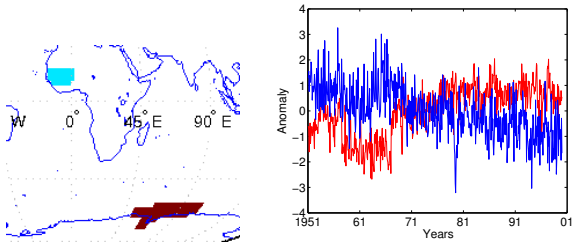


Figure 6: Dipole having an original correlation −0.25 but a residual correlation −0.39 corresponds to the known dipole AAO.



(a) Dipole having a correlation of −0.39

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

Figure 7: Dipole showing non-linear trend corresponding to abrupt change due to the Sahel drought

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 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.

$$Discard\ dipoles\ if\ |\beta_A - \beta_B| \geq \hat{\beta} \tag{11}$$

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. 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.

3.4 Comprehensive Evaluation

Table 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 8 shows

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

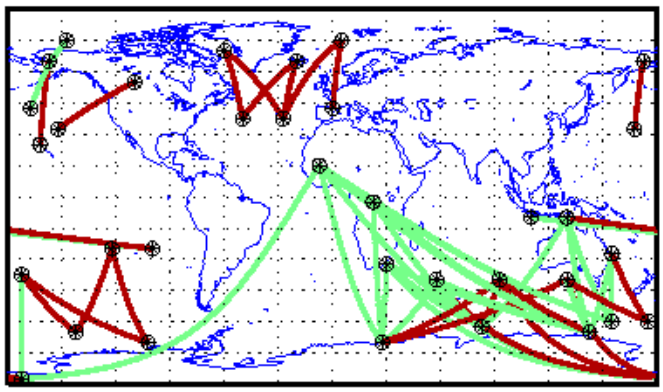


Figure 8: Dipoles declared significant in the NCEP dataset at a threshold of -0.25. Red denotes significant dipoles and green denotes insignificant dipoles.

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 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 9 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. 10 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 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. 11 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

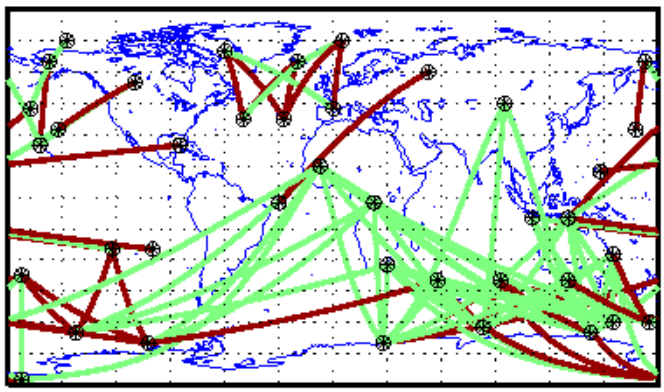


Figure 9: Dipoles declared significant in the NCEP dataset at a threshold of -0.2. Red denotes significant dipoles and green denotes insignificant dipoles.

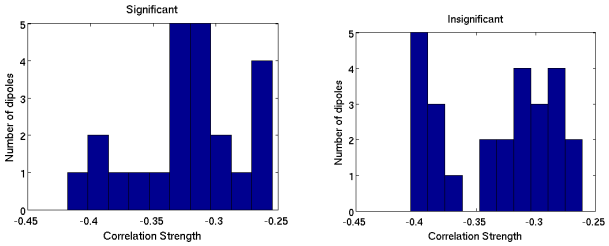


Figure 10: Histogram of correlation strengths for significant and insignificant dipoles.

NCEP data from the climate science [28] about the presence of significant spurious trends in the southern hemisphere.

Table 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 them (see Table 1). Seasonality in the dipoles could be one possible reason. Fig. 12 shows the gamma values of the dipoles. From the figure, we see that quite a few of them have significant value of gamma.

3.4.1 *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 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 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.

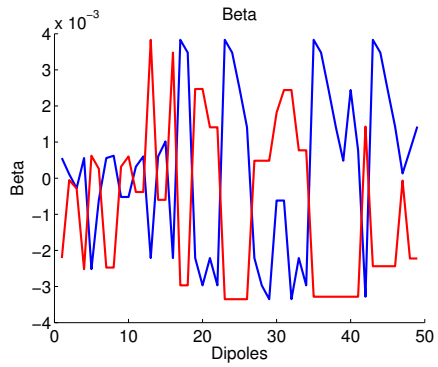


Figure 11: Beta values at the two ends of the dipoles for the NCEP dataset.

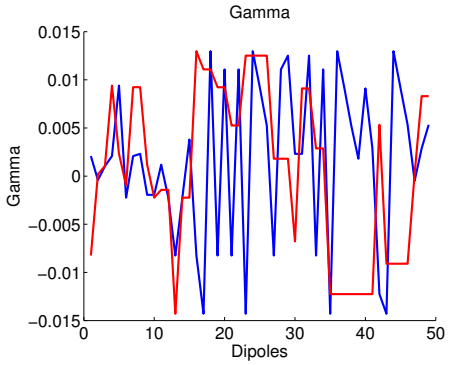


Figure 12: Gamma values at the two ends of the dipoles for the NCEP dataset.

3.4.2 *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 [26]. 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. 13 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.

3.4.3 *A new dipole ?*

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. 8 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,

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

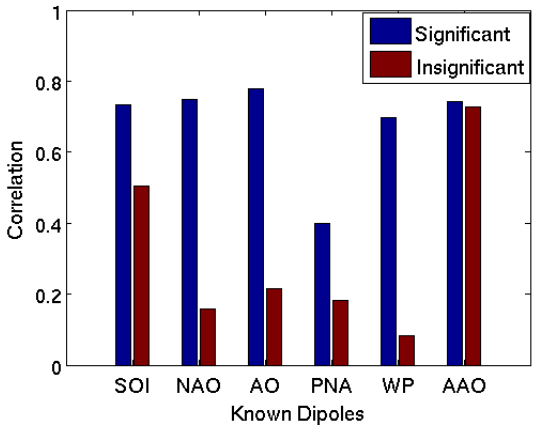


Figure 13: Maximum correlation with known indices in the two sets of dipoles.

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. 14. It appears as significant in the NCEP data and its correlation with the known indices is also very low (see Fig. 15). 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.

4. DISCUSSION AND CONCLUSION

Significance testing in spatio-temporal data presents many

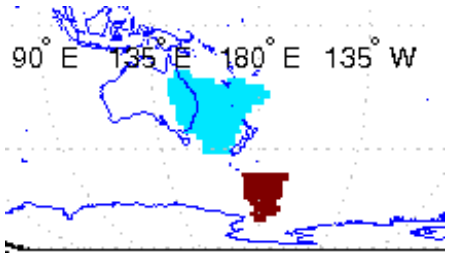


Figure 14: Dipole near Australia shows up as statistically significant.

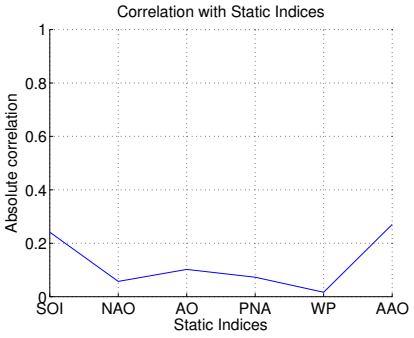


Figure 15: Correlation of the dipole near Australia with known indices

challenges due to the inherent autocorrelation dependencies in time and space. However, significance testing of spatio-temporal patterns has received little attention. In this paper, 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.

Acknowledgments

This work was partially supported by NSF grants IIS-0905581, IIS-1029711 and SES-0851705. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute.

5. REFERENCES

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

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

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

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

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

[6] E.S. Edgington. *Randomization tests*, volume 147. CRC Press, 1995.

[7] S. Hanhijarvi, G. Garriga, and K. Puolamaki. Randomization techniques for graphs. 2009.

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

[9] Nuno Castro and Paulo J. Azevedo. Time series motifs statistical significance. In *SDM*, pages 687–698, 2011.

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

[11] T. Oates. Peruse: An unsupervised algorithm for finding recurrig patterns in time series. 2002.

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

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

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

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

[16] D.S. Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

[17] R.F. Engle and C.W.J. Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, pages 251–276, 1987.

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

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

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

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

[22] R.A. Fisher et al. On the” probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

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

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

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

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

[27] J.E. Janowiak. An investigation of interannual rainfall variability in africa. *Journal of Climate*, 1:240–255, 1988.

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