Empirical Comparison of Correlation Measures and Pruning Levels in Complex Networks Representing the Global Climate System

Alex Pelan, Karsten Steinhaeuser, Nitesh V. Chawla
Deptartment of Computer Science and Engineering
Interdisc. Center for Network Science and Applications
University of Notre Dame
Notre Dame, IN, 46656
Email: {apelan,ksteinha,nchawla}@nd.edu

Dilkushi A. de Alwis Pitts
Center for Research Computing
University of Notre Dame
Notre Dame, IN, 46656
Email: dpitts@nd.edu

Auroop R. Ganguly
GIST Group, CSED
Oak Ridge National Laboratory
Oak Ridge, TN 37831
Email: gangulyar@ornl.gov

***Abstract*—Climate change is an issue of growing economic, social, and political concern. Continued rise in the average temperatures of the Earth could lead to drastic climate change or an increased frequency of extreme events, which would negatively affect agriculture, population, and global health. One way of studying the dynamics of the Earth's changing climate is by attempting to identify regions that exhibit similar climatic behavior in terms of long-term variability. Climate networks have emerged as a strong analytics framework for both descriptive analysis and predictive modeling of the emergent phenomena. Previously, the networks were constructed using only one measure of similarity, namely the (linear) Pearson cross correlation, and were then clustered using a community detection algorithm. However, nonlinear dependencies are known to exist in climate, which begs the question whether more complex correlation measures are able to capture any such relationships. In this paper, we present a systematic study of different univariate measures of similarity and compare how each affects both the network structure as well as the predictive power of the clusters.**

I. Introduction

Identifying and analyzing patterns in global climate is an important task, lending scientists a deeper understanding of the complex interactions between many variables that lead to observed climate phenomena. Complex networks have already been established as an effective means of representation of the climate [1]–[4], both for descriptive and predictive tasks [5]. These networks are constructed from gridded climate data, wherein each vertex represents a grid point (physical location in space) and weighted edges represent the climatic similarity between them (correlation in climate variability). By pruning and clustering these networks, climate scientists are able to uncover structure in the climate system and determine how different regions relate to each other.

A. Related Work

A number of publications related to this work exist in the literature, both in terms of clustering climate data in general as well as the use of complex networks to represent the climate system. For example, there are several studies on clustering climate data including applications of standard k-means clustering [6], [7] and a weighted k-means kernel with spatial constraints [8], to identify climate zones; a shared nearest neighbor method [9] to discover climate indices [10]; and a correlation-based approach of to identify multivariate clusters from data. The concept of climate networks was introduced in [3] and we were the first to apply community detection for the purpose of identifying climate regions [2], [5]. One other work considered the possibility of nonlinear correlation measures [1] in a limited context and concluded that the difference between linear and nonlinear measures was not significant. Still, a comprehensive and systematic evaluation different measures of similarity and how they affect the network structure is lacking from the literature.

B. Contributions

In this paper, we consider six different measures of similarity from four different genres: distance-based, linear, rank-based, and nonlinear measures. We create networks for three different edge densities to enable a fair comparison between the resulting networks. We calculate network statistics, perform prediction experiments (using the methodology described in [5]), and we compare the results to evaluate effects that arise from choosing different measures of similarity.

The remainder of this paper is organized as follows. In Section II, we introduce the dataset used for these experiments. In Section III, we discuss the different measures of similarity included in this study. Sections IV and V present the experimental setup and results, respectively. Section VI discusses the software we developed to easily create networks in this framework, which is available available for download with sample datasets on our website[1]. We conclude by pointing to some open challenges for future resarch in multivariate network construction and place our work in a broader context within and beyond climate science.

[1]http://www.nd.edu/ dial/software/climateNet.zip

II. Data

The Earth science data for our analysis stems from the NCEP/NCAR Reanalysis project [11], which is publicly accessible for download at [12]. This dataset is constructed by fusing and assimilating measurements from heterogeneous remote and in-situ sensors. Variable selection is an important issue in this context, one we have not yet fully explored. Previous research has relied on domain expertise for an appropriate selection [13]–[15]; however, we did not want to limit ourselves with an *a priori* selection and therefore include a wide range of variables in our study.

Specifically, for the purpose of these experiments we selected seven variables with the guidance of a domain expert; temperature (SST), sea level pressure (SLP), horizontal(HWS) and vertical (VWS) wind speed, precipitable water (PW), relative humidity (RH), and geopotential height (GH). These variables were chosen both for variety and because they are significant variables when it comes to defining climate regimes. The dataset is available on an angular evenly spaced 5° x 5° latitude-longitude grid at monthly intervals for a period of 60 years (1948-2007). Data is provided as a separate 720-element time series for each grid point.

Climate data presents some unique challenges because seasonality creates natural recurrence patterns resulting in temporal autocorrelation (Figure 1(a)). We combat this by preprocessing the raw data into anomaly series, which captures only the long-term variability, or deviation from usual behavior. Following the precedent set by related work [3], [9], we de-seasonalize the data by monthly z-score transformation and de-trending. At each grid point, we calculate for each month $m = \{1, ..., 12\}$ (i.e., separately for all Januaries, Februaries, etc.) the mean

$$\mu_m = \frac{1}{Y} \sum_{y=1948}^{2007} a_{m,y} \tag{1}$$

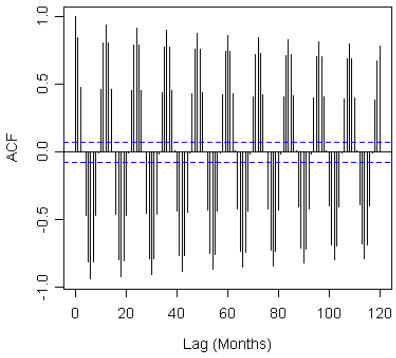
and standard deviation

$$\sigma_m = \sqrt{\frac{1}{Y-1} \sum_{y=1948}^{2007} (a_{m,y} - \mu_m)^2} \tag{2}$$

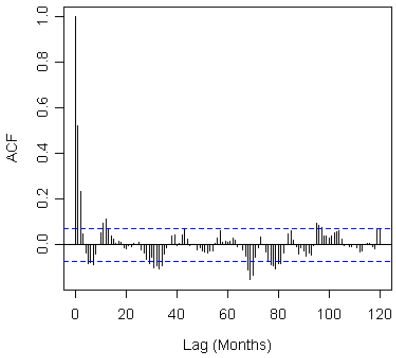
where y is the year, Y the total number of years in the dataset, and $a_{m,y}$ the value of series A at $month = m$, $year = y$. Each data point is then transformed (a^*) by subtracting the mean and dividing by the standard deviation of the corresponding month,

$$a^*_{m,y} = \frac{a_{m,y} - \mu_m}{\sigma_m} \tag{3}$$

The result of this process is illustrated in Fig. 1(b), which shows that de-seasonalized values have significantly lower autocorrelation than the raw data. In addition, we de-trend the data by fitting a linear regression model and retaining only the residuals. All data used for experiments or discussed hereafter using this method.



(a) Raw Data



(b) De-Seasonalized

Fig. 1. The de-seasonalized data (bottom) exhibits significantly lower autocorrelation due to seasonality than the raw data (top).

III. Measures of Similarity

We used six different measures of similarity in our experiments. These popular measures were selected because they offer a wide spectrum of ways to think about similarity between two time series, from linear interactions to monotonic curve fitting to general information finding methods.

A. Euclidean Distance

The Euclidean distance between two points is the normal distance, as would be measured by a ruler. In our networks we compute the distance between two time series by summing over the whole series the pair-wise distances between points taken at the same time step. We do not take the square root in this variation, and instead calculate it as

$$\sum_{i=1}^{t} (a_i - b_i)^2 \tag{4}$$

where i is a single time step and t is the length of the series.

B. Manhattan Distance

Manhattan distance is the sum of the lengths of the projections of the line segment between two points on the coordinate axes. It is also referred to as taxicab distance because it represents the minimum number of city blocks a taxi would have to travel between two points. Like the Euclidean distance,

we calculate the sum of Manhattan distance between time series as

$$\sum_{i=1}^{t} |a_i - b_i| \tag{5}$$

C. Pearson Correlation

The Pearson correlation is a measure of the linear correlation between two variables. It is one of the most commonly used measures in statistics and serves as our baseline, because Pearson correlation and variants on it have been used in most prior works [3], [5], [16]. It is computed as

$$\frac{\sum_{i=1}^{t}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{t}(a_i - \bar{a})\sum_{i=1}^{t}(b_i - \bar{b})}} \tag{6}$$

where $\bar{x}$ denotes the mean over the entire series x. Pearson correlation gives a value between -1 and +1, but since inverse relationships are of equal relevance in this particular context we use the absolute value.

D. Mutual Information

Mutual Information measures the dependence between two variables, or how how much knowing the quantity of one variable tells you about the quantity of another variable. It is calculated from the joint probability density functions of both variables as well as the marginal probability density functions of each individual variable,

$$\int\int p(a,b) log\left(\frac{p(a,b)}{p(a)p(b)}\right) dxdy \tag{7}$$

This measure is of potential significance for our experiments because it captures linear as well as nonlinear interactions. Mutual information networks were calculated using a Matlab package; for implementation details see [17].

E. Spearman's Rho

Spearman's Rho is a non-parametric, rank-based measure of dependence between two variables. It captures both linear and nonlinear interactions by seeing how well the relationship between two variables can be fit using a monotonic function, which is either strictly increasing or strictly decreasing. Like Pearson correlation, Spearman's Rho ranges from -1 and +1, and once again we take the absolute value. It is called a rank-based measure because each variable is ranked individually and then the differences between ranks for a given data point are calculated. A perfect correlation in Spearman's Rho is when the highest value of variable x is found at the same point as the highest or lowest value of variable y, the second highest x at the second highest/lowest y, and so on. To simplify our calculations, we break ties arbitrarily. We then use the following equation to calculate Spearman's Rho over an entire time series as

$$1 - \frac{\sum_{i=1}^{t}(a^i - b^i)^2}{n(n^2-1)} \tag{8}$$

where x^i is the rank of variable x.

F. Kendall's Tau

Like Spearman's Rho, Kendall's Tau is also calculated by comparing the relative ranks of the variables in two time series. Instead of checking correlation between ranks, however, it checks whether pairs are concordant or discordant. For two pairs p and q and sets of ranks x and y, if x^p is greater than x^q and y^p is greater than y^q, or x^p is less than x^q and y^p is less than y^q, then the pair is concordant (in agreement). Otherwise, there is disagreement and they are considered a discordant pair. We count the total number of concordant and discordant pairs for all possible pairs of points in the series, then calculate Kendall's Tau as

$$\frac{concordantpairs - discordantpairs}{1/2(n-1)} \tag{9}$$

Like Pearson correlation and Spearman's Rho, Kendall's Tau also ranges from -1 to +1 so we take absolute value as we are interested in both negative and positive correlation.

IV. Experimental Setup

A. Network Construction

We built networks for each of our seven variables combined with each of our six edge weighting methods. For the present purpose, we are only interested in grid points over the oceans; this decision is justified in [5]. Therefore, we first we apply a land-sea mask to the data to select the relevant data points. Then, we construct a fully connected network, meaning that we compute the edge weight between each pair of vertices in the network. Finally, we prune the networks to remove edges with low weight, which represent relatively weaker connections. Other authors have elected to use a fixed pruning threshold [3]; in prior work we used the statistical confidence in the correlation as a more rigorous method for pruning [5]. In this paper, we prune all networks to a fixed number of edges to ensure the fairest possible comparison between them, using 50,000, 100,000, and 250,000 edges (these values were determined from the significance-based pruning in [5]. Specifically, our pruning algorithm sorts the edges and then retains only the top m edges by weight, plus any ties. Thus, our experiments encompass 126 different networks (7 variabes $\times$ 6 measures $\times$ 3 pruning levels). The steps of this network construction process are visually summarized in Figure 2.

B. Prediction Experiments

The first step after we prune the networks is to detect clusters. We do this using the *Walktrap* community detection algorithm with default parameters, selected because it is computationally efficient and able to incorporate edge weights [18]. The algorithm also determines the number of clusters based on an external optimization criterion.

We are interested in testing the predictive power of the clusters [5]. We selected nine regions from a variety of different climate aspects, and the model predicts two variables for each of these (our dependent variables): air temperature and precipitation. We average the values over all grid points

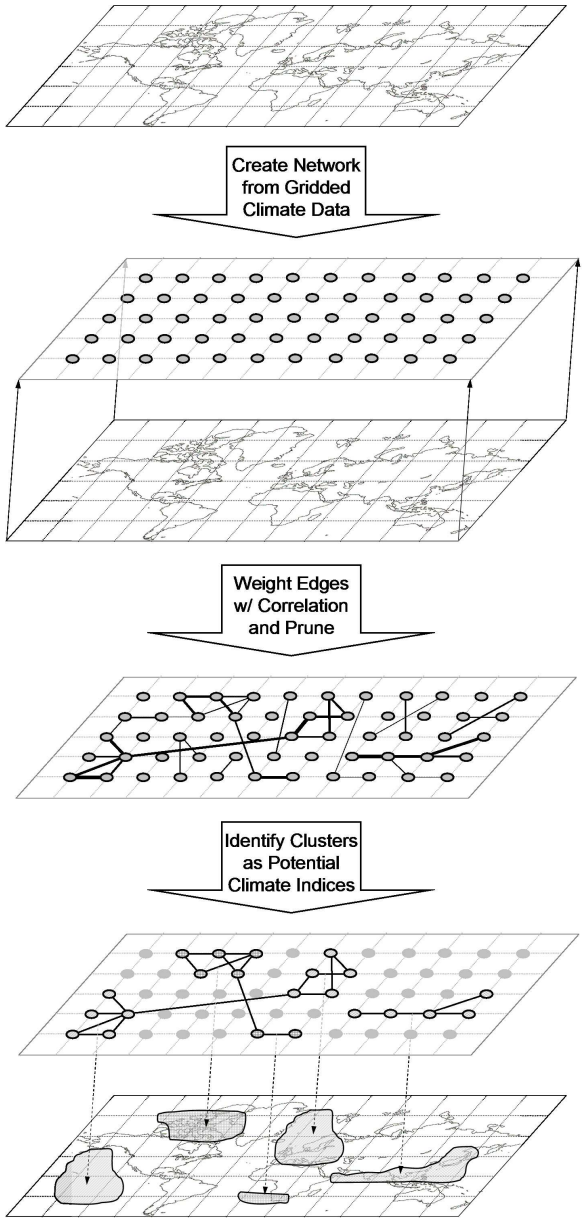


Fig. 2. Step-by-step overview of the network construction process.

in each of these regions into one time series, and these then become our dependent variables (Y).

We create our independent variables (X) from the clusters in a similar fashion. We average the anomaly values from every vertex in the cluster into one single time series for every cluster. Each cluster becomes a variable, and then we build a linear regression, attempting to predict the precipitation and temperature from the time series for the clusters. We use the first 50 years as a training set and the last 10 years as our testing set. This is representative of the predictive tasks we would want clusters to be able to do – relate the climate variability over oceans to land climate, like climate indices [10]. We use root mean square error (RMSE) to evaluate the regression experiments.

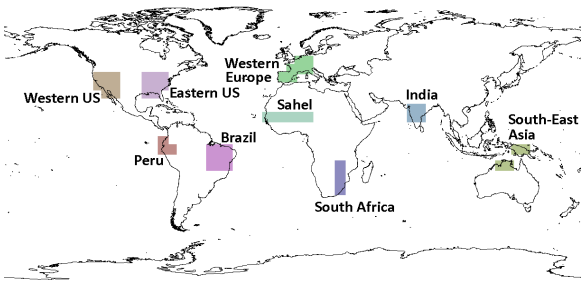


Fig. 3. Target Regions.

TABLE I
NUMBER OF VERTICES - GH NETWORK

Similarity Measure	50k Edges	100k Edges	250k Edges
Euclidean Distance	1,699	1,701	1,701
Manhattan Distance	1,699	1,701	1,701
Mutual Information	1,701	1,701	1,701
Pearson Correlation	1,701	1,701	1,701
Spearman's Rho	1,701	1,701	1,701
Kendall's Tau	1,701	1,701	1,701

After we run these experiments, we create new predictors from the best of the results. For each measure of similarity and climate variable, we select the level of pruning that performed the best. Then, for each measure of similarity, we create a new predictors wherein each variable's clusters from their best level of pruning is a variable. We run these new predictors on the same 18 variables from the first experiments. These create multivariate predictors from the univariate networks that were built.

V. RESULTS

A. Network Statistics

The network statistics we were interested in included number of vertices, number of edges, clustering coefficient, characteristic path length, and diameter. Table 1 shows the number of vertices in a typical network. There are 1701 sea data points in our data set, so every vertex is represented in almost every network. Networks that do not have 1701 vertices have vertices that had 0 edges going to or from them post pruning.

Table 2 shows the number of edges in a typical network. Although we pruned the networks to 50,000, 100,000, and 250,000 edges, there are not exactly those number of edges in the networks. This is because we included ties in the pruned networks.

TABLE II
NUMBER OF EDGES -PW NETWORK

Similarity Measure	50k Edges	100k Edges	250k Edges
Euclidean Distance	50,000	100,001	250,024
Manhattan Distance	50,002	100,002	250,003
Mutual Information	50,001	100,000	250,029
Pearson Correlation	50,001	100,001	250,000
Spearman's Rho	50,000	100,001	250,005
Kendall's Tau	50,001	100,007	250,019

TABLE III
CLUSTERING COEFFICIENT - GH NETWORK

Similarity Measure	50k Edges	100k Edges	250k Edges
Euclidean Distance	0.547	0.672	0.729
Manhattan Distance	0.549	0.672	0.727
Mutual Information	0.613	0.679	0.655
Pearson Correlation	0.605	0.680	0.673
Spearman's Rho	0.603	0.679	0.671
Kendall's Tau	0.604	0.679	0.671

TABLE IV
CLUSTERING COEFFICIENT - PW NETWORK

Similarity Measure	50k Edges	100k Edges	250k Edges
Euclidean Distance	0.615	0.593	0.558
Manhattan Distance	0.611	0.580	0.546
Mutual Information	0.561	0.386	0.316
Pearson Correlation	0.585	0.484	0.410
Spearman's Rho	0.579	0.477	0.403
Kendall's Tau	0.579	0.478	0.403

Tables 3 and 4 show two different trends noticed in the clustering coefficients of our networks. The clustering coefficient is a measure of the degree to which vertices tend to cluster together. It measures how close the neighbors of a vertex are to forming a fully-connected clique. For our networks, we took the average clustering coefficient of all vertices. In table 3, the clustering coefficient rises as the number of edges rises. This makes sense, as more edges should lead to greater clustering. In table 4, however, the clustering coefficient goes down as the number of edges rises. In previous work, edges were not pruned by a constant number of edges but rather by a significance test based method. In the network represented by table 3, this method lead to 249,322 significant edges. In the network represented by table 4, this lead to 50,835. The highest clustering coefficients in table 3 were found for 250,000 edges while the highest clustering coefficients in table 4 were found for 50,000 edges. We think this is because additional non significant edges are likely to just cluster randomly, which leads to a decrease in local clustering. It is also possible the different variables interact with local and far away vertices differently, meaning that there are possibly more local relationships for geopotential height than precipitable water.

Characteristic path length is the average distance, in number of edges, between two randomly selected vertices in the effort. Table 5 shows a typical network's characteristic path values. As would be expected, the characteristic path length goes down as more edges are added. One variable, shown in table 6, differed from the others in characteristic path length. While the trend of decreasing path length with increasing edges remained true, the paths were much longer. This is shown in table 6. The other variables had values much closer to table 5.

Diameter is the number of vertices in the shortest path between the two vertices that are furthest apart in a network. As would be expected, this decreases given an increase in the number of edges. This happens for every variable, and the GH network has higher than expected diameter, given its higher than expected characteristic path length.

TABLE V
CHARACTERISTIC PATH - OMEGA NETWORK

Similarity Measure	50k Edges	100k Edges	250k Edges
Euclidean Distance	2.320	1.971	1.827
Manhattan Distance	2.338	1.984	1.827
Mutual Information	2.549	2.129	1.829
Pearson Correlation	2.527	2.125	1.830
Spearman's Rho	2.508	2.113	1.829
Kendall's Tau	2.508	2.112	1.829

TABLE VI
CHARACTERISTIC PATH - GH NETWORK

Similarity Measure	50k Edges	100k Edges	250k Edges
Euclidean Distance	16.425	7.978	2.614
Manhattan Distance	17.054	7.98	2.601
Mutual Information	10.062	5.772	2.182
Pearson Correlation	10.508	6.414	2.430
Spearman's Rho	10.406	6.284	2.418
Kendall's Tau	10.397	6.235	2.416

B. Prediction Results

Table VII has the results of the first step of our predictive experiments. It displays the most effective pruning level for each variable and measure of similarity. The results are interesting, as it appears that regardless of how many significant edges networks had in our previous work, the best performing clusters are generally those clustered at 50,000 edges. We think this is because predictors that have more edges generally have less clusters in our experiments. This is because, at least compared to the significance-based pruning, we are forcing there to be extra edges in the networks that have more edges. This causes there to be more random edges, resulting in less networks that contain less predictive power. In addition to lower quality clusters, the larger edge networks also suffer from more noise in their linear regressions - there are less independent variables to build the model on.

We took these results and created new predictors, one for each method of edge weighting, and ran the new models on the same 18 variables as before. The results are in Table VIII. These results suggest that the difference between different methods of edge weighting are minimal. The biggest indicator of the region of the root mean square error is the prediction experiment, rather than the method of edge weighting. In addition to the experimental results, we also include the average RMSE for all of the precipitation and all of the temperature experiments. Pearson correlation performed the best in the precipitation experiments and Spearman's Rho performed

TABLE VII
MOST EFFECTIVE PRUNING

Variable	Euclid.	Man.	Mut. Info	Rho	Tau	Pearson
GH	50k	50k	50k	50k	50k	50k
VWS	100k	100k	250k	250k	250k	250k
PW	50k	50k	50k	50k	50k	50k
RH	50k	50k	100k	250k	50k	50k
SKT	50k	50k	50k	50k	50k	50k
SLP	50k	50k	100k	50k	50k	50k
HWS	50k	50k	50k	50k	50k	50k

TABLE VIII
PREDICTIVE EXPERIMENT RESULTS

Variable	Euclid.	Manhat.	Mutual Info	Rho	Tau	Pearson
air-ausindpap	0.65	0.59	0.59	0.59	0.58	0.57
air-brazil	0.51	0.46	0.51	0.49	0.50	0.49
air-india	0.73	0.65	0.59	0.59	0.61	0.64
air-peru	0.61	0.59	0.54	0.48	0.54	0.54
air-sahel	0.70	0.66	0.72	0.64	0.69	0.70
air-southafr	0.82	0.84	0.76	0.72	0.78	0.71
air-useast	0.66	0.88	0.81	0.72	0.72	0.68
air-uswest	0.64	0.64	0.72	0.59	0.61	0.63
air-weurope	0.62	0.63	0.52	0.49	0.54	0.57
prate-ausindpap	0.64	0.62	0.67	0.67	0.64	0.66
prate-brazil	0.44	0.44	0.48	0.49	0.48	0.46
prate-india	0.71	0.63	0.68	0.69	0.69	0.65
prate-peru	0.90	0.91	0.93	0.85	0.87	0.84
prate-sahel	0.53	0.52	0.52	0.58	0.59	0.54
prate-southafr	0.72	0.68	0.72	0.71	0.69	0.71
prate-useast	0.72	0.64	0.73	0.63	0.61	0.69
prate-uswest	0.62	0.62	0.62	0.56	0.57	0.53
prate-weurope	0.47	0.48	0.38	0.42	0.41	0.41
avg prate	0.64	0.62	0.64	0.62	0.62	0.61
avg air	0.66	0.66	0.64	0.59	0.62	0.61

the best in the air temperature experiments. The increase in accuracy from using these versus other measures is minimal, as the second place measure in all experiments but especially in the averages was usually less than a tenth worse. There are no real clear trends in which measure prevails - looking purely at win counts, Euclidean won once, Manhattan won six times, Mutual Information won three times, Spearman's Rho won four times, Kendall's Tau won one time, and Pearson correlation won four times. Every measure was the best in at least one experiment, and there was no dominant measure across the board at a statistical significance. In the averages, the differences were even smaller, as bad performances in one category were cancelled out by good performances in another.

VI. SOFTWARE

We have developed software that allows users to create their own networks using our method of construction. We are releasing it open source on our website along with this paper. The software is a command line utility that can be configured to run on local systems or SGE scheduler front ends. It is possible to make any of the networks that our experiments have produced or to take other datasets from the NCEP/NCAR Reanalysis project and create other networks. The software currently allows users to create univariate or multivariate networks from formatted data files and to prune network files that have already been created. Univariate networks can be made using Euclidean distance, Manhattan distance, Pearson correlation, Spearman's Rho, Kendall's Tau, and, if the user has Matlab installed, Mutual Information. Multivariate networks can be constructed with a number of different experimental methods, described in Section 7. The software is in constant development and we are trying to eventually include the whole framework, from preprocessing data to clustering to predictive experiments, in the software.

VII. MULTIVARIATE EDGE WEIGHTING EXTENSION

As alluded to in our software section, we experimented with some multivariate edge weighting techniques in addition to univariate ones. Our multivariate methods mainly come from two separate categories: pure multivariate measures that took all variables into account and produced an edge weight between locations, and multivariate methods that took completed univariate methods and combined their weights into an edge weight between locations.

Pure multivariate methods that we experimented with include Euclidean distance, Manhattan distance, and Cross Correlation function. Euclidean and Manhattan distances are simply the multivariate corollaries of the univariate methods we used in our univariate experiments. The Cross Correlation function is a way of extending the Pearson correlation to multiple dimensions. It computes univariate Pearson correlation between each possible set of variables, then finds the Euclidean distance between the points where each possible set of variables is one dimension.

Multivariate methods, where we combined univariate networks, included simple addition and Borda voting. Simple addition simply combines the edge weights from each provided network into one result network by adding them. Borda voting is a more interesting method, as it really penalizes being well correlated in one variable but not correlated at all in another. Borda voting is similar to the method used in college football polls, in which each voter ranks their top 25. The first ranked team gets 25 points, the second ranked 24, and so on all the way to the 25th place team, which gets one point. In our implementation, each network is a voter, and the edge's rank in that network garners them (size of network - rank) points in the Borda method.

It is unfortunately very difficult to interpret our results from these experiments, particularly in the domain. What does it mean if two edges have a strong correlation in a multivariate network? It could mean that they are highly related in temper-

ature, but not at all in pressure, or vice versa. It could mean that they are slightly related in all variables. While this work is beyond the scope of this paper, these questions motivate future work, especially since we have features in software to build multivariate networks.

VIII. Conclusion

We have previously shown that complex networks are an effective descriptive and predictive framework for climate variables. This paper sought to analyze the impact of different edge-weighting mechanisms. The question that we studied in this paper was: can the predictability of the network clusters improve if the network edges are weighted differently? While we found that network statistics and structure were noticeably affected by different methods of edge weighting, we did not find a significant change in the predictive power of the network clusters. This is a compelling observation, as it indicates that the network structure carries the power, and less comes from different weighting mechanisms. However, these observations are largely on univariate networks. There is still a lot of room for future work in this field, most notably in the creation and interpretation of multivariate networks from disparate climate variables.

Acknowledgment

This research was supported in part by the National Science Foundation under Grants OCI-1029584 and BCS-082695, the Center for Research Computing at the University of Notre Dame, and as part of a project titled "Uncertainty Assessment and Reduction for Climate Extremes and Climate Change Impacts", funded in FY2010 by the "Understanding Climate Change Impacts: Energy, Carbon, and Water Initiative", within the LDRD Program of the Oak Ridge National Laboratory, managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract DEAC05-00OR22725. The United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for Government purposes.

References

[1] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, "Complex networks in climate dynamics," *Eur. Phys. J. Special Topics*, vol. 174, pp. 157–179, 2009.

[2] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "An Exploration of Climate Data Using Complex Networks," *ACM SIGKDD Explorations*, vol. 12, no. 1, pp. 25–32, 2010.

[3] A. A. Tsonis and P. J. Roebber, "The architecture of the climate network," *Physica A*, vol. 333, pp. 497–504, 2004.

[4] A. A. Tsonis, K. L. Swanson, and P. J. Roebber, "What Do Networks Have to Do with Climate?" *BAMS*, vol. 87, no. 5, pp. 585–595, 2006.

[5] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "Complex Networks as a Unified Framework for Descriptive Analysis and Predictive Modeling in Climate Science," *Statistical Analysis and Data Mining*, (in press, available at doi:10.1002/sam.10100).

[6] R. G. Fovell and M.-Y. C. Fovell, "Climate Zones of the Conterminous United States Defined Using Cluster Analysis," *J. Climate*, vol. 6, no. 11, pp. 2103–2135, 1993.

[7] W. W. Hargrove and F. M. Hoffman, "Using Multivariate Clustering to Characterize Ecoregion Borders," *Comput. Sci. Eng.*, vol. 1, no. 4, pp. 18–25, 1999.

[8] M. Sap and A. Awan, "Finding spatio-temporal patterns in climate data using clustering," in *Int'l Conf. on Cyberworlds*, 2005, pp. 164–171.

[9] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter, "Discovery of Climate Indices using Clustering," in *ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, 2003, pp. 446–455.

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

[11] E. Kalnay *et al.*, "The NCEP/NCAR 40-Year Reanalysis Project," *BAMS*, vol. 77, no. 3, pp. 437–470, 1996.

[12] http://www.cdc.noaa.gov/data/gridded/ data.ncep.reanalysis.html.

[13] A. A. Tsonis and K. L. Swanson, "Topology and Predictability of El Niño and La Niña Networks," *Phys. Rev. Lett.*, vol. 100, no. 228502, 2008.

[14] A. A. Tsonis, *Nonlinear Dynamics in Geosciences*. New York: Springer, 2007, ch. 1, pp. 1–15.

[15] K. Yamasaki, A. Gozolchiani, and S. Havlin, "Climate Networks around the Globe are Significantly Affected by El Niño," *Phys. Rev. Lett.*, vol. 100, no. 22, pp. 157–179, 2008.

[16] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, "The backbone of the climate network," *Europhys. Lett.*, vol. 87, no. 4, p. 48007, 2009.

[17] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE T PATTERN ANAL*, vol. 27, no. 8, pp. 1226–1238, 2005.

[18] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *J. Graph Alg. App.*, vol. 10, no. 2, pp. 191–218, 2006.