COMPLEX NETWORKS IN CLIMATE SCIENCE: PROGRESS, OPPORTUNITIES AND CHALLENGES

KARSTEN STEINHAEUSER[1,2], NITESH V. CHAWLA[1], AND AUROOP R. GANGULY[2]

Abstract. Networks have been used to describe and model a wide range of complex systems, both natural as well as man-made. One particularly interesting application in the earth sciences is the use of complex networks to represent and study the global climate system. In this paper, we motivate this general approach, explain the basic methodology, report on the state of the art (including our contributions), and outline open questions and opportunities for future research.

1. Introduction

Datasets and systems that can be represented as interaction networks (or graphs), broadly defined as any collection of interrelated objects or entities, have received considerable attention both from a theoretical viewpoint [1, 2, 6, 8, 13, 31] as well as various application domains; examples include the analysis of social networks [30], chemical interactions between proteins [26], the behavior of financial markets [12], and many others. Recently, the study of *complex networks* – that is, networks which exhibit non-trivial topological properties – has permeated numerous fields and disciplines spanning the physical, social, and computational sciences. So why do networks enjoy such broad appeal? Briefly, it is their ability to serve at once as a data representation, as an analysis framework, and as a visualization tool. The analytic capabilities in particular are quite powerful, as networks can uncover structure and patterns at multiple scales, ranging from local properties to global phenomena, and thus help better understand the characteristics of complex systems.

We focus on one particular application of networks in the earth sciences, namely, the construction and analysis of *climate networks* [25]. Identifying and analyzing patterns in global climate is an important task of growing scientific, social, and political interest, with the goal of deepening our understanding of the complex processes underlying observed phenomena. To this end, we make the case that complex networks offer a compelling perspective for capturing the dynamics of the climate system. Moreover, the computational sciences – specifically data mining and machine learning – are able to contribute a valuable set of methods and tools ranging from pattern recognition to predictive models. Thus, in this paper we expand upon the general approach to climate networks (e.g., see [21]) and motivate a promising area of interdisciplinary research. Indeed, we believe that this marriage of analytic methods, computational tools and domain science has the long-term potential for a transformative impact on our understanding of the earth's climate system.

The remainder of the paper is organized as follows: Section 2 describes the data and basic methodology for constructing climate networks; Section 3 briefly discusses related work involving other uses of complex networks in climate; Section 4 presents an overview of the types of structural analysis performed on climate networks, including important observations; Section 5 motivates the use of clustering on climate networks; Section 6 discusses extensions to multivariate relationships and incorporating temporal dynamics; Section 7 examines information content and predictive modeling in the context of climate networks; Section 8 addresses computational issues; finally, Section 9 outlines some of the major challenges and opportunities to advance the state of the art.

[1] Department of Computer Science & Engineering, Interdisciplinary Center for Network Science & Applications, University of Notre Dame, Notre Dame, IN 46556; ksteinha@nd.edu, nchawla@nd.edu.
[2] Geographic Information Science & Technology Group, Computational Sciences & Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; gangulyar@ornl.gov.

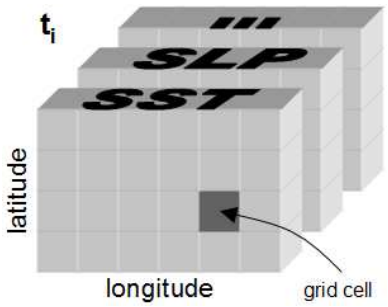


FIGURE 1. Schematic depiction of gridded climate data for multiple variables at a single timestep t_i in the rectangular plane.

2. Background and Basic Methodology

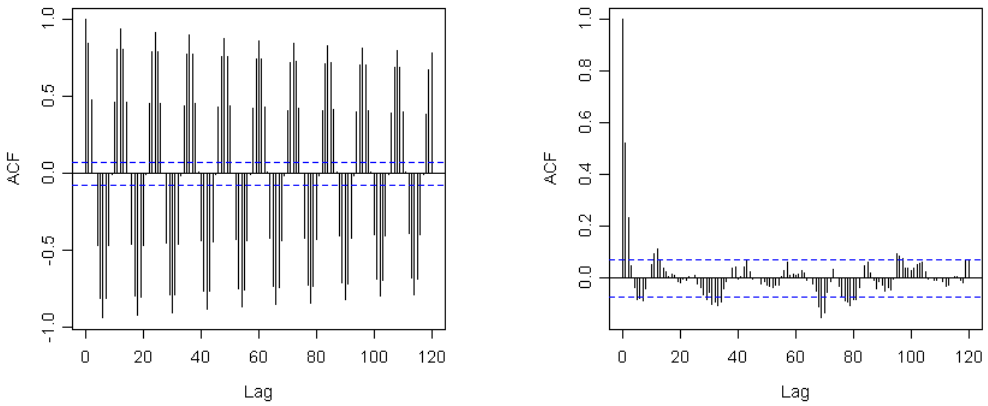
A network is any set of entities (nodes) with connections (edges) between them. The nodes can represent physical objects, locations, or even abstract concepts. Similarly, the edges can have many interpretations ranging from physical contact to mathematical relationships and conceptual affiliations. Thus, networks may take many different forms, shapes and sizes.

The concept of climate networks was first proposed by Tsonis and Roebber [21] and placed into the broader context of complex network literature in [25]. The intuition behind this methodology is that the global climate system can be represented by a set of oscillators (climate variability at different locations around the globe) interacting in some complex way. More precisely, the oscillators correspond to anomaly time series of gridded climate data (see Section 2.1) and the interactions are measured as the pairwise correlations between them [21, 25]. In the following sections, we describe the characteristics of the data and the network construction process in more detail.

2.1. **Gridded Climate Data.** The most commonly used data in climate network studies to date [3, 4, 18, 19, 20, 21, 23, 24, 25, 32, 33] stems from the NCEP/NCAR Reanalysis Project [9] (available for download at [27]). This dataset is created by assimilating remote and in-situ sensor measurements covering the entire globe and is widely recognized as one of the best surrogates for global observations as it is obviously impossible to obtain exact measurements. The data includes a wide range of surface and atmospheric variables, although prior lines of work have focused primarily on temperature [3, 24, 32] and pressure-related indicators [21, 25].

We did not want to constrain ourselves by an arbitrary *a priori* selection of variables, so in our recent work [18] we compare a wider range of climate descriptors. Specifically, we include these seven variables (abbreviation, brief definition in parentheses): *sea surface temperature* (SST, water temperature at the surface), *sea level pressure* (SLP, air pressure at sea level), *geopotential height* (Z, elevation of the 500mbar pressure level above the surface), *precipitable water* (PW, vertically integrated water content over the entire atmospheric column), *relative humidity* (RH, saturation of humidity above the surface), *horizontal wind speed* (WSPD, measured in the plane near the surface), and *vertical wind speed* (ω, measured in the atmospheric column). This is the first time such an extensive list of variables was used in a climate networks study.

These variables are available at daily intervals or as monthly averages over a period spanning more than sixty years (1948-present). However, in networks studies the goal is to capture the long-term climate variability, and therefore monthly averages are generally preferred. The data is arranged as points (grid cells) on a $2.5^\circ \times 2.5^\circ$ latitude-longitude spherical grid. In order to reduce the computational requirements (details in Section 2.3), the data may be sub-sampled to a coarser resolution (e.g., $5^\circ \times 5^\circ$ as in [19, 21]). A schematic diagram of the data for multiple variables at a single timestep t_i is depicted Fig. 1.



(a) Raw Data

(b) De-Seasonalized Data

FIGURE 2. The de-seasonlized data (right) exhibits significantly lower autocorrelation due to seasonality than the raw data (left).

2.2. **Seasonality and Autocorrelation.** The spatio-temporal nature of climate data poses a number of unique challenges. For instance, the data may be noisy and contain recurrence patterns of varying phase and regularity. Seasonality in particular tends to dominate the climate signal especially in mid-latitude regions, resulting in strong temporal autocorrelation (Fig. 2(a)). This can be problematic for identifying meaningful relationships between different locations, and indeed climate indices [28] are generally defined by the *anomaly series*, that is, departure from the "usual" behavior rather than the actual values.

Therefore, we follow precedent of related work [16, 21, 32] and remove the seasonal component from the data, specifically by monthly z-score transformation and de-trending [16]. At each grid point, we calculate for each month $m = \{1, ..., 12\}$ (i.e., separately for all Januaries, Februaries, etc.) the mean

$$\mu_m = \frac{1}{Y} \sum_{y=1948}^{2010} a_{m,y} \tag{1}$$

and standard deviation

$$\sigma_m = \sqrt{\frac{1}{Y-1} \sum_{y=1948}^{2010} (a_{m,y} - \mu_m)^2} \tag{2}$$

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. 2(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 discussed or used in the examples and case studies hereafter have been de-seasonalized and de-trended using the procedure described above.

2.3. **Network Construction.** In this section we describe the basic network construction process, which is shared by all lines of research on climate networks [3, 18, 21, 25, 32], with minor variations. Vertices of the network represent the spatial grid points of the underlying climate dataset, and weighted edges are created based on the statistical relationship between the corresponding pairs of (anomaly) time series [21]. It is important to note that the physical locality of grid points is *not* considered during network construction. Thus, any emerging cohesive patterns are the result of climatic similarity rather than spatial proximity.

2.3.1. *Estimating Link Strength.* Quantifying the relationship between a pair of vertices is critical to the network approach. Given that the data is normalized as described in Eqs. 1-3 we need not consider the mean behavior, only deviations from it. Therefore, the Pearson correlation coefficient is a logical choice as a measure of link strength [21]. For two series A and B of length t the correlation r is computed as

$$r(A,B) = \frac{\sum_{i=1}^{t}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{t}(a_i - \bar{a})^2 \sum_{i=1}^{t}(b_i - \bar{b})^2}} \tag{4}$$

where a_i is the i^{th} value in A and $\bar{a}$ is the mean of all values in the series. Note that the correlation coefficient has a range of $(-1, 1)$, where 1 denotes perfect agreement and -1 perfect disagreement, with values near 0 indicating no correlation. Since an inverse relationship is equally relevant in the present application we set the edge weight to $|r|$, the absolute value of the correlation coefficient.

We should note here that nonlinear relationships are known to exist within climate, which might suggest the use of a nonlinear correlation measure. Donges et al. [3] examined precisely this question in the context of network construction for climate and concluded that, "the observed similarity of Pearson correlation and mutual information networks can be considered statistically significant." Therefore, it seems sensible to use the simplest possible correlation measure, namely the (linear) Pearson coefficient. However, future work should further investigate this question, including a more comprehensive evaluation of different (nonlinear) correlation measures [11].

2.3.2. *Threshold Selection and Pruning.* Computing the correlation for all possible pairs of vertices results in a fully connected network but many (in fact most) edges have a very low weight, so that network pruning is desirable. And since it is impossible to determine an optimal threshold [15], we must rely on some other selection criterion. For example, Tsonis and Roebber [21] opt for a threshold of $r \geq 0.5$ while Donges et al. [3] use a fixed edge density ρ to compare different networks, noting that "the problem of selecting the exactly right threshold is not as severe as might be thought."

We would argue that a statistically principled approach is most appropriate here. Specifically, we propose using the *p-values* of the correlation coefficient to determine statistical significance [18]. Two vertices are considered connected only if the *p-value* of the corresponding correlation r is less than some (strict) threshold τ, imposing a very high level of confidence in that particular interaction. This may seem like a stringent requirement but in practice quite a large number of edges satisfy this criterion and are retained in the network.

3. Related Work

Before delving deeper into the various types of analysis performed on and corresponding insights gained from climate networks, we briefly point out two other interesting lines of research in climate science that also employ complex networks, albeit in a very different context. Both studies are fundamentally different from those discussed here in that the networks are constructed from very different types of data and designed to answer very specific questions.

The first of these involves the construction of networks from several major global climate indices, i.e., the Pacific Decadal Oscillation (PDO), the North Atlantic Oscillation (NAO), the El Niño Southern Oscillation (ENSO), and the North Pacific Oscillation (NPO) [22, 29]. Thus, the network consists of only four nodes (without any precise spatial locality) and six edges connecting them. The authors found that there are complex interactions between these indicators resulting in synchronization of the oscillations, but as the coupling strength increases the synchronous state is destroyed. This causes a major shift in global climate, and the NAO was identified as the primary participant in disturbing this process (both in observations and climate simulations).

The second study centers around hurricanes in the contintental United States [5, 7]. Specifically, networks are constructed from historical records of hurricanes that have affected multiple coastal regions. The authors find that the degree distribution is indicative of anomalous hurricane activity, and relating these anomalies to other climate events reveals strong links to sunspot activity and several of the major climate indicators. Moreover, based on these conclusions the authors discuss the potential effects of climate change on hurricane activity. The details of how the networks are constructed from observed data distinguish this as a particularly creative application of complex networks in climate science.

4. Topology and Structure at Multiple Scales

In this section, we describe several types of structural analysis for climate networks. Some are taken directly from complex networks literature, others are adapted or entirely novel to accomodate the unique properties of these spatio-temporal networks.

4.1. **Global Network Properties.**

First, one can examine the topological properties of the network at a global scale and interpret them in the context of climate [3, 18, 21, 25]. Standard measures from network analysis literature include:

- Number of nodes
- Number (or density) of edges
- Clustering coefficient (C) – indicative of the "cliquishness" of the network, this measure is computed for node i as

$$C_i = \frac{|e_{jk}|}{k_i(k_i - 1)} \tag{5}$$

  where e_{jk} is the set of all edges between first neighbors of i and k_i the degree of i, averaged over all nodes in the network.
- Characteristic path length (L) – expected distance between two randomly selected nodes in the network, computed by taking the mean over the all-pairs shortest paths.

Table 1 summarizes these for networks constructed from a wide range of climate variables. Also listed are the expected clustering coefficient and characteristic path length of a random graph with the same number of nodes and edges, estimated as

$$C_{rand} \approx \langle k \rangle / N \tag{6}$$

and

$$L_{rand} \approx ln(N)/ln(\langle k \rangle) \tag{7}$$

respectively, where $\langle k \rangle$ is the average degree and N the number of nodes in the network.

Due to the fixed data grid the number of nodes remains (nearly) constant, but the number of edges varies by as much as an order of magnitude. Nonetheless, all of the networks exhibit a high degree of clustering and short path lengths, and several researchers [3, 18, 21] have noted that climate networks of various types exhibit small-world properties [31]. Comparing the clustering coefficients and characteristic path lengths to those expected for random graphs, we find that in all cases $C \gg C_{rand}$ and $L \geq L_{rand}$, satisfying the properties of small-world networks [31].

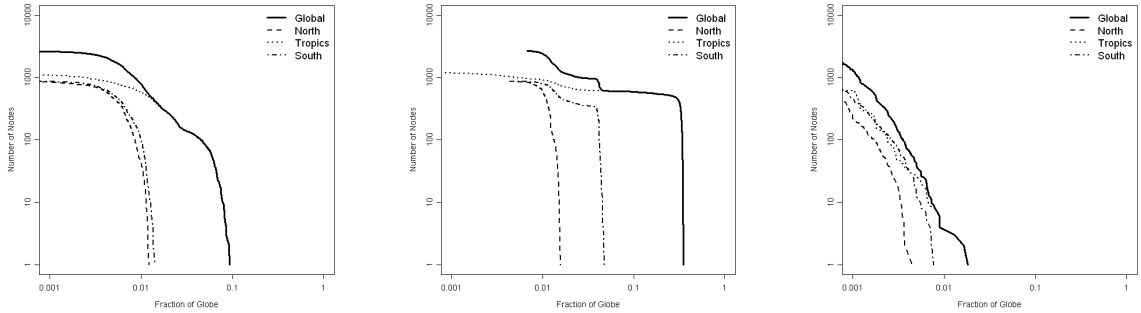
Variable	Nodes	Edges	C	L	C_{rand}	L_{rand}
SST	1,701	132,469	0.541	2.437	0.092	1.474
SLP	1,701	175,786	0.629	2.547	0.122	1.395
Z	1,701	249,322	0.673	2.436	0.172	1.310
PW	1,701	50,835	0.582	4.281	0.035	1.819
RH	1,700	25,375	0.559	4.063	0.018	2.190
WSPD	1,699	31,615	0.554	4.826	0.022	2.056
ω	1,701	71,458	0.342	2.306	0.049	1.679

TABLE 1. Summary of network properties: number of nodes/edges, average clustering coefficient (C), characteristic path length (L); expected values of C and L for random networks with the same number of nodes and edges.

While the aforementioned measures are commonly used to characterize many different kinds of networks, a quantity called *area weighted connectivity* was proposed specifically for networks constructed from data on a sphere [24]. If a node i is connected to N other nodes at λ_N latitudes, then its connectivity $\tilde{C}_i$ is computed as

$$\tilde{C}_i = \sum_{j=1}^{N} cos\lambda_j \Delta A / \sum_{over\ all\ \lambda\ and\ \varphi} cos\lambda \Delta A \tag{8}$$

where ΔA is the grid area and φ is the longitude [24]. We performed this calculation on the full network for each variable as well as for separate networks constructed from points only in the Northern (30°N-90°N), Tropical (30°S-30°N), and Southern (90°S-30S°) regions. This quantity can be plotted on a log-log plot, similar to a degree distribution; representative examples for three different variables are shown in Figure 3. Note the significant differences in distributions, which indicate that sea surface temperature and geopotential height are much more strongly connected overall than is vertical wind speed.



(a) Sea Surface Temp. (b) Geopotential Height (c) Vertical Wind Speed

FIGURE 3. Area weighted connectivity is an alternative network property for spatial data.

4.2. **Regional Network Properties.** The topological analysis can also lead to insights at the regional scale, that is, specific to certain parts of the network. For instance, the area weighted connectivity can also be plotted spatially on a map [24], as shown in Figure 4. Regions of high intensity are connected to a large fraction of the globe, and hence can be interpreted as having a significant role in the global climate system. The equatorial region spanning the Pacific Ocean, for example, is associated with the El Niño Southern Oscillation (ENSO) index [28] and therefore is

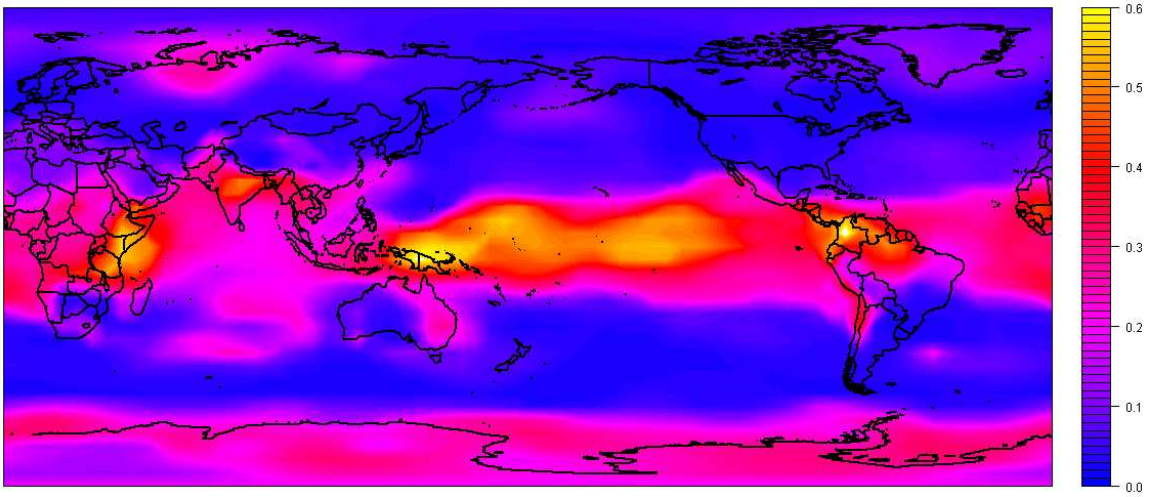


FIGURE 4. Area weighted connectivity for surface air temperature. The color scale indicates the fraction of the globe to which a point is connected via the network.

known to be one of the major global climate indicators. In fact, Tsonis and Swanson [24] have noted that the connectivity of the temperature network varies with the major El Niño and La Niña events.

Similarly, Donges et al. [3] plot other metrics such as the clustering coefficient as well as the betweenness and closeness centrality measures on a map to gain additional insights regarding the function and relative importance of different regions with respect to the global climate system.

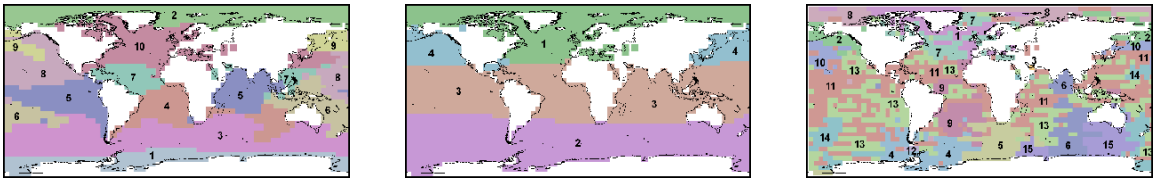
Another way that regional properties have been studied is by constructing separate networks for specific regions [32]. However, this approach is distinct from the general use of climate networks described here as the structure does not merely emerge from the properties of the network. Instead, some *a priori* knowledge is required to divide the globe (network) into meaningful partitions, usually guided by some a specific research question or hypothesis.

5. Clustering the Global Climate System

In contrast to the arbitrary partitioning of the network mentioned in Section 4.2, one may indeed be interested in clustering the climate data into regions defined by similarity in climatic variability. To this end, we have applied a community detection algorithm to climate networks [18, 19] (the term *community detection* refers to a broad class of algorithms also known as graph paratitioning, see [8, 17] for a more general description). Examples of the resulting clusters are shown in Figure 5.

The cluster structure provides rich information about the overall composition of the network and identifies closely related regions. For example, cluster 5 of sea surface temperature (Figure 5(a)) covers large portions of the Pacific and Indian Oceans, suggesting the presence of a *teleconnection* (long-range spatial dependency). In addition, comparing clusters of different variables helps in interpreting their role and relative importance in the global climate system.

In related work, Steinbach et al. [16] employed a shared nearest neighbor (SNN) algorithm to cluster climate data and demonstrated that some of the resulting clusters are significantly correlated with known climate indices while others may represent novel indicators. Although this approach does not involve climate networks in the strict sense, the SNN algorithm uses a network-like data representation. Moreover, this work was among the first to apply data mining concepts to address problems motivated by climate science.



(a) Sea Surface Temp. (b) Geopotential Height (c) Vertical Wind Speed

FIGURE 5. Clusters obtained by applying community detection on climate networks. The colors and numbers indicate unique clusters (arbitrary assignment).

6. Extending Climate Networks: Multivariate Relations and Network Dynamics

The methods discussed thus far have enabled compelling analyes and led to novel insights for the climate domain. However, they are limited in their representation of the complex relationships that are known to exist in the global climate system. We have identified two natural extensions to the general networks approch: First, the construction process should explicitly consider the possibility of multivariate relationships in climate networks. Second, climate dynamics should be incorporated by identifying, tracking, and interpreting changes in the network topology and/or cluster structure over time. In the following, we will briefly discuss each of these added dimensions, which we demonstrated in a recent case study [19] as a proof of concept.

6.1. **Multivariate Relationships in Climate.** The presence of relationships between different variables in the climate system is self-evident. In some cases, these interactions are grounded in physics and can be described by a set of equations; in other cases, the relationship may be observable but its exact nature remains unknown. Regardless, in order to create a more realistic representation of the climate system, the network model should incorporate the notion of multivariate relationsips [10]. In other words, we must replace the Pearson coefficient with an analogous measure for multivariate dependence. While conceptually intuitive, there is no obvious definition suitable in this context, and to our knowledge there are no straightfoward solutions to this problem in networks literature.

In [19], we present one (admittedly naïve) approach: we define a new feature space consisting of the pairwise correlations between a set of variables, and the network is weighted by the distance in this space. Formally, given a set of N variables one can compute $\binom{N}{2} = d$ pairwise correlations that define a corresponding feature space in $\mathbb{R}^d$. Edge weights are then calcuated as the distance (e.g., Euclidean) in this higher-dimensional space. When several variables behave similarly this distance will be small, so that a *lower* weight now indicates a stronger relationship.

Our experimental results demonstrate some success in the use of this definition of multivariate networks [19]. However, this distance measure is difficult to interpret and lacks the flexibility necessary for a general framework. Thus, univariate networks will continue to play an important role, but additional work is required in developing complementary multivariate approaches.

6.2. **Dynamics in Climate Networks.** Climate variability includes signals at annual and interannual scales, varying in both space and time, so that relationships in the climate system are constantly changing. However, the basic network model is unable to account for – much less detect – such changes in behavior.

A logical first step in addressing this issue is to construct multiple networks over time, as we have done in [19]. By dividing the data into windows and constructing a separate network at each step, we are able to measure the correspondence between consecutive windows and identify significant changes in structure. However, this case study represents a relatively simplistic approach focusing only on one particular aspect of the network structure.

7. Predictive Modeling in Climate Networks

This section highlights some of our most recent work and most important contributions in this area, which also serve as an example of advances enabled by an interdisciplinary research effort. Our motivations here were two-fold: first, a focus on the regional properties as defined by the cluster structure in climate networks (Section 5); second, a move beyond descriptive analysis and toward the development of predictive models for climate.

Our methodology rests on the observation that climate variability at different locations is intricately related, but the exact nature of these relationships is not well understood. More specifically, several major ocean climate indices are known to be strongly related with land climate [28]. These indicators are usually developed based on some observed phenomenon that is measured and quantified *a posteriori*, but what if we could extract this predictive information content from data?

In [16], the authors demonstrate that ocean clusters obtained using a traditional algorithm are correlated with known indices as well as land climate. However, climate networks enable us to answer this question more comprehensively using the same framework for descriptive analysis and predictive modeling. To this end, we construct networks consisting only of ocean regions and identify clusters using community detection. We then treat the cluster averages as potential climate indices by using them as inputs into a predictive model for land climate. Our preliminary results suggest that the ocean climate clusters contain significant information content, and that these models are better predictors of land climate than simple autoregressive methods. Thus, through the use of computational tools data mining is able to leverage the extensive corpus of observed climate data and confirm existing or even discover previously unknown relationships in the global climate system.

8. Computational Issues

There are numerous computational challenges that arise at various stages of the network construction and analysis process. First and foremost, calculating the pair-wise correlations between all grid points is a non-trivial task. In our experiments we used a coarse grid containing only $O(10^3)$ nodes, resulting in $O(10^6)$ pairs, and constructing the networks with simple Pearson correlation took several thousand CPU-hours. We used the statistical software package R[1] for our implementation and distributed the workload across 200 nodes of a dedicated high-performance computing cluster to make these operations computationally tractable.

However, multiple factors could (adversely) affect the computational demands of network construction. Using a higher-resolution spatial grid, for example, increases the number of nodes: the NCEP/NCAR Reanalysis data is available on a $2.5^\circ \times 2.5^\circ$ grid consisting of $O(10^4)$ nodes, thus resulting in $O(10^8)$ pairs. This would grow the problem size by two orders of magnitude, and even higher resolution datasets are available from other sources including those output by computational climate models. Whether such a network would yield any additional information is an open research question, but the sheer magnitude of the data makes this a challenging problem. In addition, substituting a different correlation measure could further drive up the computational requirements. For instance, one might want to estimate the mutual information between each pair to capture the nonlinear relationships in the time series. The exact computational demands would depend on the method used, but it would most certainly exceed those of the simple Pearson correlation.

Moreover, the generation of predictive models from the data poses additional challenges. Our work has focused mostly on linear regression models, computed first for only 10 regions but more recently at several hundred individual locations representing all land grid points around the globe. Still, even with several dozen input variables such models are easily built on a desktop computer. But lately we have been experimenting with more complex models such as support vector regression and neural networks, and learning these – especially in a large feature space – can become prohibitive. Thus, in addition to challenging mining and analysis tasks, there are more fundamental computer science problems regarding computing infrastructure and efficient implementation to be solved.

[1] http://www.r-project.org/

9. Future Work: Opportunities and Challenges

As outlined in this paper, the use of complex networks in climate is motivated by an acute need to fill gaps in understanding of the physical processes underlying the global climate system. Unlike traditional analysis methods, climate networks are capable of capturing complex relationships, discovering spatial structure and incorporating predictive modeling into a single framework. This network approach has already led to novel insights, and we believe it holds even greater potential. Lying at the intersection of multiple scientific disciplines, this emerging area of research is capable of bringing together experts from diverse backgrounds: climate scientists can contribute a wealth of data, domain expertise and exciting research questions; these, in turn, will motivate data miners to develop novel methods and algorithms to address the unique challenges arising from climate data.

In particular, we see three primary areas where future research has the potential for immediate and significant contributions:

(1) **Nonlinear** relationships are known to exist within climate data, but their relevance in the context of network construction have not been fully explored. As alluded to in Section 2.3, an extensive study comparing different correlation measures and their effect on network structure is needed in this regard.
(2) **Multivariate** relationships as described in Section 6.1 must be quantitatively captured and integrated with the networks to achieve a more realistic representation of the climate system. Advances in statistical and/or computational methods (e.g., see [10, 14]) may be necessary to devise a meaningful, interpretable measure of multivariate dependence.
(3) **Spatio-Temporal** relationships and network dynamics are arguably the area in most need of an interdisciplinary research effort. Changes in network structure over time should be automatically detected and, where possible, related to external events for validation or interpretation.

Advancing towards these goals will necessitate the development of novel algorithms and efficient implementations thereof. Datasets continue to increase in size, and expanding the scope of analysis to include more variables or allow for the presence of additional spatial and/or temporal lags further compounds the complexity of the problem. Therefore, it is imperative that data miners work in close collaboration with climate scientists to ensure that their solutions adequately and completely address relevant questions in the domain.

10. Acknowledgements

This work was supported in part by the National Science Foundation under Grant No. 0826958. The research was performed as part of a project titled "Uncertainty Assessment and Reduction for Climate Extremes and Climate Change Impacts", which in turn was funded in FY2009-10 by the initiative called "Understanding Climate Change Impact: 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 DE-AC05-00OR22725. 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 United States Government purposes.

References

[1] A.-L. Barabási and E. Bonabeau. Scale-free networks. *Scientific American*, 288:50–59, 2003.
[2] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.
[3] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *European Physics Journal Special Topics*, 174:157–179, 2009.
[4] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *Europhysics Letters*, 87:48007, 2009.

[5] J. B. Elsner, T. H. Jagger, and E. A. Fogarty. Visilibity network of united states hurricanes. *Geophysical Research Letters*, 36:L16702, 2009.

[6] P. Erdös and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Acadademy of Sciences*, 5:17–61, 1960.

[7] E. A. Fogarty, J. B. Elsner, T. H. Jagger, and A. A. Tsonis. Network Analysis of U.S. Hurricanes. In *Hurricanes and Climate Change*, pages 153–167. Springer Science + Business Media, LLC, 2009.

[8] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99:7821–7826, 2002.

[9] E. Kalnay et al. The NCEP/NCAR 40-Year Reanalysis Project. *BAMS*, 77(3):437–470, 1996.

[10] S.-C. Kao, A. R. Ganguly, and K. Steinhaeuser. Motivating complex dependence structures in data mining: A case study with anomaly detection in climate. In *IEEE ICDM Workshop on Knowledge Disocvery from Climate Data*, pages 223–230, 2009.

[11] S. Khan, S. Bandyopadhyay, A. R. Ganguly, et al. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2):026209, 2007.

[12] J.-P. Onnela, J. Saramäki, K. Kaski, and J. Kertész. Financial market - a network perspective. In *Practical Fruits of Econophysics*, pages 302–306. Springer, 2006.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[14] C. Schoölzel and P. Friederichs. Multivariate non-normally distributed random variables in climat eresearch – introduction to the copula approach. *Nonlin. Proc. Geophys.*, 15:761–772, 2008.

[15] A. Serrano, M. Boguna, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences USA*, 106(16):8847–8852, 2009.

[16] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. P. r. Discovery of Climate Indices using Clustering. In *ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, pages 446–455, 2003.

[17] K. Steinhaeuser and N. V. Chawla. Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31(5):413–421, 2010.

[18] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. Complex networks as a unified framework for descriptive analysis and predictive modeling in climate. Article in review.

[19] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. An exploration of climate data using complex networks. *ACM SIGKDD Explorations*, 12(1), 2010.

[20] A. A. Tsonis. *Nonlinear Dynamics in Geosciences*, chapter 1, pages 1–15. Springer, New York, 2007.

[21] A. A. Tsonis and P. J. Roebber. The architecture of the climate network. *Physica A*, 333:497–504, 2004.

[22] A. A. Tsonis, K. Swanson, and S. Kravtsov. A new dynamical mechanism for major climate shifts. *Geophysical Research Letters*, 34(L13705), 2007.

[23] A. A. Tsonis and K. L. Swanson. On the role of atmospheric teleconnections in climate. *Journal of Climate*, 21:2990–3001.

[24] A. A. Tsonis and K. L. Swanson. Topology and Predictability of El Niño and La Niña Networks. *Physical Review Letters*, 100(228502), 2008.

[25] A. A. Tsonis, K. L. Swanson, and P. J. Roebber. What Do Networks Have to Do with Climate? *Bulletin of the American Meteorological Society*, 87(5):585–595, 2006.

[26] P. Uetz, L. Giot, G. Cagney, et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, 1999.

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

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

[29] G. Wang, K. L. Swanson, and A. A. Tsonis. The pacemaker of major climate shifts. *Geophysical Research Letters*, 36:L07708, 2009.

[30] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.

[31] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.

[32] K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate Networks around the Globe are Significantly Affected by El Niño. *Physical Review Letters*, 100(22):157–179, 2008.

[33] K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate Networks Based on Phase Synchronization Analysis Track El-Niño. *Progress of Theoretical Physics*, Supplement No. 179:178–188, 2009.