2011 11th IEEE International Conference on Data Mining Workshops

Community Dynamics and Analysis of Decadal Trends in Climate Data

William Hendrix*‡, Isaac K. Tetteh†, Ankit Agrawal*,
Fredrick Semazzi†, Wei-keng Liao*, and Alok Choudhary*
*Dept. of Electrical Engineering and Computer Science
Northwestern University, Evanston, IL 60208
†Dept. of Marine, Earth, and Atmospheric Sciences
North Carolina State University, Raleigh, NC 27695-8208
‡Corresponding author: whendrix@eecs.northwestern.edu

***Abstract*—The application of complex networks to study complex phenomena, including the Internet, social networks, food networks, and others, has seen a growing interest in recent years. In particular, the use of complex networks and network theory to analyze the behavior of the climate system is an emerging topic. This newfound interest is due to the difficulty of analyzing climate data—this analysis is notoriously difficult due to the strong spatio-temporal dependencies, multivariate nature, seasonal behavior, and nonlinear phenomena inherent in the climate system. Network-based approaches model the complex long-term dependencies of weather attributes (such as temperature or air pressure) between locations on the Earth as a network of relationships and analyze these networks to gather insights about the emergent behavior of the system as a whole.**

In this paper, we describe our work-in-progress on a methodology for capturing and characterizing the evolution of the climate network. We do this by splitting the climate data into a set of overlapping decadal time windows and forming a network for each of these datasets representing the complex interdependencies in the climate system over the particular decade. We can then use this sequence of networks to characterize major patterns and anomalies in the data. We validate our methodology by identifying nontrivial events and trends in the evolution of the decadal networks and correlating these with known climatological phenomena.

***Keywords*-climate application, complex network analysis, time-evolving graphs, clustering**

I. Introduction

The application of complex networks to study complex phenomena, including the Internet, social networks, food networks, and others, has seen a growing interest in recent years [1]–[4]. In particular, the use of complex networks and network theory to analyze the behavior of the climate system is an emerging topic [4]–[11]. This newfound interest is due to the difficulty of analyzing climate data—this analysis is notoriously difficult due to the strong spatio-temporal dependencies, multivariate nature, seasonal behavior, and nonlinear phenomena inherent in the climate system. Network-based approaches model the complex long-term dependencies of weather attributes (such as temperature or air pressure) between locations on the Earth as a network of relationships and analyze these networks to gather insights about the emergent behavior of the system as a whole. This type of technique might, for example, be used to identify and evaluate the likelihood of major climate shifts in the output from various General Circulation Models (GCMs).

One of the limitations of several previous approaches is that they use all of the available data to construct these networks of dependencies. While this decision may seem to be the way to construct the most "accurate" network, it only allows the construction of a single, static network that neglects the changing nature of the climate. As such, these approaches are limited in their ability to detect emerging trends or large-scale anomalies in the behavior of the climate system. To the best of our knowledge, only one other work [10] has tried to model the evolution of the climate system as a sequence of global-scale climate networks; however, the results we present here, which are based on an analysis of the evolving structure of the correlations in air temperature values, offer a more natural physical interpretation than the clusters formed based on Euclidean distance between pairwise correlations of four variables presented in [10].

In this paper, we describe a technique for modeling and discovering patterns and anomalies in the decadal-scale character of the climate system by analyzing a sequence of complex networks formed by advancing a 10-year time window along the climate data. We describe our technique in Section II, present some preliminary observations made by applying our technique to real climate data in Section III, and conclude in Section IV.

II. Methodology

In this section, we describe our technique for analyzing the temporal structure and evolution of the climate network. This technique consists of five main steps:

1) Initially, we process the data to reduce the effects of seasonality in the data. Though the seasonal trends in the data cannot be eliminated entirely, these patterns would overwhelm other interesting patterns in the data without some form of preprocessing to limit their effects.
2) With this processed data, we create a series of overlapping 10-year time windows, advancing by one

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.51

year in between each pair of consecutive windows. By allowing a significant (9-year) overlap between consecutive time windows, we hope to see the large-scale changes in the climate network emerge gradually rather than appearing as chaotic noise.

3) Using this time-windowed data, we form a climate network for each time window to represent the large-scale dependencies in the climate patterns between locations on the Earth, generating a sequence of networks representing the evolution of the climate system over time. These networks use a set of latitude/longitude grid points as their vertices and connect vertices if the two time series associated with the points are correlated.
4) Next, we apply a clustering algorithm to each time window in order to group together sets of grid points within the window that share similar behavior.
5) Finally, we compare the networks representing consecutive time periods in order to identify clusters that appear to be *stable*, or persist from one time window to the next.

Once we have identified these stable clusters, we visualize their evolution in order to identify major trends and anomalies in the climate networks. Our results, presented in Section III, are the product of a visual analysis of the progression of these stable clusters.

A. Data

For our analysis, we analyze the monthly mean Surface Air Temperature variable from the NCEP Reanalysis project [12]. This dataset has a resolution of 2.5° latitude by 2.5° longitude, and data is available for every grid point over a period from January 1948 through December 2010 (63 years), for a time series of 756 monthly values. Data is recorded for each latitude from 90°N to 90°S, but at the two poles (90°N and 90°S), all longitude values represent the same location, so we only use data from 87.5°N to 87.5°S, for a total of 10,224 grid points (71×144).

B. Data preprocessing

While the analysis of air temperature data can yield valuable insights into the behavior of the climate system as a whole, such insights are difficult to detect in the raw data due to the presence of very strong seasonal effects. Thus, to unmask more interesting patterns and anomalies, we need to process the data to reduce the effects of this seasonality.

To reduce the effects of seasonality on the Reanalysis data, we calculate monthly *anomaly series* values, or the amount that a particular month deviates from the average behavior of the temperature for that variable. To calculate these anomaly series values, we divide the 756-month data values at each grid point into twelve 63-month time series, one corresponding to each month of the year. For each of these monthly time series, we subtract the mean and divide by the standard deviation in order to remove the annual trends and normalize before recombining the data into the full 63-year anomaly time series.

Having limited the effects of seasonality in the data through the calculation of this anomaly time series, we divide the data into overlapping ten-year time windows, advancing the time window by one year until we reach the end of our data. By allowing a significant (9-year) overlap between successive time windows, the large-scale changes in the climate system should emerge gradually enough to allow us to separate longer-term patterns and trends from noise in the data.

C. Forming the network

For each of these time windows, we wish to create a network representing the significant dependencies between the various regions of the globe. We form this network by calculating the absolute value of the Pearson correlation coefficient between every pair of edges and applying a threshold (cutoff) value so that each edge that meets or exceeds this cutoff will represent a strong positive or negative linear dependency between the grid points that the edge connects. We also apply a distance threshold to reduce the effects of spatial autocorrelation in the data, or the tendency of co-located grid points to display very similar behavior.

Even though the climate system exhibits nonlinear phenomena, such as matter and energy transfer, we use the linear Pearson correlation coefficient rather than a measure like mutual information that can capture nonlinear dependencies, as it has already been reported in literature [13] that the local and mesoscopic properties of the networks generated using these two measures are extremely similar.

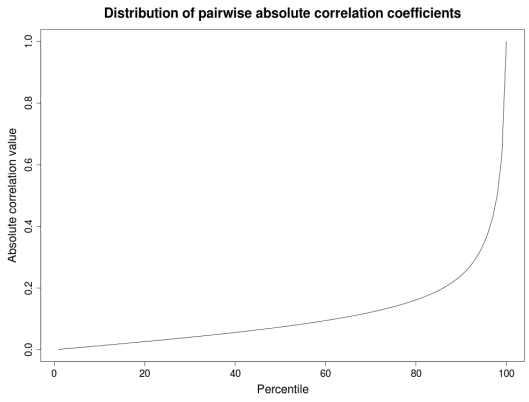
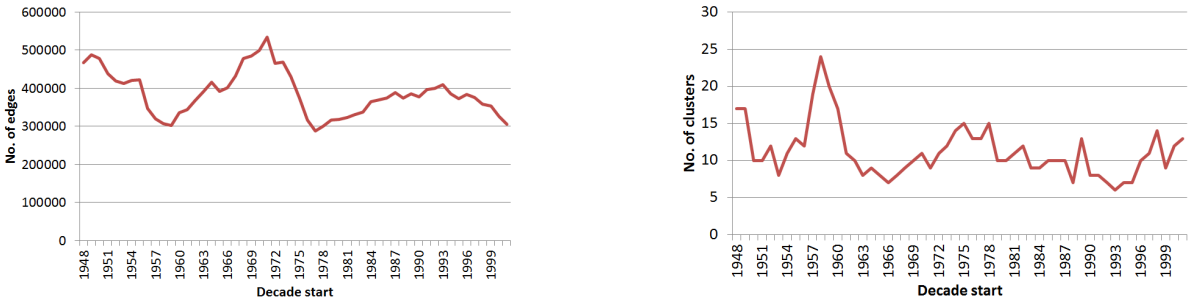


Figure 1. Distribution of the absolute value of the all-pair correlation values for all latitude/longitude grid points over the entire 63-year dataset. In this work, we use the 99th percentile correlation value (0.6307) as a cutoff value when forming the decadal climate networks.

Further, in order to ensure consistency between the edges in multiple time windows, we selected a single correlation threshold that we applied to each time window based on the distribution of the absolute value of the correlation



(a) (b)

Figure 2. Summary of the number of edges (a) and clusters (b) identified in the constructed decadal climate networks. The horizontal axis in both graphs denotes the first year of the decade.

coefficient across all pairs of grid points using the full, 63-year time series. This distribution appears in Figure 1. The threshold value we chose was the 99th percentile of the pairwise correlation values, 0.6307.

Finally, to reduce the effects of spatial autocorrelation in the data, we pruned the edges in each of the networks that represented correlations between points that were less than 500 km apart. This 500 km threshold was chosen to be large enough to eliminate edges between adjacent grid points at the equator, but small enough so as not to eliminate all of the significant correlations associated with grid points near the poles. The distance between grid points was approximated as the length of a geodesic on a sphere of radius 6371 km, using the formula

$$d = 6371 \cos^{-1}(\cos(\theta_1 - \theta_2)\cos(\varphi_1)\cos(\varphi_2) + \sin(\varphi_1)\sin(\varphi_2)),$$

where (θ_1, φ_1) are the latitude and longitude, respectively, of the first grid point, in radians, and (θ_2, φ_2) are the latitude and longitude of the second. While this technique does not eliminate the effects of spatial autocorrelation in the data (as can be seen from our results in Section III), it does help to uncover more interesting long range effects in the data.

D. Extracting the large-scale structure of the network

For each of these networks, we wish to capture some features that capture the overall structure of the time-dependent networks in such a way that we can identify the patterns and track the anomalies and emerging trends. To this end, we adopt a methodology based on clustering the networks and detecting the clusters that exhibit stability between consecutive time windows.

To cluster the networks, we applied the CNM clustering algorithm [14], an efficient modularity-based clustering algorithm implemented in the Small-world Network Analysis and Partitioning (SNAP) software package [15]. We chose CNM for its efficiency, its ready availability, and its use of modularity as an optimization criterion. In future work, we may investigate the applicability of other clustering algorithms to the climate network.

We say that clusters in consecutive time windows are related if they share an average of at least 80% of their members. More formally, clusters c_1 and c_2 are related iff

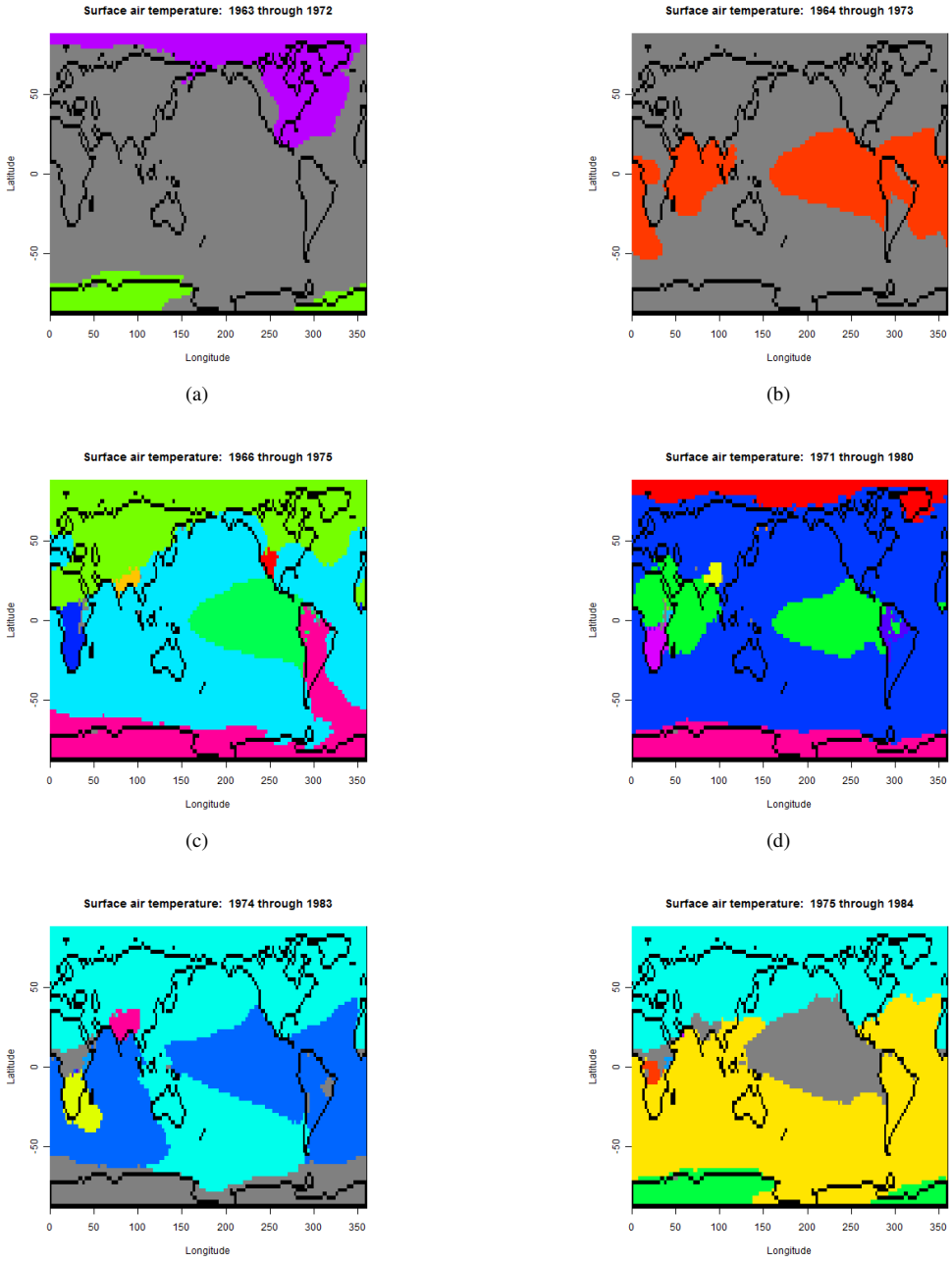
$$\frac{1}{2}\left(\frac{|c_1 \cap c_2|}{|c_1|} + \frac{|c_1 \cap c_2|}{|c_2|}\right) \geq 0.8,$$

where $|c_1|$ is the number of grid points in cluster c_1, $|c_2|$ is the number of grid points in c_2, and $|c_1 \cap c_2|$ is the number of grid points in common between clusters c_1 and c_2. We chose a threshold of 0.8 because a threshold larger than 0.75 ensures that each cluster can only be associated with one cluster in each of its adjacent time windows. (If one cluster were associated with two clusters in a single adjacent time window, the smaller of the two clusters could have an overlap of up to 50% with the original, leading to an average overlap of at most 75%.)

III. Results and Discussion

For the purpose of this work, we present some observations of the resulting network sequence and correlate these observations with known climatological phenomena. Using the methodology described in Section II, we visualized the resulting sequence of network clusters using R [16], an open source statistical software package. Unless otherwise noted, all of the clustering results we present in this section represent the stable clusters identified by our methodology, with instable clusters and singleton clusters appearing gray. Full results of our technique applied to surface air temperature data, including both the stable clusters and the full network clustering, can be viewed at http://cucis.ece.northwestern.edu/projects/Expeditions/climkd11_results.html.

First, we present some basic statistics of the time-sensitive climate networks that we constructed. Figure 2 gives a summary of how the number of clusters and the number of edges in the networks changes over time. Visually inspecting the trends in the total number of edges indicates a signal reminiscent of a low frequency phenomenon dominated by interdecadal variability, with maxima during the time windows 1949–1958, 1971–1980, and 1993–2002 and minima



(a) (b) (c) (d) (e) (f)

Figure 3. Illustration of the evolution of a stable cluster linking the surface air temperature in the Nino-3 region with the Indian Ocean during the period 1963–1983. The cluster, which first appears during the time window 1963–1972 (b), is not apparent in the time window 1962–1971 (a). The cluster becomes unstable in the time windows 1966–1975 (c) and 1971–1980 (d), though the teleconnection between the Nino-3 region and the Indian Ocean is still evident in 1971–1980. The cluster makes its final appearance in 1974–1983 (e, f). In figures (c) and (d), the various colors represent the grid points assigned to each of clusters identified in the time-dependent networks, regardless of cluster stability, and gray represents grid points not assigned to a cluster. The other images show only the stable clusters in color and depict unstable clusters in gray.

during 1959–1968 and 1977–1986. As might be expected, the number of clusters in the graph follows an inverse pattern with respect to the number of edges in the networks. While the period between the absolute maximum and minimum at 1971–1980 and 1977–1986 represents the steepest decline in the number of significant correlations, though, the largest number of clusters occurred in 1958–1967, one year before a local minimum in the number of edges, and the smallest

coincided with a local maximum in 1993–2002. These twin patterns suggest a possible link to a modulation of a planetary-scale climatic pattern. This type of pattern has been previously observed in the literature [6].

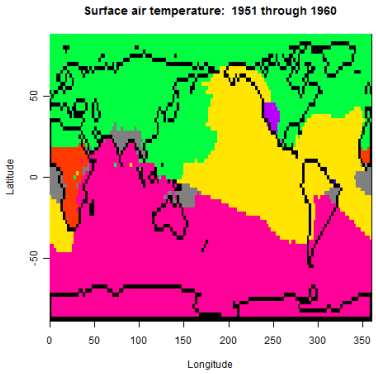
We now turn our attention to the sequence of stable clusters identified by our methodology in Section II. One notable feature of these clusters is that the vast majority are geographically contiguous, despite being formed from networks that have no inherent geographical information. Though this feature is almost certainly a product of the spatial autocorrelation in the data, these contiguous clusters correspond well to our intuition that climate zones enclose closed shapes.

Another notable feature of the resulting networks is a persistent teleconnection between the Nino-3 region of the Pacific Ocean and the Indian Ocean that first appeared in the 1963–1972 decade and lasted until 1974–1983, with two brief destabilizations in 1966–1975 and 1971–1980 (see Figure 3). This teleconnection is a known feature of the El Niño phenomenon [17], and the year 1972 corresponds to a peak in El Niño activity [18].

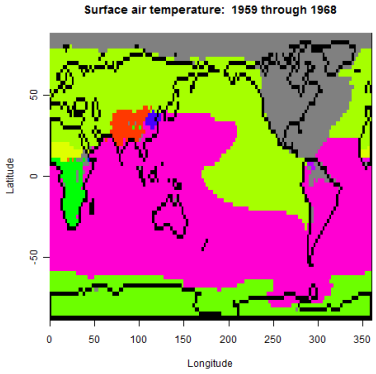
One other feature of the data feature of the data is a shift in how the clusters divide the African continent. In the time windows prior to 1953–1962, the clusters divided northern Africa from central Africa between $17.5°$N and $20°$N; however, after the decade 1962–1971, this division was moved further south, appearing consistently around $5°$N to $10°$N. In between these time periods, a separate cluster emerged, covering the area between $10°$N–$17.5°$N and $15°$W–$32.5°$E appeared from 1956–1965 until 1961–1970, excepting the window 1960–1969 (see Figure 4). This zone covers the Sahel region in Africa, the zone immediately south of the Sahara desert. The Sahel region went through a period of strong rainfall in the 1950's, but it experienced a major drought starting in 1968 and has undergone significant period of desertification in the intervening years [19], [20].

IV. Conclusions and Future Directions

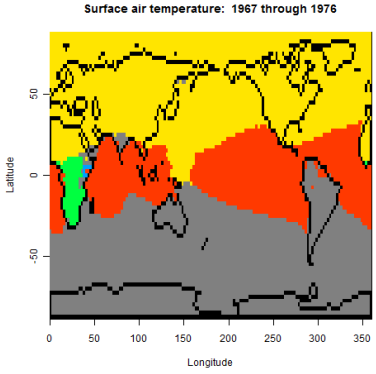
In summary, we have presented a methodology to tackle the challenging problem of identifying the patterns, trends, and anomalies in the evolution of the climate system, and we have validated our methodology by using it to identify interesting, nontrivial features of the resulting climate networks that correspond to known climatological features and events. Potential future directions of this work include the extension of our methodology to multiple variables; the evaluation of different processing methods, such as the different clustering or seasonal adjustment techniques; the development of techniques to automatically detect motifs or anomalies in the evolving climate networks; and the application of these techniques to improve existing climate models or improve forecasting.



(a)



(b)



(c)

Figure 4. Illustration of the shift of the Sahel region in Africa from being clustered with central/southern Africa (a) to being its own stable cluster (b) to being clustered with northern Africa (c). In these figures, each of the solid colors represent grid points assigned to the same cluster, with gray representing unstable clusters or grid points not assigned to a cluster.

Acknowledgments

This work is supported in part by NSF award numbers CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CCF-1029166, and OCI-1144061, and in part by

DOE grants DE-FC02-07ER25808, DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, and DE-SC0005340.

REFERENCES

[1] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, "The backbone of the climate network," *EPL (Europhysics Letters)*, vol. 87, no. 4, pp. 48 007+, Feb. 2010.

[2] A. Reka and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, pp. 47–97, Jun. 2002.

[3] J. Camacho, R. Guimerà, and L. A. N. Amaral, "Robust patterns in food web structure," *Physical Review Letters*, vol. 88, no. 22, p. 228102, 2002.

[4] A. A. Tsonis and P. J. Roebber, "The architecture of the climate network," *Physica A: Statistical and Theoretical Physics*, vol. 333, pp. 497–504, Feb. 2004.

[5] A. Gozolchiani, K. Yamasaki, O. Gazit, and S. Havlin, "Pattern of climate network blinking links follows el niño events," *EPL (Europhysics Letters)*, vol. 83, no. 2, p. 28005, 2008.

[6] A. A. Tsonis, K. Swanson, and S. Kravtsov, "A new dynamical mechanism for major climate shifts," *Geophysical Research Letters*, vol. 34, pp. L13 705+, Jul. 2007.

[7] A. A. Tsonis and K. L. Swanson, "Topology and predictability of el niño and la niña networks," *Physical Review Letters*, vol. 100, no. 22, 2008.

[8] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, "Complex networks in climate dynamics," *The European Physical Journal – Special Topics*, vol. 174, pp. 157–179, 2009.

[9] A. R. Ganguly, K. Steinhaeuser, D. J. Erickson, M. Branstetter, E. S. Parish, N. Singh, J. B. Drake, and L. Buja, "Higher trends but larger uncertainty and geographic variability in 21st century temperature and heat waves," *Proceedings of the National Academy of Sciences*, vol. 106, no. 37, pp. 15 555–15 559, Sep. 2009.

[10] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "An exploration of climate data using complex networks," in *SensorKDD '09*. ACM, 2009, pp. 23–31.

[11] ——, "Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science," *Statistical Analysis and Data Mining*, 2010.

[12] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, and et al., "The NCEP/NCAR 40-Year Reanalysis Project." *B. Am. Meteorol, Soc.*, vol. 77, pp. 437–472, Mar. 1996.

[13] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, "Complex networks in climate dynamics," *The European Physical Journal - Special Topics*, vol. 174, pp. 157–179, 2009, 10.1140/epjst/e2009-01098-2. [Online]. Available: http://dx.doi.org/10.1140/epjst/e2009-01098-2

[14] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks." *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 70, no. 6, pp. 66 111–1–6, 2004.

[15] D. A. Bader and K. Madduri, "Snap, small-world network analysis and partitioning: an open-source parallel graph framework for the exploration of large-scale networks," in *22nd IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2008.

[16] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[17] K. K. Kumar, B. Rajagopalan, M. Hoerling, G. Bates, and M. Cane, "Unraveling the mystery of indian monsoon failure during el niño," *Science*, vol. 314, no. 5796, pp. 115–119, 2006.

[18] W. H. Quinn, "The large-scale enso event, the el niño and other important regional features," *Bulletin de l'Institut Francais d'Etudes Andines*, vol. 22, no. 1, pp. 13–34, 1993.

[19] C. J. Tucker, H. E. Dregne, and W. W. Newcomb, "Expansion and contraction of the sahara desert from 1980 to 1990," *Science*, vol. 253, pp. 299–301, 1991.

[20] L. Olsson, L. Eklundh, and J. Ardo, "A recent greening of the sahel – trends, patterns, and potential causes," *Journal of Arid Environments*, vol. 63, pp. 556–566, 2005.