Complex Networks as a Unified Framework for Descriptive Analysis and Predictive Modeling in Climate Science

Karsten Steinhaeuser[1,2], Nitesh V. Chawla[1*] and Auroop R. Ganguly[2]

[1]*Department of Computer Science and Engineering, Interdisciplinary Center for Network Science and Applications, University of Notre Dame, Notre Dame, IN 46556, USA*

[2]*Geographic Information Science and Technology Group, Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*

Received 5 March 2010; revised 5 November 2010; accepted 10 November 2010
DOI:10.1002/sam.10100
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: The analysis of climate data has relied heavily on hypothesis-driven statistical methods, while projections of future climate are based primarily on physics-based computational models. However, in recent years a wealth of new datasets has become available. Therefore, we take a more data-centric approach and propose a unified framework for studying climate, with an aim toward characterizing observed phenomena as well as discovering new knowledge in climate science. Specifically, we posit that complex networks are well suited for both descriptive analysis and predictive modeling tasks. We show that the structural properties of 'climate networks' have useful interpretation within the domain. Further, we extract clusters from these networks and demonstrate their predictive power as climate indices. Our experimental results establish that the network clusters are statistically significantly better predictors than clusters derived using a more traditional clustering approach. Using complex networks as data representation thus enables the unique opportunity for descriptive and predictive modeling to inform each other. © 2010 Wiley Periodicals, Inc. Statistical Analysis and Data Mining, 2010

Keywords: complex networks, climate data, network analysis, community detection, multivariate predictive modeling

1. INTRODUCTION

Identifying and analyzing patterns in global climate is an important task, because it helps climate scientists develop a deeper understanding of the complex processes underlying observed phenomena. For much of our planet's past only rudimentary climate records are available, and efforts to study them have relied heavily on—and driven the development of—statistical methods [1] and physics-based computational models (e.g., Ref. 2), which use first principles to combine various earth system components and project future climate. However, advances in modern technology (like satellites in the 1950s) have greatly increased our ability to monitor climate and rapidly provide massive amounts of data, presenting us with an opportunity to induce transformative changes in the way we analyze and understand the earth's climate system. These data exhibit a number of traits that have the potential to not only complement hypothesis-driven research but also enable the discovery of new hypotheses or phenomena from the rich data. Specifically, these traits include: (i) greater spatial coverage and higher resolution; (ii) extended temporal span; (iii) observational records; (iv) reanalysis data, which is a hybrid of observed and model-simulated data (see Section 2); (v) multiple vetted data sources; and (vi) a vibrant research community.

Data of such extent and longitudinal character brings novel challenges for data-driven science for charting the path from data to knowledge to insight. The process of data-guided knowledge discovery will entail an integration of descriptive analysis and predictive modeling for 'useful insights' (hypotheses) from the data, which can then be validated against observed phenomena. This unified framework prompts 'networked' thinking: imagine the globe as a spatiotemporal grid. Each cell, corresponding to a region, can be represented by a node, and different nodes are connected to each other not by spatial proximity, but rather on the basis of similarity shared in climatic

Correspondence to: Nitesh V. Chawla (nchawla@nd.edu)

variability. Such interactions among nodes can be exploited to discover how regions are related and impact each other.

To this end, we make the case that complex networks offer a compelling perspective for capturing the dynamics in climate data not only for descriptive analysis but also predictive modeling. The concept of climate networks was first proposed by Tsonis and Roebber [3], but their use has been limited to describing physical properties of the climate system and comparing them to known phenomena. There also exists prior work that applies data mining techniques in climate, specifically to discover ocean climate indices from historical data via clustering and correlating the clusters with climate on land [4]; this approach is extended by Lin *et al*. [5] by employing the same clustering algorithm but building association rules between ocean and land regions. However, these approaches are limited in that they only consider a single climate variable. In contrast, we take the broader view by comparing networks constructed from several climate variables separately and capture their interactions in a multivariate predictive model, paving the path from data to knowledge to insight. We should note here that the multivariate nature of the data can be considered at several points in the process, that is, during network construction, clustering, or prediction. Here we combine the data during the predictive modeling stage as the construction of multivariate networks directly as well as the coclustering of multiple networks are themselves separate areas of active research (see, e.g., Ref. 6).

Contributions: We focus on the challenges of regional predictions and precipitation; as highlighted in a recent article [7], regional predictions and precipitation remain among the four 'real holes in climate science'. We investigate whether relatively hypothesis-free, data-guided knowledge discovery has the ability to advance the state of climate science in these areas and complement the predictive power of physics-based models. To this end, we present an innovative approach that encompasses both *descriptive analysis* and *predictive modeling*. In particular, we posit that using complex networks as data representation provides a unified framework for identifying and characterizing patterns in the data as well as developing predictive insights, while enabling the analysis and modeling tasks to inform each other in unprecedented ways. The results and analysis presented here are distinct from prior work in that we construct networks from multiple climate models and compare/contrast their properties, we use community detection to identify homogeneous climate regions, and we learn multivariate predictive models. The technical and methodological contributions can be summarized as follows:

- Complex networks from a wide range of climate variables (Sections 2 and 3).
- Analysis of the properties of networks to gain insights in the climate domain (Section 4).
- Derivation of multivariate ocean climate indices from network clusters and show that they are *statistically significantly* better predictors of land climate than clusters obtained with a traditional clustering method (Section 5).
- Offer compelling perspective on complex networks as unifying framework and the unique opportunity for descriptive analysis and predictive modeling tasks to inform each other (Section 6).

2. CLIMATE DATA

In the following, we describe the characteristics of the dataset used in this study as well as the preprocessing steps required for our analysis.

2.1. Dataset Description

The data stems from the NCEP/NCAR Reanalysis Project [8] (available at Ref [9]). This reanalysis 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 (it is obviously impossible to obtain exact measurements). We did not want to constrain ourselves by an arbitrary *a priori* selection of variables, so we consider a wide range of surface and atmospheric 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* (GH, elevation of the 500 mbar 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* (HWS, measured in the plane near the surface), and *vertical wind speed* (VWS, measured in the atmospheric column). This is the first time such a wide range of climate variables is used in a climate networks study.

These variables are available as monthly averages over a period of 60 years (1948–2007), for a total of 720 data points. The data is spatially arranged as points (grid cells) and we consider a $5^\circ \times 5^\circ$ latitude–longitude spherical grid. A schematic diagram of the data for a single timestep t_i depicted in the rectangular plane is shown in Fig. 1. For the purpose of this study, we only use the data over the oceans as we are ultimately interested in finding relationships between ocean indicators and land climate (see Section 5).

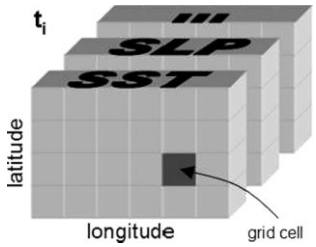

Fig. 1 Schematic depiction of gridded data for a single timestep t_i in the rectangular plane

2.2. Countering Seasonality and Autocorrelation

The spatiotemporal 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 prediction, and indeed climate indices [10] are generally defined by the *anomaly series*, that is, departure from the 'usual' behavior rather than the actual values.

Therefore, we follow the precedent of related work and compute anomaly values [11–13]. Specifically, we remove the seasonal component by monthly z-score transformation and detrending as described in Ref. 4. At each grid point, we calculate for each month $m = \{1, \ldots, 12\}$ (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. 2(b), which shows that deseasonalized values have significantly lower autocorrelation than the raw data. In addition, we detrend the data by fitting a linear regression model and retaining only the residuals. For the remainder of this article, all data used for experiments or discussed hereafter have been deseasonalized and detrended as described above.

3. CLIMATE NETWORKS

The global climate system can be represented by a network of interacting regions, connected by relationships derived from their climatic variability. The intuition behind this methodology is that the dynamical behavior of the system can be captured in the local and global topology of a complex network. Nodes of such a network correspond to spatial grid points of the underlying dataset, and weighted edges are created based on the statistical relationship between the corresponding pairs of (anomaly) time series [3,11,12,14]. In this section, we describe the network construction process in more detail.

3.1. Estimating Link Strength

Quantifying the relationship between a pair of nodes is critical to the network approach. Given that the data is normalized as described in Eqs. 1–3, we need not consider

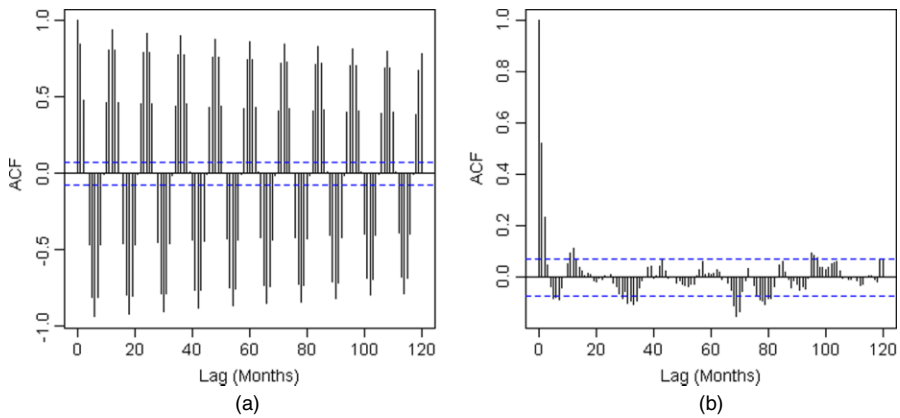

Fig. 2 The deseasonalized data (b) exhibits significantly lower autocorrelation due to seasonality than the raw data (a)

the mean behavior, only deviations from it. Therefore, the Pearson correlation coefficient is a logical choice as a measure of link strength [3]. For two series A and B of length t the correlation r is computed as

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

where a_i is the ith value in A and $\overline{a}$ 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 a couple of caveats here: (i) given that the networks span the entire globe, temporal lags may affect the correlation strength over long distances. However, we ignore lags in our analysis as the use of monthly data should mitigate their influence to some degree, though future work could investigate the possibility of lags and their effect on network structure. Second, nonlinear relationships are known to exist within climate, which might suggest the use of a nonlinear correlation measure. However, Donges et al. [15] 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 is sensible to use the simplest possible correlation measure, namely the (linear) Pearson coefficient.

3.2. Threshold Selection and Pruning

Computing the correlation for all possible pairs of nodes 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 [16], we must rely on some other selection criterion. For example, Tsonis and Roebber [3] opt for a threshold of $r \geq 0.5$ while Donges et al. [15] 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'. However, we believe that a statistical approach is very principled and most appropriate here. Specifically we use the *p-values* of the correlation coefficients, computed using a two-sided t-test with confidence intervals based on the Fisher transformation, to determine statistical significance. Two nodes are considered connected only if the *p-value* of the corresponding correlation r is less than 1×10^{-10}, imposing a very high level of confidence in that relationship. This may seem like a stringent requirement but, as shown in Section 4, quite a large number of edges satisfy this criterion and are retained in the final network. The result of these construction and pruning procedures is a (weighted) simple graph, that is, one which does not contain self-loops or multiple edges between any pair of vertices.

4. DESCRIPTIVE ANALYSIS

Using the weighting and pruning methods described in Section 3, we construct a separate network for each of the seven climate variables. In this section, we present several examples of descriptive analysis enabled by climate networks. The results combine quantitative evaluation with qualitative interpretation within the climate domain.

4.1. Network Topology and Clustering

First, we compute several structural properties of each network, which provide clues about the dynamics of the climate system:

- Number of nodes and edges.
- Clustering coefficient (C)—indicative of the 'cliquishness', this measure is the mean of all clustering coefficients C_i in the network, computed for each node i as

$$C_i = \frac{2|e_{jk}|}{k_i(k_i - 1)}, \qquad (5)$$

  where e_{jk} is the set of all edges between first neighbors of i and k_i is the degree of i (the number of edges connecting to node i).
- 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 computed with the algorithm described in Ref. 17.
- Also included are the expected clustering coefficient and characteristic path length of a random graph with the same number of nodes and edges, estimated as

$$C_{\text{rand}} \approx \langle k \rangle / N, \qquad (6)$$

  and

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

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

Variable	Nodes	Edges	C	L	C_{rand}	L_{rand}
SST	1701	132 469	0.541	2.437	0.092	1.474
SLP	1701	175 786	0.629	2.547	0.122	1.395
GH	1701	249 322	0.673	2.436	0.172	1.310
PW	1701	50 835	0.582	4.281	0.035	1.819
RH	1700	25 375	0.559	4.063	0.018	2.190
HWS	1699	31 615	0.554	4.826	0.022	2.056
VWS	1701	71 458	0.342	2.306	0.049	1.679

respectively, where $\langle k \rangle$ is the average degree, and N is the number of nodes in the network [18].

These properties are summarized in Table 1. Note that the slight variation in the number of nodes is due to the fact that in two cases a small number of nodes have no edges connecting to them after the pruning process and hence are removed from the network.

Second, we perform clustering to extract spatial patterns from the data. The goal here is to identify homogeneous regions, as defined by similarity in the long-term climate variability; note that spatial proximity is *not* taken into consideration. Such regions, or *clusters*, are of interest both to reveal inherent structure within the climate system (for example evidence of spatial autocorrelation or teleconnections) as well as to use as potential climate indicators (see Section 5).

Several methods have been used for clustering climate data, for instance k-means [19] and a shared-nearest neighbor approach [4]. However, we note that this task can be accommodated directly within our framework. In network literature, the clustering process is also called *community detection* due to its origins in social network analysis [20], and a number of algorithms have been applied in various settings; examples include the discovery of functional modules in protein–protein interactions [21], characterization of transportation networks [22], and many more [23,24]; for an extensive survey of methods and selected applications, see Ref. 25. To our knowledge, we were the first to publish on the application of community detection to climate networks and analysis of the resulting structure [14], and we are only aware of one other very recently accepted work in this area [26].

Since it considers 'network distance', in contrast to the pair-wise distances used by traditional approaches, it should be suitable for partitioning such a spatiotemporal dynamical system. In choosing an appropriate algorithm for this study, three requirements guided our selection: (i) the ability to utilize edge weights, (ii) suitability for dense networks, and (iii) overall computational efficiency. The former two constraints are motivated by the physical properties of the networks themselves, namely, the inherent presence of a large number of edges with varying connection strengths. The first requirement, in particular, eliminates a large number of algorithms from consideration as they only work with unweighted networks. Thus, the results presented here were obtained with the *Walktrap* algorithm described in Ref. 27, which meets all the above criteria. Specifically, the algorithm is based on the intuition that a random walker will get trapped in a dense part of the network (cluster). The authors define a node distance metric based on a large number of short random walks, which is used to discern the overall cluster structure. We used the default parameter settings: walk lengths of $t = 4$ steps and global maximum modularity to determine the best partition. The resulting clusters are visualized in Fig. 3. Another benefit of this algorithm is its ability to determine the number of clusters from the data, which varies from 4 to 18 across the different variables.

4.2. Domain Interpretation of Climate Networks

In the following, we examine the network properties more closely and interpret them in the context of domain knowledge. We should note here that nodes with no remaining edges after pruning are removed from the network, which explains the minor variations for relative humidity and wind speed; but in general the number of nodes is fixed (Table 1). In contrast, the number of edges varies widely, ranging by nearly one order of magnitude. The fact that the two hydrological variables (precipitable water, relative humidity) and wind components have a smaller number of edges can be attributed to these generally being related to more localized activity, whereas temperature and pressure—especially over the oceans—participate in larger-scale phenomena. We see additional evidence of this interpretation reflected in the geographic properties of the networks in Section 4.3.

Moving to the clustering coefficient we find that, despite varying numbers of edges, most networks exhibit a high degree of clustering; vertical wind speed is the exception. Since we do not impose any spatial constraints on the networks, there is no guarantee that clusters will be geographically cohesive, but as shown in Fig. 3 this is often the case. In fact, we observe a relationship between clustering coefficient and the clusters we discovered, namely, the higher the value of C, the fewer clusters exist in the network and the more geographically coherent they are. Thus, while geopotential height with $C = 0.673$ only has four very uniform clusters, vertical wind speed with $C = 0.342$ has 15 clusters, many of which are spatially disjoint.

A closer look reveals that the clusters reflect many known relationships among climate variables, but also a

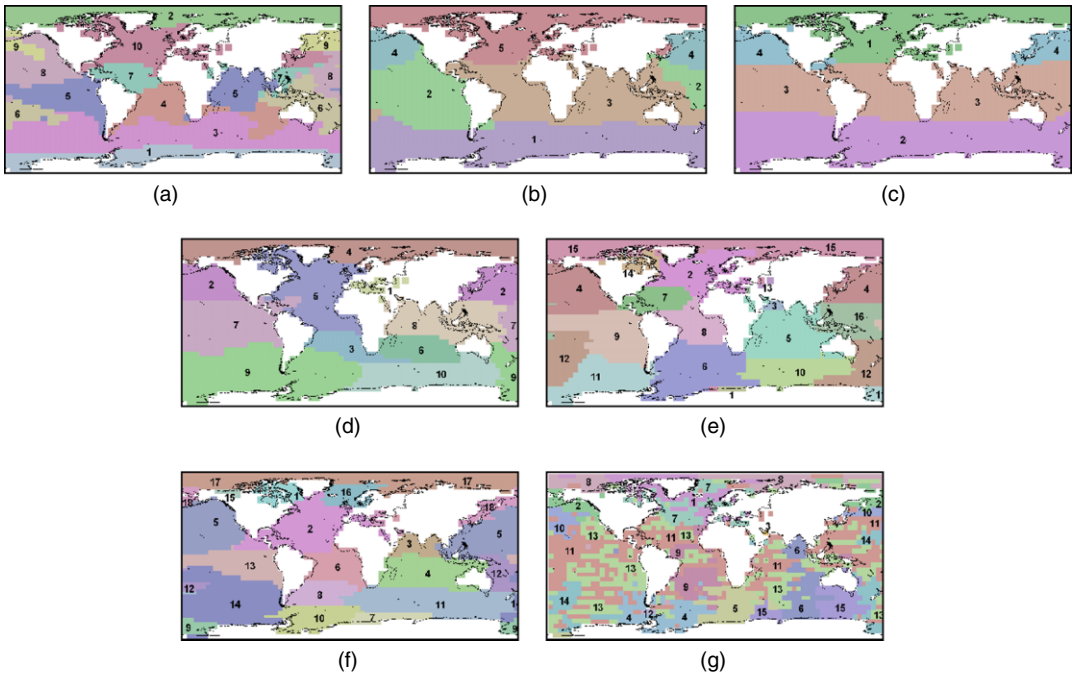

Fig. 3 Clusters obtained by applying community detection to climate networks (number of clusters in parentheses), numbers identify individual clusters (arbitrary assignment); best viewed in color. (a) Sea surface temperature (10); (b) sea level pressure (5); (c) geopotential height (4); (d) precipitable water (10); (e) relative humidity (16); (f) horizontal wind speed (18); (g) vertical wind speed (15)

few that are not as obvious and hence may be of interest to climate scientists. For example, cluster 5 in panel (a) of Fig. 3 seems to capture a well-known teleconnection between the El Niño Southern Oscillation and the Indian Ocean [28,29]. Moreover, panels (a), (d)–(f) of Fig. 3 all look remarkably similar. The close correspondence between sea surface temperature and precipitable water is explained by the Clausius–Clapeyron relation [30], which describes the increased water-holding capacity of air with increasing temperature. Relative humidity in turn is a function of temperature and atmospheric water content, explaining its relationship with both sea surface temperature and precipitable water. Lastly, the connection of wind speed with this group is not apparent, but a search of domain literature revealed that there is in fact a known relationship between surface winds and sea surface temperature [31]. Geopotential height depends on sea level pressure, which explains the similarity in their clusters. We also observe distinctive latitudinal bands, likely a result of the interplay between the wind belts that make up the global atmospheric circulation and these two pressure-related variables. Interestingly, it was shown in Ref. [3] that the tropics and extratropics consist of two separate networks with fundamentally different properties, which validates the presence of distinct communities in these regions. Finally, vertical wind speed looks unlike any other variable as it does not form geographically cohesive clusters. But this behavior is not surprising due to its involvement in convection, a highly localized activity that remains one of the most difficult atmospheric processes to model [32].

Characteristic path lengths range from 2.3 to 4.8, implying high overall connectivity. Comparing the clustering coefficients and characteristic path lengths to those expected for random graphs, we find that in all cases $C \gg C_{\text{rand}}$ and $L \geq L_{\text{rand}}$, satisfying the small-world properties [18] as may be expected in correlation-based networks [33]. In the context of climate, this suggests that there should be a relatively small number of distinct clusters, or climate regions, for each variable, which is precisely what we observed in Fig. 3.

4.3. Geographic Network Properties

While the above measures are commonly used to characterize all kinds of complex networks, a quantity called *area weighted connectivity* was proposed specifically to correct for the bias in the degree sequence induced by the angularly even spacing of grid points on the sphere [12]. If a node i is connected to k_i other nodes, then its connectivity $\tilde{C}_i$ is computed as

$$\tilde{C}_i = \sum_{j=1}^{k_i} \cos\lambda_j \Delta A \Big/ \sum_{\text{over all } \lambda \text{ and } \varphi} \cos\lambda \Delta A, \qquad (8)$$

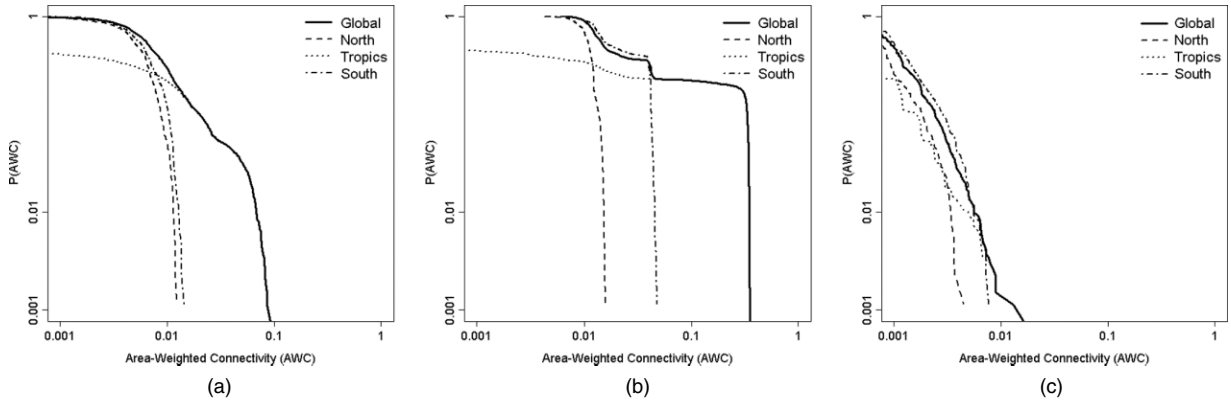

Fig. 4 Area weighted connectivity is an alternative network property for spatial data. (a) Sea surface temperature; (b) geopotential height; (c) vertical wind speed

where ΔA is the grid area, λ is the latitude and φ is the longitude [12]. In addition to the full network, we performed the same calculation for separate networks constructed from points only in the Northern (30°N–90°N), Tropical (30°S–30°N), and Southern (90°S–30S$^\circ$) regions. This quantity can be plotted against its observed probability on a log–log plot, similar to the degree distribution. Representative examples for three different variables are depicted in Fig. 4.

Panel (a) of Fig. 4 is based on sea surface temperature data very similar to Fig. 2 in Ref 34, and indeed the solid black lines agree quite closely. In their article, the authors assert that for most of the scales involved the distribution is a power law, but we would caution that this is only true for a limited subset in the center of the graph selected for display; when extending the axes to their full extent, this property no longer seems to hold. Moreover, for different variables we see quite different distributions. For example, for geopotential height more nodes are connected to a larger fraction of the globe—this is especially true for the tropical region, where one would expect relatively homogeneous pressure levels. In contrast, the connectivity of the network for vertical wind speed is consistently lower, supporting the notion that wind patterns are a localized activity.

For a more detailed understanding of the geographic network properties, we also examine the relative frequency distribution of link lengths for each variable. The great-circle distance (shortest path between two points on a sphere) was computed using Vincenty's formula [35], which is robust to the rounding errors of other formulae when performing the computations on a computer. Figure 5 shows the distributions for each variable. To aid in our interpretation, Panel (h) of Fig. 5 shows an annotated schematic of an idealized histogram which highlights two regions: a large number of short edges indicates that a variable has high spatial autocorrelation, a property embodied by Tobler's First Law of Geography, which states that 'everything is related to everything else, but near things are more related than distant things' [36]; a 'fat tail' suggests the presence of teleconnections within that same variable, that is, climatologically similar behavior in locations that are geographically separated (see also Ref 37). We should note here that all of the histograms show a lower value for very short distances. Though counterintuitive, these values are in fact correct: because the nodes are evenly spaced angularly on the sphere, grid points near the equator are much further apart (>500 km) than points near the poles, and hence will never appear in the first two bins.

Examining the individual panels of Fig. 5, we find that all but one of the variables peak at short distances, exhibiting a 'nearest neighbor' property reflective of the spatial autocorrelation one would expect in climate data. Vertical wind speed is the exception, which shows much lower correlation overall (panel (h) of Fig. 5) due to its association with very localized (sub-grid scale) convective activity [32], resulting in poor clustering (Panel (h) of Fig. 3). In contrast, variables in the middle row of Fig. 5 do exhibit some spatial autocorrelation but virtually no long-range connections. Thus, the corresponding networks resemble a mesh structure wherein nodes are connected to their closest neighbors, which, as shown in panels (d)–(f) of Fig. 3, has favorable implications for clustering. In fact, the two hydrological variables and wind speed show some similar patterns, for example tropical clusters stretching from the Western coast of South America (near Peru) all the way into South-East Asia (potentially related to the Intertropical Convergence Zone). Variables in the top row of Fig. 5 exhibit both the highest local connectivity as well as varying degrees of teleconnection strengths—properties associated with small-world behavior and reflective of the network topology seen in Section 4.1. For example, recall that the clusters for sea surface temperature visually appear similar to those of several other variables. However,

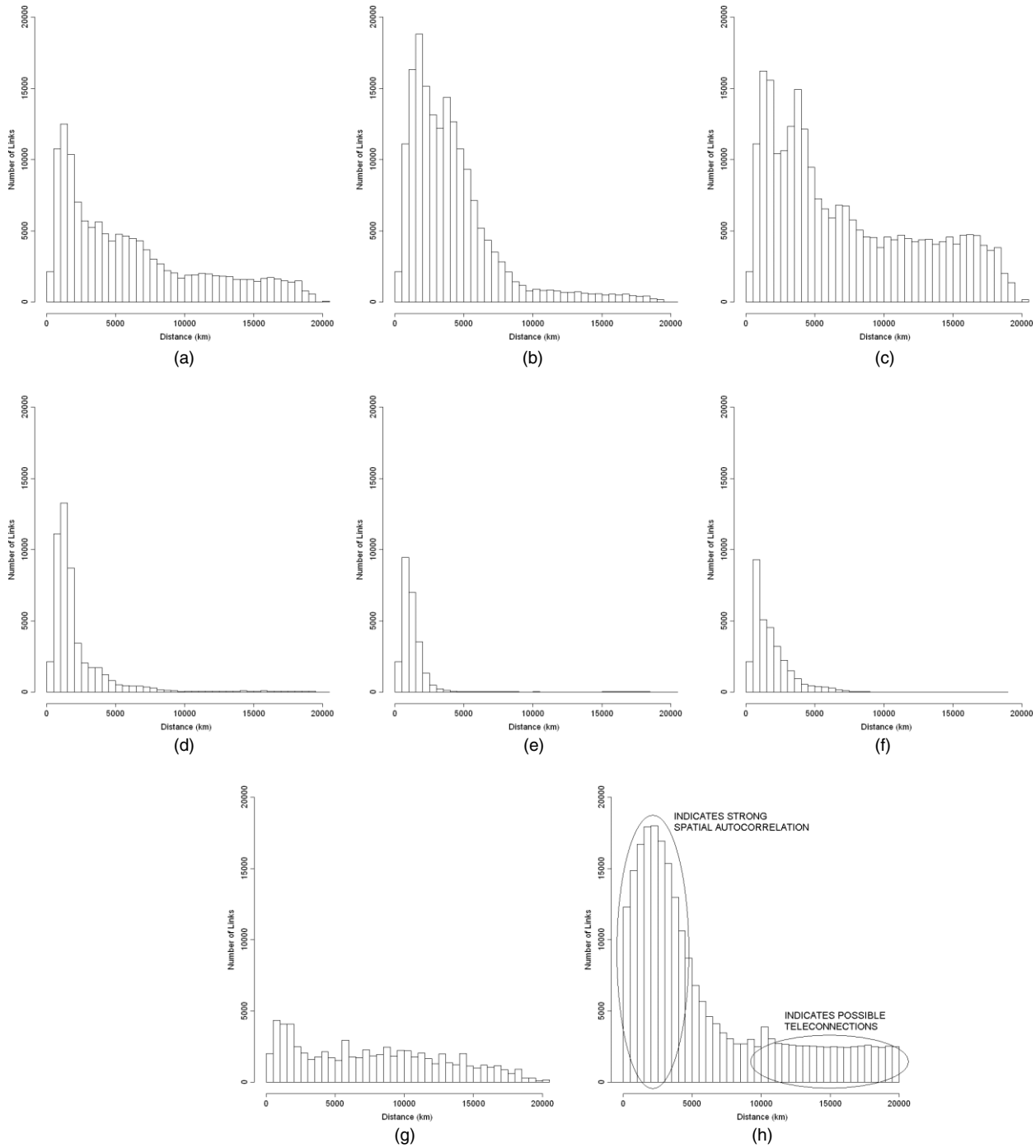

Fig. 5 Relative frequency distributions of link lengths. Variations in the distribution for different variables can be explained by their climatological interpretation. (a) Sea surface temperature; (b) sea level pressure; (c) geopotential height; (d) precipitable water; (e) relative humidity; (f) horizontal wind speed; (g) vertical wind speed; (h) ideal interpretation

communities 5, 7, and 10 (panel (a) of Fig. 3) each consist of two disjoint regions, which account for the 'fat tail' in the distribution.

In this section, we demonstrated that complex networks are a suitable representation for climate data as they provide a means for various types of *descriptive* analysis, giving rise to interesting domain insights. Building on the results obtained thus far—most notably the clusters—we now shift focus to the *predictive* aspect of our proposed framework.

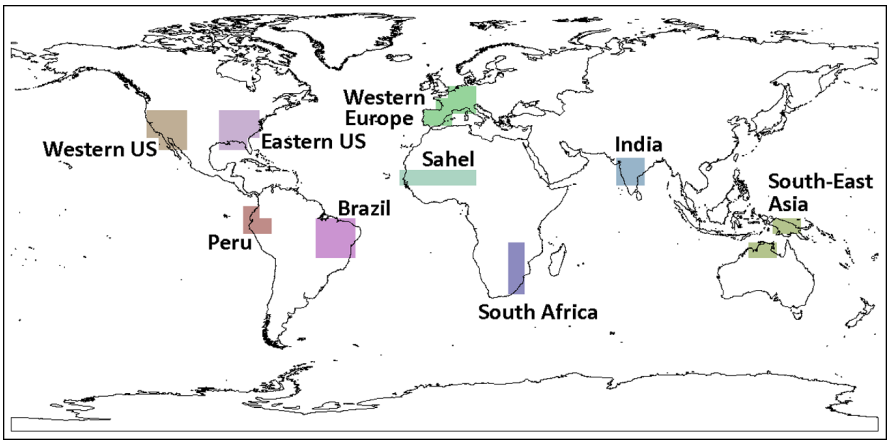

Fig. 6 Target regions for climate indices

5. PREDICTIVE MODELING

Here we consider one specific predictive task, namely the extraction of climate indices from observed data. A climate index can be defined as a 'diagnostic tool used to describe the state of the climate system and monitor climate' [38], that is, it summarizes climatic variability at local or regional scale into a single time series and relates these values to other events. One of the most studied indices is the southern oscillation index (SOI), which is strongly correlated with the El Niño phenomenon [39] and is predictive of climate in many parts of the world; see Ref. 10 for other examples. Thus, ocean dynamics are known to have strong influence over climate processes on land, but the nature of these relationships is not always well understood. This motivates the solution strategy for our predictive modeling task, which lies in leveraging the descriptive insights from the previous section for prediction. Specifically, we are able to re-use the clusters obtained in Section 4.1 as potential climate indices for predicting the behavior of variables on land. Other researchers have also reported on relationships between the properties of climate networks and the El Niño index, albeit in slightly different contexts [13,34,40]. Note that there may be other variables that contain additional predictive value, for example, auxiliary measurements over land (see, e.g., Refs. 41, 42). However, we focus specifically on information content in oceanic indices for land climate (teleconnections); identifying and evaluating these complimentary sources of additional predictive information is itself a nontrivial problem and thus beyond the scope of this article.

5.1. Experimental Setup

The upcoming report from the Intergovernmental Panel on Climate Change (expected in 2013) calls for greater attention to regional assessments, so we focus on prediction at regional scales. Accordingly, we selected nine target regions illustrated in Fig. 6. Some of these, like Peru and the Sahel, are known to have relationships with major climate indices; others were included to provide a representative set of regions around the world. Here we consider two climate variables in each region, air temperature and precipitation (also obtained from the NCEP/NCAR Reanalysis Project [8]), for a total of 18 response variables (nine regions $\times$ two variables). We chose these primarily for their relevance to human interests as they directly influence our health, environment, infrastructures, and other man-made systems.

Recall that we constructed a separate network for each of seven variables over the sea and applied community detection to obtain clusters (Fig. 3). The number of clusters is different for each variable as it depends on network properties, for a total of 78 clusters altogether. In the following, we outline the step-by-step procedure used to obtain our experimental results.

Step 1: For each network cluster, create a corresponding predictor by averaging over all grid points within the cluster.

Step 2: Similarly, for each target region, create two response variables by computing the average temperature and precipitation over all grid points.

Step 3: Divide the data into 50-year training set (1948–1997) and 10-year test set (1998–2007).

Step 4: For each of the 18 response variables, build a linear regression model on the training set and generate predictions for the test set.

While it is conceivable to use other learning algorithms in Step 4, we opted for linear regression as it maintains the interpretability of the model, which is important to domain scientists. Nonetheless, future work could explore alternate prediction algorithms in this context.

For comparison, we also cluster the data using k-means with Euclidean distance, a clustering technique named one of the top ten data mining algorithms [43] and standard method used in the climate domain (see, e.g., Refs. 19,44,45). Since determining an appropriate number of clusters k is itself a difficult problem, we perform a comprehensive test using three different settings: $k = 5$, $k = 10$, and k equal to the number of clusters k_n obtained using community detection to assure the fairest possible comparison (where k differs between variables). Steps 1–4 are then repeated using each of these alternate clusterings.

To quantify performance, we calculate root mean square error (RMSE) between predictions and actual (observed) data. Unlike simple correlation, which measures on covariance between two series, RMSE incorporates notions of both variance and estimator bias in a single metric.

5.2. Experimental Results

The RMSE scores for the prediction experiments are summarized in Table 2; the lowest (best) and highest (worst) scores in each row are shown in **bold** and *italics*, respectively. First, we note that the clusters extracted from climate networks consistently produce as comparable or better predictions than the clusters obtained from k-means. Moreover, no one setting of k seems to work best in all cases. Network clusters also have the lowest mean RMSE across both temperature and precipitation, affirming that they are effective in various predictive settings. To support this notion, we evaluate network-based clusters relative to the other methods using the Hochberg procedure of the Friedman test [46] at 95% confidence intervals—a nonparametric way to determine statistical significance of performance rankings across multiple experiments. The outcomes are included at the bottom of Table 2; a checkmark (✓) denotes that the network-based clusters are significantly better than the clusters in that column and, indeed, we find this to be true in all three cases.

Table 2. RMSE scores for predictions of temperature and precipitation using network clusters and k-means clusters for $k = 5$, $k = 10$, and $k = k_n$, the number of network clusters for each variable. The best (bold) and worst (italic) scores in each row are indicated. A checkmark (✓) at the bottom of a column denotes that the network-based clusters are significantly better according to the Friedman test of ranks at 95% confidence.

			k-Means		
	Region	Network clusters	$k = 5$	$k = 10$	$k = k_n$
Air temperature	SE Asia	**0.541**	0.629	0.694	*0.886*
	Brazil	0.534	*0.536*	0.532	**0.528**
	India	**0.649**	0.784	*1.052*	0.791
	Peru	**0.468**	0.564	*0.623*	0.615
	Sahel	**0.685**	0.752	0.750	*0.793*
	S Africa	0.726	**0.711**	*0.968*	0.734
	East US	0.815	0.824	*0.844*	**0.811**
	West US	**0.767**	0.805	0.782	*0.926*
	W Europe	0.936	*1.033*	**0.891**	0.915
	Mean	**0.680**	0.737	*0.793*	0.778
	StdDev	±0.150	±0.152	±*0.165*	**±0.135**
Precipitation	SE Asia	**0.665**	0.691	*0.700*	0.684
	Brazil	**0.509**	0.778	*0.842*	0.522
	India	**0.672**	0.813	0.823	*0.998*
	Peru	**0.864**	*1.199*	1.095	1.130
	Sahel	**0.533**	*0.869*	0.856	0.593
	S Africa	**0.697**	*0.706*	0.705	0.703
	East US	0.686	0.750	*0.808*	**0.685**
	West US	**0.605**	0.611	*0.648*	0.632
	W Europe	**0.450**	*0.584*	0.549	0.542
	Mean	**0.631**	0.778	*0.781*	0.721
	StdDev	**±0.124**	±0.182	±0.156	±*0.207*
	Friedman test ($\alpha = 0.05$)		✓	✓	✓

Predictive Power: Having established network-based clusters as the better candidate indices, we are faced with the question whether they offer any true predictive power. An answer must ascertain that the clusters contain useful information. We do this by comparing directly to two baseline approaches, namely by calculating the 'lift' over random predictions as well as a univariate predictor. Lift for region R and variable v is defined as percent improvement of the network-based predictions,

$$\text{Lift}(R, v) = \frac{\hat{y}_{\text{alt}} - \hat{y}_{\text{net}}}{\hat{y}_{\text{alt}}}, \tag{9}$$

where $\hat{y}_{\text{net}}$ is the RMSE obtained using network-based clusters and $\hat{y}_{\text{alt}}$ is the RMSE from the alternate method, that is, random or univariate prediction as explained below.

First, we perform a randomization experiment, wherein we scramble the order of the test data and recompute RMSE. More specifically, we randomly rearrange the time series for the 10-year test period and compute the RMSE to the network-based predictions. Results are shown in Table 3 under the heading 'Random'; each reported value represents an average taken over 10 000 runs with different random seeds. For temperature, we observe gains across all regions, with lifts ranging from 8 to 60%. For precipitation improvements are more modest, ranging from 5 to 26% in seven of the nine regions. The two exceptions are the Sahel and South Africa, where precipitation is generally low and anomalies are infrequent, making prediction more difficult. Nonetheless, these results suggest that the clusters do in fact have some predictive power.

Second, we compare the network-based predictions to a univariate predictor. Univariate here means that only the response variable itself is used to generate predictions. The procedure starts by building a linear regression model on

Table 3. RMSE scores and lift (% improvement) of network-based clusters over univariate and random predictions of temperature and precipitation. The best score in each row is indicated in bold.

		Network	Random		Univariate	
	Region	RMSE	RMSE	Lift (%)	RMSE	Lift (%)
Air temperature	SE Asia	**0.541**	0.791	32	0.621	13
	Brazil	**0.534**	0.834	36	0.659	19
	India	0.649	**0.634**	18	**0.634**	−2
	Peru	**0.468**	1.181	60	0.722	35
	Sahel	**0.685**	0.964	29	0.758	10
	S Africa	**0.726**	0.900	19	0.766	5
	East US	**0.815**	0.963	15	0.862	5
	West US	**0.767**	0.917	16	0.805	5
	W Europe	0.936	1.019	8	**0.917**	−2
	Mean	**0.680**	0.929		0.749	
	±StdDev	0.150	0.123		**0.102**	
Precipitation	SE Asia	**0.665**	0.809	18	0.732	9
	Brazil	**0.509**	0.575	12	0.787	35
	India	0.672	0.757	11	**0.638**	−5
	Peru	0.864	0.909	5	**0.841**	−3
	Sahel	0.533	0.521	−2	**0.520**	−2
	S Africa	0.697	0.710	2	**0.669**	−4
	East US	0.686	0.781	12	**0.665**	−3
	West US	**0.605**	0.720	16	0.649	7
	W Europe	**0.450**	0.604	26	0.492	8
	Mean	**0.631**	0.666		0.710	
	±StdDev	0.124	**0.113**		0.124	

the training set of the target series and making a prediction for the next month. The new value is then added to the training data and the process repeated until predictions are made for the entire testing period. This provides a measure of predictability in the time series itself, that is, absent any help from climate indices.

The results are shown in Table 3 under the heading 'Univariate'. We find that, for temperature, the network clusters improve scores by 5–35% in seven of the nine regions and thus definitely add predictive power. In contrast, for precipitation the answer is not so straightforward as scores actually decrease slightly in five of the nine regions, with low to moderate gains in the other four. We postulate that this discrepancy is largely a reflection of inherent differences between the two variables which make patterns in precipitation generally more difficult to model and predict. Still, it is apparent that network-based clusters add predictive power above and beyond the baseline.

Model Parsimony and Interpretability: We have shown convincingly the value of network-based clusters for prediction in climate, yet it is important to effectively communicate our results and make them accessible to climate scientists. One of the biggest challenges stems from the fact that many different clusters contribute to each index, while traditional indices are composed of at most a handful measurements. Therefore it is pertinent that we address the issue of model parsimony. Drawing on our experience in data mining, the intuitive solution is feature selection. We apply a filter-based technique called *correlation-based feature selection (CFS)* that finds subsets of features, which are highly correlated with the dependent variable but uncorrelated with each other [47]. Specifically, it uses a heuristic to simultaneously find the subset of predictors (clusters) that maximize predictability while minimizing the correlation among them. Consequently, the number of clusters may vary for each response variable. We repeat all predictive tasks after CFS and compute RMSE scores and lift as before; the results are shown in Table 4. We find that the number of selected clusters ranges from 6 to 18, a significant reduction from the original total of 78. Although not all components contribute equally, this number of clusters is much more manageable for inspection by a domain expert (see Section 5.3).

Furthermore, we observe another well-known advantage of feature selection: the parsimonious model generalizes better to the unseen data, resulting in lower RMSE scores in many cases (hence the negative lift values). So not only does this postprocessing step enhance interpretability of the regression models, but it has the potential to increase predictive power on held-out data at the same time. In the following, we reiterate why the methods and results of this

Table 4. RMSE scores and lift (% improvement) of all network-based clusters over 'selected' subsets of size k obtained using feature selection. The best score in each row is indicated in bold.

		Network	Selected		
	Region	RMSE	k	RMSE	Lift (%)
Air temperature	SE Asia	0.541	9	**0.539**	0
	Brazil	**0.534**	11	0.595	10
	India	0.649	17	**0.532**	−22
	Peru	**0.468**	9	0.524	11
	Sahel	0.685	15	**0.678**	−1
	S Africa	0.726	13	**0.690**	−5
	East US	0.815	9	**0.730**	−12
	West US	0.767	12	**0.764**	0
	W Europe	0.936	9	**0.855**	−9
	Mean	0.680		**0.656**	
	±StdDev	0.150		**0.117**	
Precipitation	SE Asia	0.665	10	**0.610**	−9
	Brazil	**0.509**	15	0.591	14
	India	0.672	12	**0.617**	−9
	Peru	0.864	18	**0.779**	−11
	Sahel	0.533	18	**0.496**	−7
	S Africa	0.697	8	**0.679**	−3
	East US	0.686	15	**0.639**	−7
	West US	**0.605**	18	**0.605**	0
	W Europe	0.450	6	**0.446**	−1
	Mean	0.631		**0.607**	
	±StdDev	0.124		**0.097**	

article are relevant to the climate domain and explain how they might be used to promote our current understanding.

5.3. Domain Interpretation: Advancing Climate Science

Due to space constraints we cannot investigate each individual clustering with respect to its climatological interpretation, but we present several illustrative examples. To this end, we perform a thought experiment that compares our findings with a baseline performance estimate one might expect based on domain knowledge alone. Specifically, based on expertise one would likely choose a much smaller subset of variables so we ask the question, *How do clusters obtained with our data mining method compare to those supported by domain knowledge?*

Temperature in Peru: First, we focus on the prediction of air temperature in Peru. We chose this example because it is closely related to the El Niño phenomenon, and therefore domain insights on climate predictability in this region are plentiful. The predictions for all 'selected' network clusters are shown in Fig. 7, along with the actual (observed) data. It is apparent that the predictive model works quite well here, capturing all the major variability. In fact, the RMSE score of 0.468 is among the lowest of any prediction task (Table 2).

The following nine clusters were selected within our framework: SST-5,6; GH-4; PW-7; HWS-1; VWS-1,11,12,14 (refer to Fig. 3). But a domain expert would likely select only three of these: SST-5, SST-6, and PW-7, partly because they are in close proximity to Peru, but also cover the El Niño regions. We repeat the regression using only these three clusters and obtain an RMSE of 0.552. This value is lower than any of the k-means clusters, showing that domain knowledge outperforms a naïve data mining method in this case. However, the network clusters perform better than domain insights alone, suggesting that there is additional predictive power to glean from the data. For example, based on our observations in Section 4.1 one would be unlikely to ever select any clusters based on VWS, but including them seems to improve predictive ability here. This poses an open question for domain scientists, namely whether there exist any convective patterns that may have some predictability.

Temperature in India: Second, we look at the prediction of air temperature in India (Fig. 8), because here the cluster selection provided the largest absolute improvement in performance. A total of 17 clusters were selected within our framework, namely: SST-4,5,7; GH-3; PW-1,6,7; RH-1,2,10,15; HWS-3; VWS-9,11,13,14,15 (refer to Fig. 3).

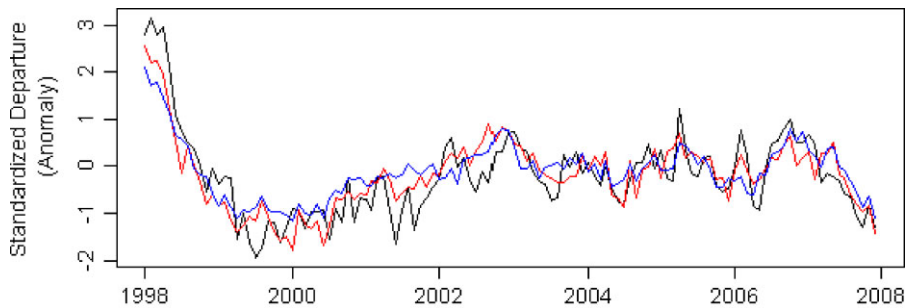

Fig. 7 Prediction of air temperature in Peru using all (red) and 'selected' (blue) network clusters compared to observations (black). Best viewed in color

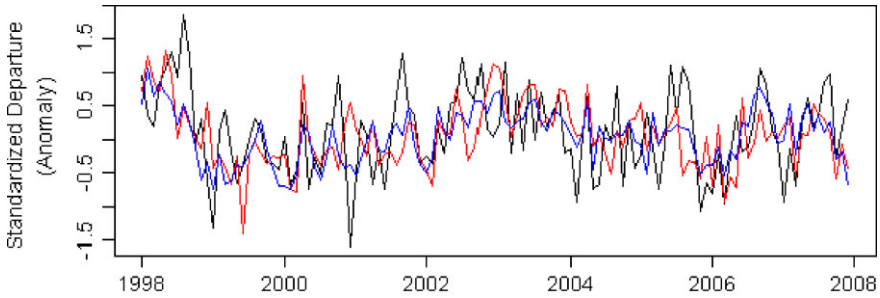

Fig. 8 Prediction of air temperature in India using all (red) and 'selected' (blue) network clusters compared to observations (black). Best viewed in color

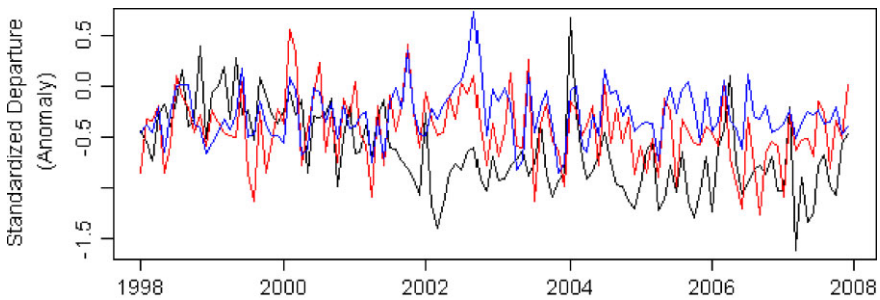

Fig. 9 Prediction of precipitation in Brazil using all (red) and 'selected' (blue) network clusters compared to observations (black). Best viewed in color

Of these, domain expertise would dictate that we select all three of the SST clusters because they surround other monsoon regions, GH-3, which captures the tropics in general, and HWS-3 due to spatial proximity. Repeating the regression using only these four clusters yields an RMSE of 0.572, which is again significantly lower than any of the k-means clusters. It is also lower than using all network-based clusters but the feature selection is able to improve upon it, supporting the use of this postprocessing step within our framework.

Precipitation in Brazil: Third, we consider one example of precipitation in Brazil (Fig. 9). Here 15 clusters were selected: SST-7,10; PW-3,9,10; RH-13; HWS-8,13,14; VWS-1,9,10,11,13,15. On the basis of domain knowledge, one would certainly select VWS-9 as well as PW-3 and PW-9, which relate to atmospheric water content and convective activity in the immediate area. In addition, one may also include SST-7 and HWS-8 for proximity as well as HWS-13 because of known climatic ties to the El Niño region. Using these six clusters we obtain an RMSE of 0.659. This value is higher than both k-means and network-based clusters, suggesting that domain knowledge cannot improve upon the indices extracted from data.

Discussion: The domain knowledge-based predictions were guided in their choice of predictors by the network clusters after feature selection. However, even this guided prediction performed worse than the best-performing data mining approach, suggesting that the networks succeed in capturing certain relationships that are lost on more naïve data mining methods, and which may be nonobvious or even nonintuitive to a domain expert. We conjecture that the ability of networks to represent more complex (e.g., many-to-many) relationships enhances the resulting clusters, which in turn constitute more powerful climate indices, but additional research is required to better understand the underlying processes. In two of the three case studies, domain knowledge improved predictive skills over k-means, and in one case over complex networks alone without the benefit of feature selection. However, for the limited number of examples studied here the network clusters obtained via feature selection achieve the highest overall predictability. As such, this work represents an important step toward knowledge discovery from climate data. We have confirmed well-understood insights (e.g., similarity in the spatial patterns of the clusters of temperature and precipitable water, as suggested by the Clausius–Clapeyron relation [30]), discovered nonintuitive hypotheses that were verified by domain science (e.g., relationship between temperature and surface winds [31]), and demonstrated the potential for discovering knowledge from data in a relatively hypothesis-free manner, which may advance climate science (e.g., possible long-range spatial associations in convective patterns and their potential predictive power on land climate).

6. BENEFITS OF THE UNIFIED FRAMEWORK

The network-based framework developed in this article provides a unique way for descriptive analysis and predictive modeling to inform each other. Specifically:

- We have demonstrated that climate networks offer a way to improve predictive skills by extracting predictors based on cluster attributes. Thus, in the context of climate, we have shown that network-based clusters yield informative predictors via statistical averages of the variables within each cluster.

- We argue that the predictive information content of the clusters, especially the ones that are nonobvious or nonintuitive, may lead to more robust descriptive analysis. We have shown that clusters which may not be intuitive to a domain expert can add predictive power beyond baseline approaches, which in turn would suggest that the clusters are not spurious.

7. FUTURE WORK

Further research is needed to investigate the following related areas: (i) combining variables (univariate versus multivariate) in conjunction with different functions for weighting edges, integrating multiple edge weights into a single network with consideration given to the interpretability of results, and the sensitivity of these approaches to the selection of pruning threshold; (ii) comparing the network communities to clusters obtained using other unsupervised learning techniques beyond k-means clustering; (iii) the ability to interpret improvements in the predictive skills of parsimonious representations of predictors derived from complex network clusters, which needs to account for both improvements over climate domain knowledge (e.g., known ocean-based climate indices [10]) as well as the value-add over other clustering approaches; iv incorporating temporal lags, both in the correlation measure used for network construction as well as in the predictive models; (v) the possibility of leveraging data-guided insights about climate variables obtained from observations to improve climate model projections in the future and reduce the associated uncertainty; (vi) extending the predictive modeling and descriptive analysis to be able to predict not just mean processes but climate extremes (e.g., significant change in regional climate patterns or the recurrence patterns of extreme events).

ACKNOWLEDGMENTS

The authors would like to thank Prof. Vipin Kumar for the engaging discussions. We also thank the editor and our four anonymous reviewers for their insightful comments and helpful suggestions. This research was supported in part by the National Science Foundation under Grants OCI-1029584 and BCS-0826958. The research was performed 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 nonexclusive, 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] H. von Storch and F. W. Zwiers, Statistical Analysis in Climate Research, Cambridge University Press, Cambridge, UK, 1995.
[2] W. D. Collins, The Community Climate system Model: CCSM3, J Clim 19(11) (2006), 2122–2143.
[3] A. A. Tsonis and P. J. Roebber, The architecture of the climate network, Physica A 333 (2004), 497–504.
[4] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter, Discovery of climate indices using clustering, In ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2003, 446–455.
[5] F. Lin, X. Jin, C. Hu, X. Gao, K. Xie, and X. Lei, Discovery of teleconnections using data mining technologies in global climate datasets, Data Sci J 6(Supplement) (2007), S749–S755.
[6] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, Science 328(5980) (2010), 876–878.
[7] Q. Schiermeier, The real holes in climate science, Nature 463 (2010), 284–287.
[8] E. Kalnay, The NCEP/NCAR 40-Year Reanalysis Project, BAMS 77(3) (1996), 437–470.
[9] http://www.cdc.noaa.gov/data/gridded/data.ncep.reanalysis.html [Last Accessed March 2010].
[10] http://www.cgd.ucar.edu/cas/catalog/climind/ [Last Accessed March 2010].
[11] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, The backbone of the climate network, Europhys Lett 87(4) (2009), 48007.
[12] A. A. Tsonis, K. L. Swanson, and P. J. Roebber, What do networks have to do with climate? BAMS 87(5) (2006), 585–595.
[13] K. Yamasaki, A. Gozolchiani, and S. Havlin, Climate networks around the globe are significantly affected by El Niño, Phys Rev Lett 100(22) (2008), 157–179.
[14] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, An exploration of climate data using complex networks, In ACM SIGKDD Workshop on Knowledge Discovery from Sensor Data, 2009, 23–31, subsequently published in ACM SIGKDD Explor 12(1) (2010), 2532.

[15] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, Complex networks in climate dynamics, Eur Phys J Special Topics 174 (2009), 157–179.
[16] A. Serrano, M. Boguna, and A. Vespignani, Extracting the multiscale backbone of complex weighted networks, PNAS 106(16) (2009), 8847–8852.
[17] R. W. Floyd, Algorithm 97: Shortest Path, Comm ACM 5(6) (1962), 345.
[18] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998), 440–442.
[19] R. G. Fovell and M.-Y. C. Fovell, Climate zones of the conterminous united states defined using cluster analysis, J Clim 6(11) (1993), 2103–2135.
[20] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, UK, 1994.
[21] S. Asur, D. Ucar, and S. Parthasarathy, An ensemble framework for clustering protein-protein interaction graphs, Bioinformatics 23(13) (2007), 29–40.
[22] R. Guimerá, S. Mossa, A. Turtschi, and L. A. N. Amaral, The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles, Proc Natl Acad Sci U S A 102(22) (2005), 7794–7799.
[23] M. E. Newman, Finding and evaluating community structure in networks, Phys Rev E 69 (2003), 026113.
[24] K. Steinhaeuser and N. V. Chawla, Identifying and evaluating community structure in complex networks, Pattern Recognition Lett 31(5) (2010), 413–421.
[25] S. Fortunato, Community detection in graphs, Phys Rep 486 (2010), 75–174.
[26] A. A. Tsonis, G. Wang, K. L. Swanson, F. A. Rodrigues, and L. da Fontura Costa, Community structure and dynamics in climate networks, Clim Dyn (2010) Online First, doi: 10.1007/s00382-010-0874-3.
[27] P. Pons and M. Latapy, Computing communities in large networks using random walks, J Graph Algebra Appl 10(2) (2006), 191–218.
[28] D. P. Chambers, B. D. Tapley, and R. H. Stewart, Anomalous warming in the Indian Ocean coincident with El Niño, J Geophys Res 104(C2) (1999), 3035–3047.
[29] K. K. Kumar, B. Rajagopalan, M. Hoerling, G. Bates, and M. Cane, Unraveling the mystery of Indian monsoon failure during El Niño, Science 314(5796) (2006), 115–119.
[30] R. R. Rogers and M. K. Yau, Short Course in Cloud Physics, (3rd ed.), Butterworth-Heinemann, Woburn, MA, 1989.
[31] L. W. O'Neill, D. B. Chelton, and S. K. Esbensen, Observations of SST-induced perturbations of the wind stress field over the southern ocean on seasonal timescales, J Clim 16(14) (2003), 2340–2354.
[32] K. A. Emanuel, Atmospheric Convection, Oxford University Press, New York, NY, 1994.
[33] S. Bialonski, M.-T. Horstmann, and K. Lehnertz, From brain to earth and climate systems: small-world interaction networks or not? Chaos 20 (2010), 013134.
[34] A. A. Tsonis and K. L. Swanson, Topology and Predictability of El Niño and La Niña Networks, Phys Rev Lett 100 (2008), 228502.
[35] T. Vincenty, Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations, Surv Rev 23(176) (1975), 88–93.
[36] W. R. Tobler, A computer movie simulating urban growth in the Detroit region, Econ Geogr 46(2) (1970), 234–240.
[37] A. A. Tsonis, K. L. Swanson, and G. Wang, On the role of atmospheric teleconnections in climate, J Clim 21(12) (2008), 2990–3001.
[38] http://cdiac.ornl.gov/climate/indices/indices_table.html [Last Accessed March 2010].
[39] C. F. Ropelewski and P. D. Jones, An extension of the Tahiti-Darwin southern oscillation index, Mon Weather Rev 115 (1987), 2161–2165.
[40] A. Gozolchiani, K. Yamasaki, O. Gazit, and S. Havlin, Pattern of climate network blinking links follows El Niño events, Europhys Lett 83(2) (2008), 28005.
[41] P. A. O'Gorman, and T. Schneider, The physical basis for increases in precipitation extremes in simulations of 21st-century climate change, Proc Natl Acad Sci U S A 106(35) (2009), 14773–14777.
[42] M. Sugiyama, H. Shiogama, and S. Emori, Precipitation extreme changes exceeding moisture content increases in MIROC and IPCC climate models, Proc Natl Acad Sci U S A 107(2) (2010), 571–575.
[43] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, et al. Top ten algorithms in data mining, Knowl Inf Syst 14(1) (2007), 1–37.
[44] H. F. Diaz, M. P. Hoerling, and J. K. Eischeid, ENSO variability, teleconnections and climate change, Int J Climatol 21 (2001), 1845–1862.
[45] F. M. Hoffman, W. W. Hargrove, and D. J. Erickson, Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models, Earth Interact 9(10) (2005), 1–27.
[46] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Mach Learn Res 7 (2006), 1–30.
[47] M. A. Hall and L. A. Smith, Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper, In International Florida AI Research Society Conference, 1999, 235–239.