Comparing Predictive Power in Climate Data: Clustering Matters

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
E-Mail: {ksteinha,nchawla}@nd.edu
[2] Computational Sciences and Engineering Division,
Oak Ridge National Laboratory, Oak Ridge TN 37831, USA
E-Mail: gangulyar@ornl.gov

Abstract. Various clustering methods have been applied to climate, ecological, and other environmental datasets, for example to define climate zones, automate land-use classification, and similar tasks. Measuring the "goodness" of such clusters is generally application-dependent and highly subjective, often requiring domain expertise and/or validation with field data (which can be costly or even impossible to acquire). Here we focus on one particular task: the extraction of ocean climate indices from observed climatological data. In this case, it is possible to quantify the relative performance of different methods. Specifically, we propose to extract indices with complex networks constructed from climate data, which have been shown to effectively capture the dynamical behavior of the global climate system, and compare their predictive power to candidate indices obtained using other popular clustering methods. Our results demonstrate that network-based clusters are statistically significantly better predictors of land climate than any other clustering method, which could lead to a deeper understanding of climate processes and complement physics-based climate models.

1 Introduction

Cluster analysis is an unsupervised data mining technique that divides data into subsets (called *clusters*) of elements that are – in some way – similar to each other [18]. As such, it is a versatile analysis tool that has been employed in a wide range of application settings including image segmentation [36], text and document analysis [27], and bioinformatics [6]. Clustering has also been applied for mining climate, ecological, and other environmental data. Examples include definition of ecoregions via multivariate clustering [16], automatic classification of land cover from remotely sensed data [21], and the definition of climate zones [10]; for a more complete survey see [15].

In the domain of climate data sciences, clustering is especially useful for discovery or validation of climate indices [28]. A *climate index* summarizes variability at local or regional scales into a single time series and relates these values to other events [38]. Let us consider one particular task: the extraction of ocean climate indices from historical data. For instance, one of the most studied indices is the Southern Oscillation Index (SOI), which is strongly correlated with the El Niño phenomenon and is predictive of climate in many parts of the world [25]; see [40] for other examples. Thus, ocean dynamics are known to have a strong influence over climate processes on land, but the nature of these relationships is not always well understood.

In fact, many climate indices – including the SOI – were discovered through observation, then developed more formally with hypothesis-guided analysis of data. However, given the increasing availability of extensive datasets, climate indices can also be extracted in a data-driven fashion using clustering [28, 24]. This approach presents a unique set of challenges including data representation, selection of a clustering method, and evaluation. Because it is such a difficult problem, climate scientists often resort to relatively simple algorithms such as k-means [10, 21].

For example, as anecdotal evidence of these challenges, Loveland et al. reported that in developing a global land cover dataset, "The number of clusters created for each continent was based on the collective judgment of the project team" [21]. This and similar accounts therefore beg the fundamental question, *What is the best clustering method for climate datasets?* In response we posit that deriving clusters from complex networks, which have been shown to capture the dynamical behavior of the climate system [8, 29, 31, 33, 37], may be an effective approach. This raises the issue of evaluation and validity of discovered clusters as climate indices. As typical of any clustering task, evaluation is highly subjective, relying on the judgment of a domain expert as field data for validation can be costly or even impossible to acquire. Instead of evaluating clusters directly, however, one can measure performance in terms of an external criterion, i.e., their predictive power.

Contributions We combine these two challenges – "choice of clustering" and "evaluation" – in a comprehensive comparative study within the context of climate indices. This paper expands upon the general methodology outlined in our prior work [29, 30] but focuses on comparing different clustering algorithms as well as evaluating different regression algorithms for their predictability on climate indices. Specifically, the contributions of the present work can be summarized as follows. We extract ocean climate indices from historical data using traditional clustering methods in addition to network-based clusters (Sections 3 & 4). We then generate predictive models for land climate at representative target regions around the globe using the clusters as predictors (Section 5), using the same process as before [29]. We compare the clustering methods based on their ability to predict climate variability and demonstrate that the network-based indices have significantly more predictive power (Section 6). Finally, we provide domain interpretations of the clustering results for a selected case study to illustrate the potential value of data mining methodologies for climate science (Section 7).

To our knowledge, this is the first study to systematically address the problem of clustering climate data for the purpose of discovering climate indices.

In particular, we cluster a large corpus of ocean climate data using various popular clustering algorithms, in addition to the clusters obtained from complex networks constructed with these data. Each set of clusters then serves as input to a predictive model for land climate of the general form $f : \mathbf{x} \rightarrow y$, where $\mathbf{x}$ represents a set of cluster centroids and y is given by one of two climate variables (temperature and precipitation) at nine target regions around the globe.

Our experimental results demonstrate that the network-based clusters are statistically significantly better predictors (climate indices) than clusters obtained using traditional clustering methods. In comparing different regression algorithms, we also note that more complex methods do not necessarily improve performance for this particular predictive task.

2 Climate Data

In the following, we briefly describe the characteristics of the dataset used in our analysis as well as the pre-processing steps required for the purpose of discovering climate indices [29, 30].

2.1 Dataset Description

The climate data stems from the NCEP/NCAR Reanalysis Project [19] (available at [39]). This dataset is constructed by assimilating remote and in-situ sensor measurements from around the world and is widely recognized as one of the best available proxies for global observations (it is obviously impossible to obtain exact data for the entire globe).

Although most climate indices are defined for temperature and/or pressure-related variables, we did not want to constrain ourselves by an *a priori* selection of variables. In fact, one question of interest was whether other variables also have predictive power, and hence we consider a wide range of surface and atmospheric measurements. Specifically, we include the following seven variables (abbreviation, 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 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* (HWS, measured in the plane near the surface), and *vertical wind speed* (VWS, measured in the atmospheric column).

This line of research (including [29, 30]) is the first to use such a wide range of variables in climate networks studies. The data are available as monthly averages for a period of 60 years (1948-2007), for a total of 720 data points. Measurements are sampled at points (grid cells) on a $5° \times 5°$ latitude-longitude spherical grid. A schematic diagram of the data for a single time step t_i is shown in Fig. 1.

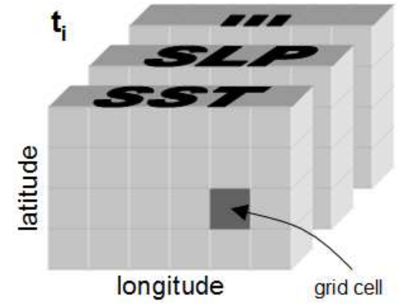


Fig. 1. Schematic depiction of gridded climate data for multiple variables at a single timestep t_i in the rectangular plane.

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. This can be problematic for prediction, and indeed climate indices are generally defined by the *anomaly series*, that is, departure from the "usual" behavior rather than the actual values. We follow precedent of related work [28–30, 32] and remove the seasonal component by monthly z-score transformation and de-trending.

At each grid point, we calculate for each month $m = \{1, ..., 12\}$ (i.e., separately for all Januaries, Februaries, etc.) the mean

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

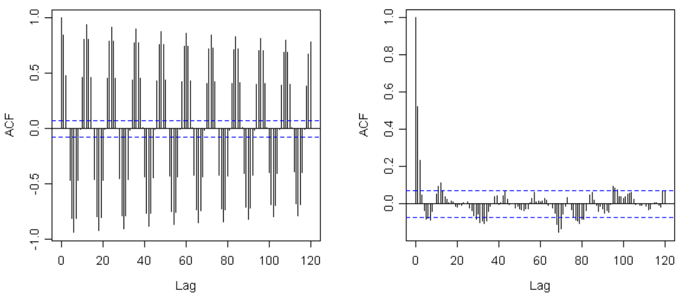
and standard deviation

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

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

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

The result of this process is illustrated in Fig. 2(b), which shows that de-seasonalized series has significantly lower autocorrelation than the raw data. In addition, we de-trend the data by fitting a linear regression model and retaining only residuals. For the remainder of this paper, all data used in experiments or discussed hereafter have been de-seasonalized and de-trended as just described.



(a) Raw Data (b) De-Seasonalized

Fig. 2. The de-seasonlized data (b) exhibits significantly lower autocorrelation than the raw data (a).

2.3 Data Representation

In this paper, we employ two distinct representations for the different analysis methods. The first is used strictly to construct climate networks; it considers each grid point as a network vertex and the corresponding data as a time series (Section 3). The second is used for the traditional clustering methods (Sections 4.2-4.5) and consists of a flat-file format, wherein each grid point is considered as an instance (row) and each time step as an attribute (column); the temporal nature of the data as well as certain aspects of the relationships between grid points is lost.

3 Climate Networks

The intuition behind this methodology is that the dynamics in the global climate system can be captured by a complex network [8, 29, 33]. Vertices represent spatial grid points, and weighted edges are created based on statistical relationships between the corresponding pairs of time series.

3.1 Estimating Link Strength

When dealing with anomaly series we need not consider the mean behavior, only deviations from it. Therefore, Pearson correlation (r) is a logical choice as measure of the edge strength [28–30, 32], computed for two series A and B of length t as

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

where a_i is the i^{th} value in A and $\bar{a}$ is the mean of all values in the series. Note that r 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.

We should point out here that nonlinear relationships are known to exist within climate, which might suggest the use of a nonlinear correlation measure. However, Donges et al. [8] examined precisely this question and concluded that, "the observed similarity of Pearson correlation and mutual information networks can be considered statistically significant." Thus it is sensible to use the simplest possible measure, namely (linear) Pearson correlation.

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 pruning is desirable. Since there is no universally optimal threshold [26], we must rely on some other criterion. For example, Tsonis and Roebber [32] opt for a threshold of $r \geq 0.5$ while Donges et al. [8] use a fixed edge density ρ to compare networks, noting that "the problem of selecting the exactly right threshold is not as severe as might be thought."

We believe that a significance-based approach is more principled and thus appropriate here. Specifically, we use the *p-value* of the correlation to determine statistical significance. Two vertices 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 quite a large number of edges satisfy this criterion and are retained in the final network.

In [29], we examined the topological and geographic properties of these networks in some detail. Suffice it to say here that for all variables the networks have a high average clustering coefficient and a relatively short characteristic path length, suggesting that there is indeed some community structure; more on this in the following section.

4 Clustering Methods

In this section we provide succinct descriptions of the clustering methods used in this comparative study; for algorithms we defer the reader to the original works, as cited. Sections 4.1 & 4.2 were developed in [29] but are included for completeness. Note that climate networks employ a network-based data representation whereas traditional clustering methods use a flat-file representation of time series at each grid cell, as described in Section 2.3.

4.1 Network Communities

This method is based on the climate networks described in Section 3. There exists a rich body of literature on the theory and applications of clustering in networks, also called *community detection* due to its origins in social network analysis [34]; other examples include discovery of functional modules in protein-protein interactions [6], characterization of transportation networks [13], and many more. However, to our knowledge we are the first to apply community detection in climate networks [29].

In choosing an appropriate algorithm for this study, three constraints guided our selection: *(i)* the ability to utilize edge weights, *(ii)* suitability for relatively dense networks, and *(iii)* overall computational efficiency. The first requirement in particular eliminates a large number of algorithms from consideration as they only work with unweighted networks. Thus, all results presented here were obtained with the algorithm described in [23] using the default parameter settings, which meets all the above criteria (tests with other algorithms produced comparable results). A fringe benefit of this algorithm is an option to determine the number of clusters from the data.

4.2 K-Means Clustering

The k-means algorithm is one of the oldest and well-known methods for cluster analysis, and several refinements have been proposed over the years. We use the implementation described in [17]. Its fundamental aim is to partition the data into k distinct clusters such that each observation belongs to the cluster with the "closest" mean, where closeness is measured by the Euclidean distance function. Due to its simplicity the algorithm is popular and enjoys widespread use, but it also suffers from drawbacks including the need to specify the number of clusters k *a priori* as well as sensitivity to noise, outliers, and initial conditions (mitigated by running the algorithm multiple times).

4.3 K-Medoids Clustering

This algorithm is a variation on k-means clustering in that it also seeks to partition the data into k clusters by minimizing the distance to the cluster centers, except that the data points themselves are chosen as centers (called *medoids*). We use an implementation known as Partitioning Around Medoids (PAM) [20]. It is subject to some of the same problems as k-means but is more robust to outliers and noise in the data.

4.4 Spectral Clustering

This term refers to a class of clustering techniques that utilize the eigenvalues of a similarity matrix constructed from the data (called the *spectrum*, hence the name) for dimensionality reduction and then find clusters in the lower-dimensional space. The method used to compute the similarity matrix is also referred to as kernel function. Data can be partitioned either into two parts – recursively if necessary – or directly into k subsets. We use the algorithm described in [22], which utilizes multi-way partitioning and was shown to yield good results on a wide variety of challenging clustering problems.

4.5 Expectation Maximization

An expectation-maximization (EM) algorithm is a general technique for finding maximum likelihood parameter estimates in statistical models, and cluster analysis is one of its most common applications. In general, EM methods are computationally expensive but work well in a variety of application settings. We use the algorithm described in [11], which implements EM for a parameterized mixture of k Gaussians and is reasonably efficient.

5 Experimental Setup

This section explains how the various algorithms are used to obtain potential climate indices and how we compare them in a predictive setting.

5.1 Extracting Candidate Indices

Recall the definition of a climate index, that is, a summary of climate variability over one or more ocean regions, which is related to climate on land. Our first task is to extract potential indices from historical data using clustering. We run each algorithm described in Sec. 4 on all data corresponding to ocean grid points. With the exception of the network-based approach, the number of clusters k must be specified *a priori*. Therefore, we perform a comprehensive set of experiments by running each algorithm for $k = 5$, $k = 10$, and k equal to the number of clusters k_n obtained using community detection in networks to assure the fairest possible comparison (78 clusters total where k differs between variables).

5.2 Evaluating Predictive Power

The upcoming report from the Intergovernmental Panel on Climate Change (expected 2013) calls for attention to regional assessments of climate, so we focus on prediction at regional scale. We use nine target regions covering every continent, illustrated in Figure 3 (consistent with [29] for comparison). Some, like Peru and the Sahel, have known relationships with major climate indices; others were included to provide a representative set of regions around the world.

Moreover, we consider two climate variables in each region, temperature and precipitation, for a total of 18 response variables (9 regions × 2 variables). We chose these variables primarily for their relevance to human interests: they directly influence our health and well-being as well as our environment, infrastructures, and other man-made systems. Precipitation obtained from reanalysis has potential issues but is used here to develop initial insights and to compare with

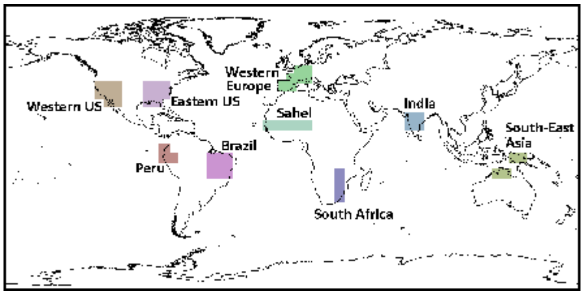


Fig. 3. Target regions for climate indices.

temperature, which is considered more reliable. In the following, we outline the step-by-step procedure used for the predictive modeling [29]:

1. For each of the algorithm-parameter combinations described above, create a corresponding set of predictors ($\mathbf{x}$) consisting of the cluster centroids by averaging the time series for all grid points assigned to each cluster.
2. Similarly, for each target region, create two response variables (y) by computing average temperature / precipitation over all grid points in the region.
3. Divide the data into a 50-year training set (1948-1997) and a 10-year test set (1998-2007).
4. For each of the 18 response variables, build a regression model $f : \mathbf{x} \rightarrow y$ on the training data and generate predictions for the unseen test data using each set of predictors from Step 1 in turn.

While it is conceivable to use any number of machine learning algorithms in Step 4, we start with linear regression in Section 6.1 as it gives us a performance baseline while maintaining interpretability of the model, which is important to domain scientists. In Section 6.3 we then go on to explore alternate prediction algorithms in this context.

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

6 Experimental Results

In this section, we present our empirical comparison of clustering algorithms and evaluate the predictive power of the derived climate indices.

6.1 Comparing Clustering Algorithms

First we seek to answer the question, *Which clustering method produces the best candidate indices?* The RMSE scores for all prediction tasks are summarized in

		Networks	**K-Means**			**K-Medoids**			**Spectral**			**Expectation-Max.**		
	Region	k_n	$k = 5$	$k = 10$	$k = k_n$	$k = 5$	$k = 10$	$k = k_n$	$k = 5$	$k = 10$	$k = k_n$	$k = 5$	$k = 10$	$k = k_n$
Air Temperature	SE Asia	**0.541**	0.629	0.694	0.886	0.826	0.973	*1.009*	0.751	0.731	0.827	0.760	0.653	0.703
	Brazil	0.534	0.536	0.532	0.528	**0.509**	0.512	0.539	0.579	0.519	*0.582*	0.577	0.553	0.562
	India	0.649	0.784	*1.052*	0.791	0.685	**0.587**	0.627	0.597	0.698	0.618	0.612	0.977	0.796
	Peru	**0.468**	0.564	0.623	0.615	0.524	0.578	0.510	0.716	0.585	0.676	0.685	*0.837*	0.793
	Sahel	0.685	0.752	0.750	0.793	**0.672**	0.733	0.752	0.866	0.801	0.697	0.820	*0.969*	0.854
	S Africa	0.726	0.711	*0.968*	0.734	**0.674**	0.813	0.789	0.692	0.857	0.690	0.761	0.782	0.892
	East US	0.815	0.824	0.844	0.811	*0.908*	**0.742**	0.798	0.848	0.839	0.799	0.768	0.753	0.846
	West US	0.767	0.805	0.782	0.926	0.784	*1.021*	**0.744**	0.777	0.810	0.755	0.780	0.811	0.766
	W Europe	0.936	1.033	0.891	0.915	0.950	1.071	*1.116*	**0.868**	0.898	0.962	0.947	0.975	0.986
	Mean	**0.680**	0.737	0.793	0.778	0.726	0.781	0.765	0.744	0.749	0.734	0.746	*0.812*	0.800
	±StdDev	0.150	0.152	0.165	0.135	0.155	*0.204*	0.200	**0.109**	0.128	0.117	0.111	0.148	0.120
Precipitation	SE Asia	**0.665**	0.691	0.700	0.684	0.694	0.695	0.699	0.727	0.673	0.706	*0.739*	0.736	0.719
	Brazil	**0.509**	0.778	0.842	0.522	0.817	0.986	1.110	1.272	*1.549*	1.172	1.353	1.351	1.284
	India	0.672	0.813	0.823	0.998	0.798	1.072	*1.145*	**0.654**	1.018	0.820	0.714	0.820	0.720
	Peru	0.864	1.199	1.095	1.130	1.064	0.934	*1.227*	0.859	0.994	**0.836**	0.872	0.845	0.837
	Sahel	**0.533**	0.869	0.856	0.593	1.043	0.847	0.648	0.838	*1.115*	0.846	0.804	1.047	0.963
	S Africa	0.697	0.706	0.705	0.703	0.702	0.706	*0.772*	0.704	**0.655**	0.683	0.753	0.729	0.736
	East US	0.686	0.750	0.808	0.685	0.814	0.679	*0.851*	0.686	0.789	0.711	**0.645**	0.685	0.745
	West US	0.605	0.611	0.648	0.632	0.617	0.610	0.599	**0.587**	0.635	0.645	0.599	0.646	*0.656*
	W Europe	**0.450**	0.584	0.549	0.542	*0.720*	0.581	0.605	0.545	0.632	0.493	0.532	0.651	0.673
	Mean	**0.631**	0.778	0.781	0.721	0.808	0.790	0.851	0.764	*0.896*	0.768	0.779	0.835	0.815
	±StdDev	**0.124**	0.182	0.156	0.207	0.154	0.175	0.247	0.217	*0.307*	0.188	0.239	0.230	0.199
Friedman Test ($\alpha = 0.05$)			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. Comparison of clustering methods: RMSE scores for predictions of temperature and precipitation using candidate indices obtained via community detection in networks as well as k-means, k-medoids, spectral, and expectation-maximization clustering 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.

Table 1; the lowest (best) and highest (worst) score in each row is shown in **bold** and *italic*, respectively.

Examining the results in some detail, we note that network-based clusters achieve the best score on 6 of 18 prediction tasks, more than any other algorithm-parameter combination. This is a first indication that climate networks may be well-suited for the discovery of climate indices. Network clusters also have the lowest mean RMSE across both temperature and precipitation, affirming that they are effective for diverse predictive tasks; even in cases where networks are not the outright best option they seem to offer competitive performance. To support this notion, we evaluate network-based clusters relative to the other methods using the Hochberg procedure of the Friedman test [7] at 95% confidence intervals – a non-parametric way to determine statistical significance of performance rankings across multiple experiments.

The outcomes are included at the bottom of Table 1; a checkmark (✓) denotes that the network-based clusters are significantly better than the clusters in that column. Indeed we find this to be unequivocally the case, suggesting that networks capture the complex relationships in climate quite well. It is worth noting that clusters obtained with k-medoids and spectral clustering achieve scores comparable to, or even better than, the much more complex expectation-maximization algorithm. Regardless, we are led to conclude that community detection in climate networks yields the best candidate indices.

6.2 Validating Predictive Skill

Now that we established network-based clusters as having the highest predictability we must address the question, *Do these indices offer any true predictive power?* The answer was provided in [29]. To ascertain that the network clusters indeed contain useful information, we showed that they provide “lift” over random predictions as well as a simple univariate predictor. Our experimental results demonstrated that the network clusters do in fact have some predictive power, improving on the baseline by as much as 35%. Moreover, we can further enhance performance through feature selection [29].

6.3 Prediction Algorithms

Knowing that there is predictive information in the ocean clusters begs yet another question, namely, *What type of model can best harness this predictive power?* As alluded to in Section 5, linear regression merely provided a baseline comparison; it is entirely possible that other machine learning algorithms are better suited for modeling the processes connecting ocean and land climatology, for example, more sophisticated regressors that are able to capture nonlinear relationships.

Therefore, we also compare several fundamentally different prediction algorithms. In particular, we include neural networks (NN), regression trees (RTree) and support vector regression (SVR). The RMSE scores for the corresponding

	Region	LR	NN	RTree	SVR
Air Temperature	SE Asia	**0.541**	0.629	0.743	**0.541**
	Brazil	**0.534**	0.568	0.686	0.570
	India	0.649	0.646	0.704	**0.595**
	Peru	0.468	**0.459**	0.616	0.589
	Sahel	0.685	0.866	0.983	**0.662**
	S Africa	0.726	0.838	0.849	**0.714**
	East US	0.815	0.895	1.060	**0.773**
	West US	0.767	0.835	0.860	**0.755**
	W Europe	0.936	1.018	0.014	**0.890**
	Mean	0.680	0.750	0.835	**0.677**
	±StdDev	0.150	0.182	0.159	**0.116**
Precipitation	SE Asia	0.665	0.703	0.791	0.653
	Brazil	**0.509**	0.547	0.771	0.597
	India	0.672	0.809	1.045	**0.646**
	Peru	0.864	1.006	0.960	**0.842**
	Sahel	**0.533**	0.785	0.663	0.542
	S Africa	0.697	0.787	0.767	**0.684**
	East US	0.686	0.684	0.771	**0.649**
	West US	0.605	0.647	0.696	**0.603**
	W Europe	0.450	0.522	0.569	**0.448**
	Mean	0.631	0.721	0.782	**0.629**
	±StdDev	0.124	0.148	0.145	**0.107**
Friedman ($\alpha = 0.05$)			✓	✓	

Table 2. RMSE scores for predictions with network clusters using linear regression (LR), neural networks (NN), regression trees (RTree) and support vector regression (SVR). The best score in each row is indicated in **bold**.

prediction tasks are summarized in Table 2; the lowest (best) score in each row is shown in **bold**.

Most obviously, we find that support vector regression achieves the lowest RMSE in 13 of 15 cases (including one tie), while linear regression comes in close second; the actual scores of these two methods are generally quite close. In contrast, neural networks and regression trees perform notably worse. These observations are confirmed by the statistical significance test (Hochberg procedure of the Friedman test [7] at 95% confidence) included at the bottom of Table 2; a checkmark (✓) denotes that linear regression performs significantly better than the algorithm in that column. Note: repeating the significance test relative to the SVR scores does not change the results, i.e., they are significantly better than NN and RTree, but *not* LR.

It is prudent not to draw any general conclusions from these results, but empirical evidence suggests that the more complex regression models *do not* necessarily improve performance. We conjecture that the reason for this is a combination of high-frequency noise and a relatively small number of training samples in the data, which collectively can lead to overfitting with more complex modes. Thus, for the sake of computational efficiency as well as interpretability of results, it is advisable to use a linear regression model.

However, given larger datasets or slightly different prediction tasks, it is possible that the gains from alternate prediction algorithms – including but not limited to those compared in this paper – would indeed be more substantial.

7 Domain Interpretation

Due to space constraints we cannot examine every set of clusters with respect to its climatological interpretation, but we present one case study using Peru as illustrative example (focused only on prediction, a descriptive analysis is provided in our prior work [29]).

Air Temperature in Peru. We chose this region because it is related to the El Niño phenomenon and hence domain knowledge in this area is plentiful. The predictions for temperature using all and “selected” [29] network clusters are shown in Figure 4, along with the actual (observed) data. It is apparent that the predictive model works quite well here, capturing all major variations. In fact, the RMSE score of 0.468 is among the lowest of any prediction task (Table 1).

Examining the nine “selected” clusters in more detail, we find that this particular index is composed of the following variables: 2 SST, 1 GH, 1 PW, 1 HWS and 4 VWS. For reference, the clusters of SST and VWS are depicted in Figure 5. It comes as no surprise that the selected clusters of sea surface temperature are numbers 5 (containing the areas that define several prominent El Niño indices) and 6 (the equatorial Pacific stretching into South-East Asia). However, VWS clusters 1, 11, 12 and 14 are also included. This is curious as vertical wind speed – convective activity over the oceans – is not thought to have any predictive power in climate, yet our findings seem to suggest otherwise.

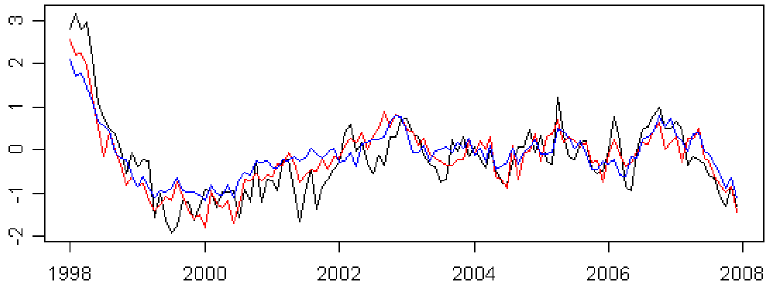
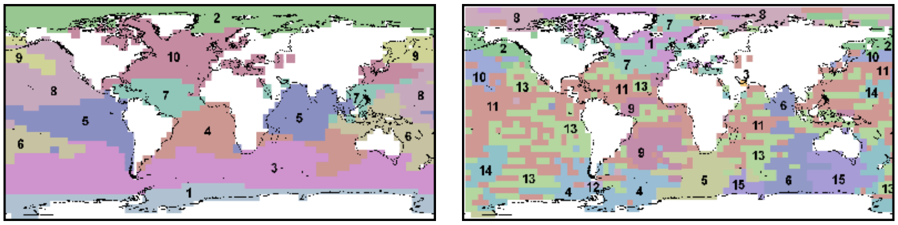


Fig. 4. Prediction of air temperature in Peru with all (red) and “selected” (blue) network clusters compared to observations (black). Best viewed in color.



(a) SST clusters

(b) VWS clusters

Fig. 5. Depictions of sample clusters (reproduced from [29]). Best viewed in color.

We contemplate this possibility with a thought experiment: *How do clusters obtained with our data mining approach compare to those supported by domain knowledge?*

To answer this question, we asked a domain expert to narrow down the clusters to only those intuitively expected to be of relevance. SST-5 and SST-6 were chosen based on known relationships, as well as PW-7 due to spatial proximity. We repeat the regression with using only these three clusters and obtain an RMSE of 0.552. This score is lower than most of the methods included in our comparison (Table 1), meaning that traditional clustering methods cannot match current domain knowledge. But *network-based clusters* significantly improve the score, suggesting that climate networks glean additional predictive power from the data. Whether and to what extent this holds in other situations remains an open question for climate scientists.

Precipitation in Peru. The predictions for precipitation using all and "selected" [29] network clusters are shown in Figure 6, along with the actual (observed) data. In contrast to temperature, the RMSE score of 0.864 for this region is on the low end across all prediction tasks (Table 1). While the very low-frequency signal is predicted to some degree, the observed data has much more variability not captured by the model. This is generally true for precipitation, namely, that

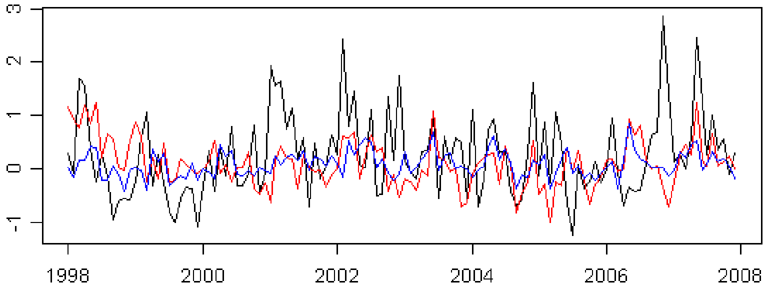


Fig. 6. Prediction of precipitation in Peru with all (red) and "selected" (blue) network clusters compared to observations (black). Best viewed in color.

the mean behavior is represented reasonably well while the model fails to predict the more sporadic short-duration, large-magnitude events. Accordingly, it is relatively more difficult to improve upon baseline methods (Table 2). Nonetheless, in some cases we observed considerable gains, prompting a more thorough investigation of the circumstances under which predictions of precipitation are improved by climate indices.

8 Discussion & Future Work

In this paper, we presented an empirical comparison of clustering methods for climate data on the basis of their ability to extract climate indices. Our experimental results demonstrate that ultimately the choice of algorithm is quite important: clustering matters! More specifically, community detection in climate networks stands out among competing methods as the superior approach across a diverse range of test cases, thereby reinforcing the notion that networks are able to effectively capture the complex relationships within the global climate system. In contrast, the prediction algorithm itself had a relatively smaller impact on quality of the predictions.

Consequently, the application of network-theoretical concepts could have far-reaching implications for climate science, e.g., studying properties of the climate system, detecting changes over time, and complementing the predictive skill of physics-based models with data-guided insights. In addition, complex networks may also prove useful in other applications involving ecological, environmental and/or social data, helping us understand the behavior of and interactions between these systems.

9 Acknowledgments

The authors would like to thank Prof. Vipin Kumar for the helpful discussions. Prof. Nitesh Chawla and Karsten Steinhaeuser are affiliated with the Interdisciplinary Center for Network Science & Applications (iCeNSA) at the University of Notre Dame. This work was supported in part by the NSF Grant 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 non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for Government purposes.

References

1. Clarke, F., Ekeland, I.: Nonlinear oscillations and boundary-value problems for Hamiltonian systems. Arch. Rat. Mech. Anal. **78** (1982) 315–333
2. Clarke, F., Ekeland, I.: Solutions périodiques, du période donnée, des équations hamiltoniennes. Note CRAS Paris **287** (1978) 1013–1015
3. Michalek, R., Tarantello, G.: Subharmonic solutions with prescribed minimal period for nonautonomous Hamiltonian systems. J. Diff. Eq. **72** (1988) 28–55
4. Tarantello, G.: Subharmonic solutions for Hamiltonian systems via a $\mathbb{Z}_p$ pseudoindex theory. Annali di Matematica Pura (to appear)
5. Rabinowitz, P.: On subharmonic solutions of a Hamiltonian system. Comm. Pure Appl. Math. **33** (1980) 609–633
6. S. Asur, D. Ucar, and S. Parthasarathy. An ensemble framework for clustering protein-protein interaction graphs. *Bioinformatics*, 23(13):29–40, 2007.
7. J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Mach. Learn. Res.*, 7:1–30, 2006.
8. J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *Eur. Phs. J. Special Topics*, 174:157–179, 2009.
9. R. W. Floyd. Algorithm 97: Shortest Path. *Comm. ACM*, 5(6):345, 1962.
10. R. G. Fovell and M.-Y. C. Fovell. Climate Zones of the Conterminous United States Defined Using Cluster Analysis. *J. Climate*, 6(11):2103–2135, 1993.
11. C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, 92:611–631, 2002.
12. M. H. Glantz, R. W. Katz, and N. Nicholls. *Teleconnections linking worldwide climate anomalies: scientific basis and societal impact.* Cambridge University Press, 1991.
13. 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. Nat. Acad. Sci. USA*, 102(22):7794–7799, 2005.
14. M. A. Hall and L. A. Smith. Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper. In *Int'l Florida AI Research Society Conf.*, pages 235–239, 1999.
15. J. Han, M. Kamber, and A. K. H. Tung. *Spatial Clustering in Data Mining: A Survey*, pages 1–29. Taylor and Francis, 2001.
16. W. W. Hargrove and F. M. Hoffman. Using Multivariate Clustering to Characterize Ecoregion Borders. *Comput. Sci. Eng.*, 1(4):18–25, 1999.
17. J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, (28):100–108, 1979.
18. A. K. Jain, N. N. Murty, and P. J. Flynn. Data clustering: A Review. *ACM Computing Surveys*, 31(3):264–323, 1999.
19. E. Kalnay et al. The NCEP/NCAR 40-Year Reanalysis Project. *BAMS*, 77(3):437–470, 1996.
20. L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Clustering Analysis.* Wiley, 1990.
21. T. R. Loveland et al. Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *Int. J. Remote Sensing*, 21(6-7):1303–1330, 2000.
22. A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2001.

23. P. Pons and M. Latapy. Computing communities in large networks using random walks. *J. Graph Alg. App.*, 10(2):191–218, 2006.
24. C. Race, M. Steinbach, A. R. Ganguly, F. Semazzi, and V. Kumar. A Knowledge Discovery Strategy for Relating Sea Surface Temperatures to Frequencies of Tropical Storms and Generating Predictions of Hurricanes Under 21st-century Global Warming Scenarios. *NASA Conf. on Intelligent Data Understanding*, Mountain View, CA, 2010.
25. C. F. Ropelewski and P. D. Jones. An Extension of the Tahiti-Darwin Southern Oscillation Index. *Mon. Weather Rev.*, 115:2161–2165, 1987.
26. A. Serrano, M. Boguna, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *PNAS*, 106(16):8847–8852, 2009.
27. M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. In *ACM SIGKDD Workshop on Text Mining*, 2000.
28. M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of Climate Indices using Clustering. In *ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, pages 446–455, 2003.
29. K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. Complex Networks as a Unified Framework for Descriptive Analysis and Predictive Modeling in Climate. *Technical Report TR-2010-07*, University of Notre Dame, 2010.
30. K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. Complex Networks in Climate Science: Progress, Opportunities and Challenges. *NASA Conf. on Intelligent Data Understanding*, Mountain View, CA, 2010.
31. K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. An Exploration of Climate Data Using Complex Networks. *ACM SIGKDD Explorations*, 12(1):25-32, 2010.
32. A. A. Tsonis and P. J. Roebber. The architecture of the climate network. *Physica A*, 333:497–504, 2004.
33. A. A. Tsonis, K. L. Swanson, and P. J. Roebber. What Do Networks Have to Do with Climate? *BAMS*, 87(5):585–595, 2006.
34. S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.
35. D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.
36. Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE T. Pattern Anal.*, 15(11):1101–1113, 1993.
37. K. Yamasaki and A. G. amd S. Havlin. Climate Networks around the Globe are Significantly Affected by El Niño. *Phys. Rev. Lett.*, 100(22):157–179, 2008.
38. http://cdiac.ornl.gov/climate/indices/indices_table.html.
39. http://www.cdc.noaa.gov/data/gridded/data.ncep.reanalysis.html.
40. http://www.cgd.ucar.edu/cas/catalog/climind/.