Response-Guided Community Detection: Application to Climate Index Discovery

Gonzalo A. Bello[1], Michael Angus[1], Navya Pedemane[1], Jitendra K. Harlalka[1], Fredrick H.M. Semazzi[1], Vipin Kumar[2], and Nagiza F. Samatova[1,3(✉)]

[1] North Carolina State University, Raleigh, NC, USA
samatova@csc.ncsu.edu
[2] University of Minnesota, Minneapolis, MN, USA
[3] Oak Ridge National Laboratory, Oak Ridge, TN, USA

Abstract. Discovering climate indices–time series that summarize spatiotemporal climate patterns–is a key task in the climate science domain. In this work, we approach this task as a problem of *response-guided community detection*; that is, identifying communities in a graph associated with a response variable of interest. To this end, we propose a general strategy for response-guided community detection that explicitly incorporates information of the response variable during the community detection process, and introduce a graph representation of spatiotemporal data that leverages information from multiple variables.

We apply our proposed methodology to the discovery of climate indices associated with seasonal rainfall variability. Our results suggest that our methodology is able to capture the underlying patterns known to be associated with the response variable of interest and to improve its predictability compared to existing methodologies for data-driven climate index discovery and official forecasts.

Keywords: Community detection · Spatiotemporal data · Climate index discovery · Seasonal rainfall prediction

1 Introduction

Detecting communities in real-world networks is a key task in many scientific domains. Oftentimes, domain scientists are particularly concerned with finding communities associated with a response variable of interest that can be used to analyze or predict this response variable. For example, in climate science, such communities may represent spatiotemporal climate patterns associated with a particular weather event [24], while in biology, they may represent groups of functionally associated genes associated with a particular phenotype [12].

However, community detection techniques are traditionally unsupervised learning methods, and thus do not take into account the variability of the response variable of interest. Therefore, the communities identified may not necessarily be associated with this response variable. Furthermore, even though semi-supervised methods have been proposed to incorporate prior knowledge

A. Appice et al. (Eds.): ECML PKDD 2015, Part II, LNAI 9285, pp. 736–751, 2015.
DOI: 10.1007/978-3-319-23525-7_45

to the community detection process, these methods do not consider a response variable either and require partial information about the community memberships, which may not be available [6]. For this reason, we introduce the problem of *response-guided community detection*–that is, identifying communities in a graph associated with a response variable of interest–and study its application to the discovery of *climate indices*, an important task in the climate science domain.

Climate indices are time series that summarize spatiotemporal patterns in the global climate system. These patterns are often associated with temperature, pressure, and wind anomalies, which can have a significant impact on regional climate. Consequently, climate indices are frequently used to analyze and predict regional weather events. For example, climate indices defined for El Niño Southern Oscillation (ENSO) are used to forecast Atlantic hurricane activity [9].

Climate indices were traditionally the product of hypothesis-driven research. However, the increasing amount of climate data available has led to the adoption of data-driven approaches to guide and accelerate climate index discovery, most commonly by using Principal Component Analysis (PCA) to identify major modes of variability in the data. Nonetheless, the use of PCA has important limitations in regards to the physical interpretability of the climate indices obtained and its ability to detect weaker patterns [22].

As an alternative, the application of clustering techniques, such as Shared Nearest Neighbor (SNN) clustering, to identify regions of homogeneous long-term variability in climate data has been proposed [22]. More recently, a network representation of the data has been adopted to better capture the dynamics of the global climate system [23–25]. Then, the climate index discovery task has been approached as a community detection problem [24]. The validity of the clusters or communities identified as climate indices has been evaluated in terms of their ability to predict a response variable of interest [22,24]. However, since these are unsupervised learning methodologies, the climate indices discovered may not necessarily be good predictors.

Therefore, to discover climate indices associated with a response variable of interest, we propose a methodology that explicitly incorporates information of this response variable during the discovery process by using response-guided community detection. We apply this methodology to the discovery of climate indices associated with seasonal rainfall variability in the Greater Horn of Africa, and validate the climate indices discovered in terms of their predictive power and climatological relevance. Discovering climate indices associated with a response variable of interest allows us to identify its sources of variability. Moreover, using these climate indices as predictors allows us to improve forecasts of this response variable, which is one of the major current challenges in climate science [20].

The main contributions of this paper are as follows. First, we formulate the problem of response-guided community detection (Section 2.1) and propose a general strategy to identify communities in a graph associated with a response variable of interest by explicitly incorporating information of this response variable during the community detection process (Section 2.2).

And second, we propose a methodology to discover climate indices associated with a response variable of interest from multivariate spatiotemporal data by using response-guided community detection (Section 3). As part of this methodology, we introduce a network representation of multivariate spatiotemporal data that, unlike existing network construction methodologies [23–25], builds the network in a response-guided manner, while also incorporating multiple covariates, spatial neighborhood information, and multiple related response variables to the network construction process (Section 3.1).

Finally, we should note that in this paper we only demonstrate the value of response-guided community detection in the context of climate index discovery. Its application to other problems and domains is the subject of future work.

2 Response-Guided Community Detection

In this section, we formally define the problem of response-guided community detection (Section 2.1), describe a general strategy for response-guided community detection, and present two examples of community detection algorithms that can be adapted to identify communities highly associated with a response variable of interest (Section 2.2).

2.1 Problem Statement

Let $X = \{x_{t,d,f} \in \mathbb{R} \,|\, t \in T, d \in D, f \in F\}$ be a multivariate spatiotemporal data set and $Y = \{y_t \in \mathbb{R} \,|\, t \in T\}$ be a response variable, where T is a set of time steps, D is a set of spatial points, and F is a set of covariates. For our motivating application of climate index discovery, X may be a global climate data set for a given month, Y may be the total rainfall at a target region for a given season, T may be a set of years, D may be a set of global coordinates, and F may be a set of climate variables (e.g., temperature, pressure, humidity).

Let data set X be represented as a graph $G = (V, E)$, where $V \subseteq D$ is the set of vertices, E is the set of edges, and each edge $(d_1, d_2) \in E$ is defined based on a domain-specific relationship between the data at spatial points d_1 and d_2 for all covariates $f \in F$ and over all time steps $t \in T$. For our motivating application of climate index discovery, an edge (d_1, d_2) may represent a statistically significant correlation between the data at spatial points d_1 and d_2.

Informally, we define *response-guided community detection* as the task of partitioning graph G into a set of communities C, such that every community $c_i \in C$ is highly *associated* with the response variable Y. To quantify this association, we construct an *index* for each community.

Definition 1. Given a community c_i, the *index* constructed for c_i using covariate $f \in F$, $I_{i,f}$, is defined as

$$I_{i,f}(t) = \frac{1}{|c_i|} \sum_{d \in c_i} x_{t,d,f} \quad \forall t \in T \tag{1}$$

Definition 2. Given a community c_i, the *association* of c_i with the response variable Y, ϕ_{c_i}, is defined as

$$\phi_{c_i} = \max_{f \in F} |r_{I_{i,f},Y}| \tag{2}$$

where $r_{I_{i,f},Y}$ is the Pearson's linear correlation coefficient between index $I_{i,f}$ and the response variable Y over all time steps $t \in T$.

Finally, we formally define the problem of *response-guided community detection*: Given a graph $G = (V, E)$ and a response variable Y, partition G into a set of communities $C = \{c_1, c_2, ..., c_{|C|}\}$, where $c_i \subseteq V$ for all $c_i \in C$, $c_i \cap c_j = \emptyset$ for all $c_i, c_j \in C$ with $i \neq j$, and $\bigcup_{i=1}^{|C|} c_i = V$, such that the average association with the response variable Y over all $c_i \in C$, $\bar{\phi}_C$, is maximized.

2.2 Algorithms for Response-Guided Community Detection

Community detection is one of the most widely studied topics in graph data analytics and, as a result, numerous methods have been proposed for this problem [8,11]. A common approach to community detection is to find the set of communities that maximizes a given quality function that measures the "goodness" of the partition of the graph. For traditional community detection, a "good" partition of the graph is generally such that there are many edges within the communities but few edges among them. However, for response-guided community detection, our goal is to identify communities highly associated with a response variable of interest. Therefore, we must maximize not only the "goodness" of the partition of the graph, but also the association of the communities in the partition with this variable.

To this end, we introduce a joint optimization criterion, $\mathscr{F}$, given by

$$\mathscr{F} = \alpha \cdot q(C) + (1 - \alpha) \cdot \bar{\phi}_C \tag{3}$$

where C is a set of communities, $q(C)$ is a function of the "goodness" of C, $\bar{\phi}_C$ is the average association of the communities in C with the response variable of interest (see Definition 2), and α is a tuning parameter to balance the trade-off between the "goodness" of C and the association of the communities with the response variable.

The "goodness" function is typically a metric that quantifies some structural properties of the partition of the graph. In this paper, we choose modularity– "by far the most used and best known quality function" for community detection [8]–as the "goodness" function. The modularity of a given partition of a graph is defined as the difference between the number of edges within the communities and the expected number of such edges in a random graph with the same degree distribution [17]. For a simple graph $G = (V, E)$ which vertices are partitioned into communities, the modularity Q [16] of the partition is given by

$$Q = \frac{1}{2m} \Sigma_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(v, w) \tag{4}$$

where A is the adjacency matrix of the graph (that is, A_{vw} is 1 if vertices v and w are connected and 0 otherwise), $m = \frac{1}{2}\Sigma_{vw}A_{vw}$ is the number of edges in the graph, $k_v = \Sigma_w A_{vw}$ is the degree of vertex v, and $\delta(i,j)$ is the Kronecker delta function (that is, $\delta(i,j)$ is 1 if i and j belong to the same community and 0 otherwise). Modularity optimization is an NP-complete problem [3], but many heuristic algorithms have been proposed [8].

A general strategy for response-guided community detection is to adapt modularity optimization algorithms by replacing modularity with the joint optimization criterion $\mathscr{F}$ defined in Equation 3 as the objective function. To illustrate this strategy, we next present two algorithms that can be adapted in this way to identify communities highly associated with a response variable of interest: the Louvain method, a very efficient greedy algorithm for modularity optimization, and simulated annealing, a computationally demanding but potentially more accurate optimization technique.

Greedy Algorithms for Response-Guided Community Detection. In general, greedy algorithms for modularity optimization identify communities by iteratively merging vertices or communities that result in the largest increase in the modularity of the graph partition [2,4].

In this paper, we focus on the Louvain method [2], a well-known greedy algorithm that has been shown to outperform other community detection algorithms in empirical comparative studies [15]. The Louvain method is adapted for response-guided community detection by using the joint optimization criterion $\mathscr{F}$ as the objective function.

Initially, each vertex is assigned to a different community. In the first phase of the algorithm, each vertex is iteratively and sequentially assigned to the community that yields the highest positive gain in the joint optimization criterion, $\Delta\mathscr{F}$, given by

$$\Delta\mathscr{F} = \alpha \cdot \Delta Q + (1-\alpha) \cdot \Delta\bar{\phi} \tag{5}$$

where ΔQ and $\Delta\bar{\phi}$ are the gain in modularity and the gain in average association with the response variable of interest over all communities resulting from the change in the communities, respectively.

In the second phase of the algorithm, a new graph is constructed by aggregating the vertices in each community into a single meta-vertex. These two phases are repeated iteratively until no further improvement of the joint optimization criterion $\mathscr{F}$ can be achieved.

Simulated Annealing for Response-Guided Community Detection. Another strategy that has been employed for modularity optimization is simulated annealing [14], an optimization technique that avoids local optima by incorporating stochastic noise into the search procedure. The level of noise is defined by a computational temperature $\mathcal{T}$, which decreases after each iteration.

In this paper, the simulated annealing algorithm proposed by Guimerà et al. [10] is adapted for response-guided community detection by using the joint optimization criterion $\mathscr{F}$ as the objective function.

Initially, each vertex is assigned to a different community. At each temperature $\mathcal{T}$, the algorithm performs (typically) n^2 random local movements (i.e., moving a vertex to another community) and n random global movements (i.e., merging two communities and splitting a community in two). Each of these local and global movements is accepted with probability

$$p = \begin{cases} 1, & \text{if } \Delta\mathscr{F} \geq 0 \\ \exp\left(\dfrac{\Delta\mathscr{F}}{\mathcal{T}}\right), & \text{if } \Delta\mathscr{F} < 0 \end{cases} \tag{6}$$

where $\Delta\mathscr{F}$ is the gain in the joint optimization criterion resulting from the change in the communities, as defined in Equation 5.

After all local and global moves have been evaluated, the current temperature $\mathcal{T}$ is decreased to $\mathcal{T}' = c \cdot \mathcal{T}$, where $c \in (0, 1)$ is a cooling parameter (typically between 0.990 and 0.999). The algorithm stops when a minimum temperature is reached or when there is no change in the joint optimization criterion $\mathscr{F}$ for a given number of consecutive iterations.

3 Climate Index Discovery

In this section, we describe our proposed methodology for the discovery of climate indices associated with a response variable of interest from multivariate spatiotemporal data by using response-guided community detection.

Our proposed methodology is comprised of two main steps. First, we represent the multivariate spatiotemporal data as a graph using our proposed network construction methodology (Section 3.1). Second, we identify communities in this graph using one of our adapted algorithms for response-guided community detection (see Section 2.2). For each community c_i identified, we construct an index $I_{i,f_i{}^*}$ (see Definition 1) potentially associated with the response variable, where $f_i{}^*$ is the *representative covariate* of the community, defined as

$$f_i{}^* = \arg\max_{f \in F} |r_{I_{i,f},Y}| \tag{7}$$

3.1 Network Construction Methodology

Spatiotemporal data can be represented as a graph, where each vertex is a spatial point and each edge indicates a significant relationship between a pair of spatial points. This type of representation has been adopted to model climate data, because it captures the dynamical behavior of the data's underlying system [23–25]. Furthermore, communities in these networks often have a higher association with the response variable of interest than clusters obtained using traditional clustering techniques, such as spectral clustering and the k-means clustering algorithm [24].

In this paper, we propose a methodology for the construction of climate networks associated with a response variable of interest. The key features of this methodology are as follows.

First, we construct the network in a response-guided manner. Existing methodologies for climate network construction consider all the spatial points in the data set as vertices and build the network by computing the correlation between every pair of vertices [24,25], which can be computationally expensive. In contrast, we only consider as vertices the spatial points associated with the response variable.

Second, we incorporate multiple covariates to the network construction process. Some existing methodologies have incorporated multiple covariates by defining a cross correlation function to weight the edges of the network [23]. Here, instead, we leverage the information of multiple covariates to assess the statistical significance of each edge in the network.

And third, we incorporate spatial neighborhood information and multiple related response variables to the network construction process, to increase its robustness in the case of data sets with small sample size.

Selecting the Set of Vertices. The set of vertices V of the network is selected based on the statistical significance of the relationship between each spatial point in the data set and the response variable of interest for multiple covariates. To assess this statistical significance, we first calculate the Spearman's rank correlation coefficients between the time series for each covariate at each spatial point and the response variable. Spearman's rank correlation is used to capture nonlinear relationships known to exist in climate data.

For each spatial point d, the p-values of the Spearman's rank correlation coefficients computed for each covariate are combined using Fisher's $\mathcal{X}^2$ test [7]; that is, by calculating the p-value of the test statistic given by

$$-2 \sum_{f \in F} ln(p_{X_{d,f},Y}) \tag{8}$$

where $p_{X_{d,f},Y}$ is the p-value of the Spearman's rank correlation coefficient between the time series for covariate f at spatial point d, $X_{d,f}$, and the response variable Y, over all time steps $t \in T$. The use of this combined probability test allows us to capture relationships between multiple covariates and the response variable. Finally, the set S of spatial points with a statistically significant combined p-value ($p < 0.01$) is selected as the set of vertices V of the network (i.e, spatial points potentially associated with the response variable of interest).

Defining the Set of Edges. The set of edges E of the network is defined based on the statistical significance of the relationship between each pair of spatial points in V for multiple covariates. To assess this statistical significance, we first calculate the Pearson's linear correlation coefficients between the time series for each covariate at each pair of spatial points. Climate networks constructed using Pearson's linear correlation coefficient have been shown to be highly similar to those constructed using nonlinear measures, such as mutual information [5].

For each pair of spatial points d_1, $d_2 \in V$, the p-values of the Pearson's linear correlation coefficients computed for each covariate are combined using Fisher's $\mathcal{X}^2$ test [7]; that is, by calculating the p-value of the test statistic given by

$$-2 \sum_{f \in F} ln(p_{X_{d_1,f}, X_{d_2,f}}) \tag{9}$$

where $p_{X_{d_1,f}, X_{d_2,f}}$ is the p-value of the Pearson's linear correlation coefficient between the time series for covariate f at spatial point d_1, $X_{d_1,f}$, and at spatial point d_2, $X_{d_2,f}$, over all time steps $t \in T$. Finally, an edge $(d_1, d_2) \in E$ is defined for every pair of spatial points d_1, $d_2 \in V$ with a statistically significant combined p-value ($p < 10^{-10}$, as defined in previous studies [24]).

Incorporating Spatial Neighborhood Information and Multiple Response Variables. Data sets with small sample size, such as the ones used in this study, can often lead to the selection of spatial points with spurious associations with the response variable of interest as vertices. To increase the robustness of the vertex selection in these cases, we leverage the spatial structure of the data and the information of multiple related (i.e., highly correlated) response variables (e.g., seasonal rainfall at multiple stations in the same region) by finding a consensus set of spatial points, S^*, given by

$$S^* = \bigcap_{j=1}^{h} S_j \cup \{N(d) \mid d \in S_j\} \tag{10}$$

where h is the number of response variables, S_j is the set of spatial points potentially associated with the j^{th} response variable and $N(d)$ indicates the spatial points spatially adjacent to spatial point d. We incorporate spatial neighborhood information because, given the strong spatial autocorrelations present in spatiotemporal data, it is likely that if a spatial point is associated with the response variable of interest, then its spatially adjacent points will also be associated with the response variable.

We then construct a climate network for the multiple related response variables using the previously described methodology with the consensus set of spatial points S^* as the set of vertices V of the network. Note that the rest of our proposed methodology for climate index discovery, including the response-guided community detection algorithms, can also be extended to incorporate multiple related response variables. In this case, the association of a community c_i, ϕ_{c_i} (see Definition 2), is redefined as the average association of c_i over all response variables Y_j for $j = 1, 2, ..., h$.

4 Experimental Evaluation

In this section, we describe the experimental evaluation of our proposed methodology for climate index discovery and report the results obtained. We applied

our proposed methodology to the discovery of climate indices associated with October to December (OND) rainfall variability in the Greater Horn of Africa (GHA), using data from four (4) stations with highly correlated rainfall patterns located in the North Eastern Highlands of Tanzania (Arusha, Kilimanjaro, Moshi, and Same).

4.1 Data Description

We used monthly gridded ocean data for the following climate variables: Sea Surface Temperature (SST), obtained from the NOAA Extended Reconstructed Sea Surface Temperature version 3 (ERSST V3) data set (data available from 1854 to present at 2° latitude-longitude resolution) [21], and Sea Level Pressure (SLP), Geopotential Height at 500 mb (GH), Relative Humidity at 850 mb (RH) and Precipitable Water (PW), obtained from the NCEP/NCAR Reanalysis 1 data set (data available from 1948 to present at 2.5° latitude-longitude resolution) [13]. SST, SLP, and GH are the most frequently used variables in identifying global climate patterns. We also include RH and PW as secondary variables for the temperature and water vapor content of the atmosphere.

Monthly rainfall data (52 years, from 1960 to 2011) and seasonal rainfall forecasts (14 years, from 1998 to 2011) for stations in Tanzania were provided by the Tanzania Meteorological Agency (TMA). Data was divided into a training set (38 years, from 1960 to 1997) and a test set (14 years, from 1998 to 2011). Note that only the training set was used to construct the climate networks and discover the climate indices presented in Section 4.3 and Section 4.4, respectively.

4.2 Data Preprocessing

Climate data exhibits complex characteristics, such as seasonal trends and strong spatial and temporal autocorrelations, that may hinder the performance of data mining techniques. To remove seasonality and minimize autocorrelations, we normalized the data using monthly z-scores transformations by subtracting the mean and dividing by the standard deviation of the data over the training set [24]. Since the focus of this study is on interannual variability, we also linearly detrended the data. Furthermore, all experiments were performed using a spatial resolution of 10° latitude-longitude for the gridded ocean data.

4.3 Climate Networks Constructed

Climate networks were constructed using our proposed network construction methodology with OND rainfall variability in the GHA as the response variable of interest (see Section 3.1). To capture time-lagged relationships, which are often present in climate data, five (5) climate networks were constructed, one for each month, starting four (4) months before the season (June) until the first month of the season (October). It is worth noting that when constructing a climate network for the month of May, no spatial points were selected as potentially

associated with the response variable, suggesting that this month may be too early before the season to yield significant climate indices.

Each climate network was constructed by leveraging the information of four (4) related stations in the North Eastern Highlands of Tanzania. Since these stations are located in the same climatological region and exhibit highly correlated rainfall patterns, they are expected to be associated with the same global climate patterns. Hence, the use of the consensus set allows us to filter out spatial points with potentially spurious associations with the response variable. Interstation variability is due to local factors, which are out of the scope of this paper.

4.4 Climate Indices Discovered

Communities associated with OND rainfall variability in the GHA were identified in the climate networks constructed using both the Louvain method and the simulated annealing algorithm adapted for response-guided community detection (see Section 2.2). As previously explained, we use a tuning parameter α to balance the trade-off between the modularity of the network partition and the association of the communities with the response variable of interest. For this experimental evaluation, we set the value of α to the multiple of 0.05 in the interval $[0.75, 1]$ that yields the set of communities with the highest average association with the response variable over the training set. Lower values of α were not considered to ensure a good modularity value. For each community identified, a climate index was constructed by computing the spatial average over the community of its representative climate variable (see Figure 1).

We compare our climate indices with those discovered using a baseline methodology and the state of the art [24]. For the baseline methodology, communities were identified in multivariate climate networks (i.e., one network was constructed for all covariates via a combined probability test, as described in Section 3.1) using both the original Louvain method [2] and the original simulated annealing algorithm for community detection [10]. For the state of the art [24], communities were identified in univariate climate networks (i.e., one network was constructed for each covariate) using Walktrap, a community detection algorithm based on random walks [19]. In both cases, the community detection and the network construction were performed in an unsupervised manner.

Table 1 summarizes the properties of the climate networks constructed and the climate indices discovered using each methodology. Given that our response-guided community detection algorithms do not exclusively optimize the "goodness" of the network partitions, our climate networks exhibit a lower modularity than those constructed using unsupervised methodologies (0.34 vs. 0.74, 0.75, and 0.59). However, our communities have a higher internal density (0.62 vs. 0.29, 0.28, and 0.47) and a lower internal variability (0.63 and 0.62 vs. 0.77, 0.78, and 0.74), indicating a well-defined structure.

We also observe that, unlike most of the climate indices discovered using the baseline and the state of the art, the majority of our climate indices (66.67%) have a statistically significant linear correlation ($p < 0.01$) with the response variable of interest over the training set. Moreover, our proposed methodology

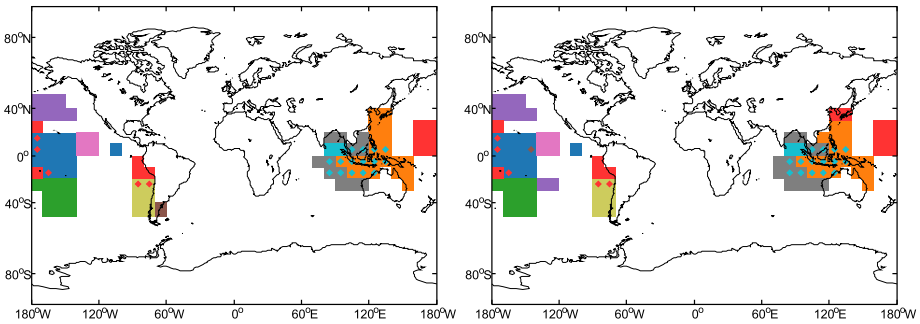

Fig. 1. Climate indices discovered using our proposed methodology with the response-guided community detection algorithm based on the Louvain method (left) and simulated annealing (right), respectively, and with OND rainfall variability in the GHA as the response variable of interest. Each color represents a different index, and diamonds indicate overlaps between indices. To improve visualization, only the top 10 indices with the highest association with the response variable over the training set are shown in each figure. Best viewed in color.

Table 1. Properties of networks constructed and climate indices discovered for OND rainfall variability in the GHA, using the proposed, baseline, and state-of-the-art (SOTA) [24] methodologies with the Louvain method (LM), simulated annealing (SA) and Walktrap as the community detection algorithms: number of networks (Num Nets), average number of vertices and edges per network (Avg Vtxs, Avg Edges), average modularity (Avg Mod), number of indices (Num Idxs), average number of vertices, standard deviation, and internal density per index (Avg Vtxs, Avg Std, Avg Dens), and percentage of indices with a statistically significant ($p < 0.01$) linear correlation with the response variable of interest (% Idxs). Best values are highlighted in bold.

Method	Algorithm	Networks				Indices				Significant Indices	
		Num Nets	Avg Vtxs	Avg Edges	Avg Mod	Num Idxs	Avg Vtxs	Avg Std	Avg Dens	Num Idxs	% Idxs
Proposed	Adapted LM	5	40.80	169.20	0.34	18	11.33	0.63	**0.62**	12	**66.67**
	Adapted SA	5	40.80	169.20	0.34	18	11.33	**0.62**	**0.62**	12	**66.67**
Baseline	Original LM	5	446.00	2614.60	0.74	49	45.51	0.77	0.29	6	12.24
	Original SA	5	446.00	2614.60	**0.75**	50	44.60	0.78	0.28	4	8.00
SOTA	Walktrap	25	444.80	7493.80	0.59	265	41.96	0.74	0.47	6	2.26

Table 2. Average linear correlation with OND rainfall at each station and at the GHA region, over the training set and the test set, of climate indices discovered for OND rainfall variability in the GHA using the proposed, baseline, and state-of-the-art (SOTA) [24] methodologies with the Louvain method (LM), simulated annealing (SA) and Walktrap as the community detection algorithms. Check marks (✓) indicate that our proposed methodology performs significantly better according to a two-way ANOVA at the 95% confidence level. Best values are highlighted in bold.

Station	Proposed				Baseline				SOTA	
	Adapted LM		Adapted SA		Original LM		Original SA		Walktrap	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Arusha	**0.4436**	**0.2999**	0.4431	0.2848	0.2496	0.2495	0.2489	0.2639	0.1481	0.2261
Kilimanjaro	0.4103	**0.3752**	**0.4300**	0.3583	0.2586	0.2437	0.2629	0.2525	0.1567	0.2230
Moshi	0.3629	**0.2980**	**0.3764**	0.2791	0.2404	0.2552	0.2317	0.2501	0.1393	0.2481
Same	0.4292	**0.3403**	**0.4341**	0.3119	0.2574	0.2111	0.2572	0.2429	0.1589	0.2148
GHA	0.4502	**0.3478**	**0.4614**	0.3272	0.2763	0.2356	0.2749	0.2497	0.1558	0.2219
Two-way ANOVA ($\alpha = 0.05$)					✓	✓	✓	✓	✓	✓

performs significantly ($p < 0.05$) better than the baseline and the state of the art across all stations in terms of the average linear correlation between the climate indices and the response variable of interest over the training set and the test set (see Table 2). This shows that, as expected, our proposed methodology is able to discover climate indices more highly associated with the response variable of interest than those discovered using unsupervised methodologies.

4.5 Seasonal Rainfall Prediction

We validate the climate indices discovered with our proposed methodology by assessing their predictive power for OND rainfall in the GHA. To this end, we trained linear regression models to predict rainfall at each station, and average rainfall at the region, using our climate indices as predictors. As specified in Section 4.1, data from 1960 to 1997 was used for training and data from 1998 to 2011 was used for testing. For comparison, linear regression models were also built using the climate indices discovered with the baseline and state-of-the-art [24] methodologies introduced in Section 4.4.

In order to avoid overfitting given the small sample size of the data sets, only the top six (6) climate indices with the highest average correlation with OND rainfall in the GHA over the training set were used to build the models. This number of predictors was selected because it yielded relatively stable performance over the training set across all methodologies (see Figure 2). Furthermore, to evaluate the ability of the models to make predictions before the start of the OND rainfall season, all experiments were preformed using data up to the month of August (one-month lead time). Climate indices discovered for the months of September and October were reconstructed using August data.

The correlations between predicted and true rainfall and the root mean squared errors (RMSE) obtained for each methodology are shown in Table 3.

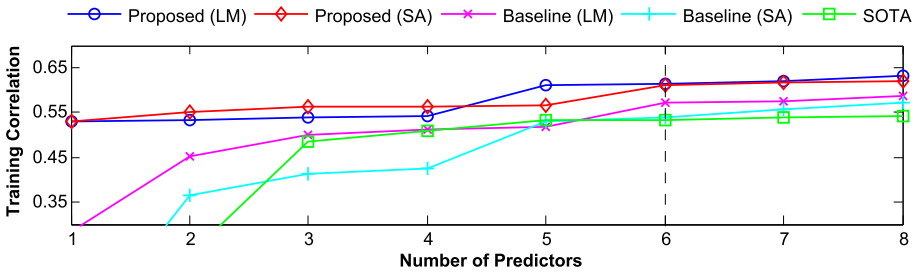

Fig. 2. Average linear correlation between true and predicted rainfall for predictions of OND rainfall at each station in the GHA region over the training set using the proposed, baseline, and state-of-the-art (SOTA) [24] methodologies with the Louvain method (LM), simulated annealing (SA) and Walktrap as the community detection algorithms vs. the number of predictors used to build the regression models. The dashed line indicates the number of predictors selected for further analysis.

Table 3. Linear correlation between true and predicted rainfall (Corr) and RMSE scores for predictions of OND rainfall at each station and at the GHA region from 1998 to 2011 obtained using the proposed, baseline, and state-of-the-art (SOTA) [24] methodologies with the Louvain method (LM), simulated annealing (SA) and Walktrap as the community detection algorithms. Check marks (✓) indicate that our proposed methodology performs significantly better according to a two-way ANOVA at the 95% confidence level. Best values are highlighted in bold.

Station	Proposed				Baseline				SOTA	
	Adapted LM		Adapted SA		Original LM		Original SA		Walktrap	
	Corr	RMSE	Corr	RMSE	Corr	RMSE	Corr	RMSE	Corr	RMSE
Arusha	**0.7143**	**0.5017**	0.5869	0.5215	0.2462	0.5023	0.3432	0.6853	0.2034	0.5779
Kilimanjaro	**0.7629**	**0.5034**	0.6736	0.5619	0.1844	1.0432	0.2053	0.7477	0.2940	0.7874
Moshi	0.6561	0.4719	**0.6564**	**0.4664**	-0.0319	0.5937	0.1059	0.7055	0.3088	0.6231
Same	**0.7237**	0.4779	0.6896	**0.4749**	0.1470	0.6796	0.1806	0.6929	0.2575	0.7121
GHA	**0.7722**	0.4133	0.7425	**0.4007**	0.1501	0.6316	0.2135	0.6390	0.2665	0.6053
Two-way ANOVA ($\alpha = 0.05$)					✓	✓	✓	✓	✓	✓

We observe that the models built using our climate indices yield a significantly ($p < 0.05$) higher correlation and lower RMSE than those built using climate indices discovered using unsupervised methodologies. This suggests that climate indices more highly associated with the response variable of interest, as the ones discovered using our proposed methodology, have greater predictive power.

We further assess the predictive power of our climate indices by comparing our predictions with the official forecasts of the OND rainfall season issued by the TMA every year on September. To this end, the rainfall season for each year was categorized according to the guidelines of the TMA as *below normal*, *normal*, or *above normal* (rainfall below 75%, between 75% and 125%, or above 125% of long-term averages, respectively). Long-term averages were computed using the training set. Similarly to the regression models, decision trees to classify the

OND rainfall season at each station were trained using data up to the month of August and considering only the top six (6) climate indices discovered with our proposed methodology as predictors. The decision trees were built using the Gini index as the split criterion and pruning to avoid overfitting.

The classification accuracies obtained are shown in Figure 3. We observe that the accuracy of the decision trees built using our climate indices is higher than that of the official forecasts for three (3) out of four (4) stations. This suggests that the use of the climate indices discovered using our proposed methodology can potentially improve forecasts of the response variable of interest.

4.6 Physical Interpretation of Climate Indices Discovered

Finally, we discuss the climate indices discovered in terms of their climatological relevance. Rainfall variability in the GHA is known to be mainly associated with

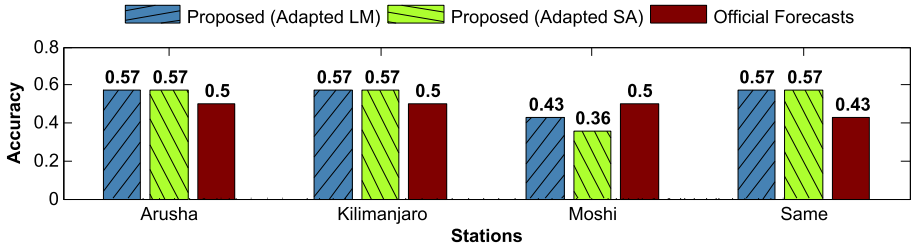

Fig. 3. Classification accuracy of the prediction of the OND rainfall season at each station in the GHA region from 1998 to 2011 obtained using the proposed methodology with the Louvain method (LM) and simulated annealing (SA) as the community detection algorithms, as well as official forecasts issued by the TMA.

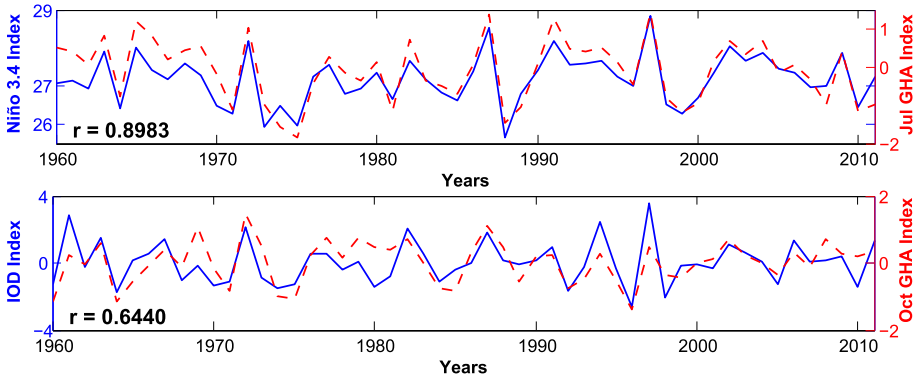

Fig. 4. Time series of the Niño 3.4 index (upper, solid line) and the IOD index (lower, solid line) with climate indices discovered in July (upper, dashed line) and October (lower, dashed line) using our proposed methodology with the adapted Louvain method as the community detection algorithm and OND rainfall variability in the GHA as the response variable of interest. The linear correlation between the time series is shown in the lower left corner of each figure.

ENSO in the equatorial Pacific Ocean [18] and the Indian Ocean Dipole (IOD) in the tropical Indo-Pacific Ocean [1].

Climate indices significantly correlated ($p < 0.01$) with ENSO, in particular with the Niño 3.4 index, were discovered in June, July, August, September, and October using both adapted community detection algorithms (for example, see Figure 4). The representative climate variable selected for these climate indices is mostly either SST or PW, a close proxy of SST in the equatorial Pacific Ocean in the NCAR/NCEP Reanalysis 1 data set. Higher SSTs in the equatorial Pacific Ocean are associated with a suppression of East African rainfall, by modulating the strength of the global upper level wind flow [18].

Climate indices significantly correlated ($p < 0.01$) with the IOD were discovered in July, August, September and October using both adapted community detection algorithms (for example, see Figure 4). These climate indices were generally discovered closer to the onset of the OND rainfall season than the ones in the equatorial Pacific Ocean, as the IOD exerts its influence on East African rainfall on a shorter timescale through local wind anomalies [1].

5 Conclusions

In this paper, we introduced the problem of response-guided community detection through its application to the task of climate index discovery. We proposed a methodology for the discovery of climate indices associated with a response variable of interest from multivariate spatiotemporal data, the contribution of which is twofold. First, we proposed a general strategy for response-guided community detection, and second, we introduced a network representation of the data that incorporates information from multiple variables.

We applied our proposed methodology to the discovery of climate indices associated with seasonal rainfall variability in the GHA. The climatological relevance of the climate indices discovered is supported by domain knowledge, as evidenced by their association with traditional climate indices known to be related to seasonal rainfall in the region. Furthermore, our results show that our methodology improves the forecast skill for this response variable with respect to existing methodologies for climate index discovery, as well as official forecasts.

Acknowledgments. This material is based upon work supported in part by the Laboratory for Analytic Sciences, the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, and NSF grant 1029711.

References

1. Black, E., Slingo, J., Sperber, K.R.: An observational study of the relationship between excessively strong short rains in coastal East Africa and Indian Ocean SST. Mon. Weather Rev. **131**(1), 74–94 (2003)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. **2008**(10), P10008 (2008)

3. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On finding graph clusterings with maximum modularity. In: Brandstädt, A., Kratsch, D., Müller, H. (eds.) WG 2007. LNCS, vol. 4769, pp. 121–132. Springer, Heidelberg (2007)
4. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**(6), 066111 (2004)
5. Donges, J.F., Zou, Y., Marwan, N., Kurths, J.: Complex networks in climate dynamics. The European Physical Journal-Special Topics **174**(1), 157–179 (2009)
6. Eaton, E., Mansbach, R.: A spin-glass model for semi-supervised community detection. In: Proc. of the 26th AAAI Conference on Artificial Intelligence, pp. 900–906. AAAI (2012)
7. Fisher, R.A.: Statistical methods for research workers. Edinburgh (1934)
8. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3), 75–174 (2010)
9. Gray, W.M.: Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb quasi-biennial oscillation influences. Mon. Weather Rev. **112**(9), 1649–1668 (1984)
10. Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Nature **433**(7028), 895–900 (2005)
11. Harenberg, S., Bello, G.A., Gjeltema, L., et al.: Community detection in large-scale networks: a survey and empirical evaluation. WIREs Comput. Stat. (1939-0068) (2014)
12. Harenberg, S., Seay, R.G., Ranshous, S., et al.: Memory-efficient query-driven community detection with application to complex disease associations. In: Proc. of the 2014 SIAM Int. Conf. on Data Mining, pp. 1010–1018. SIAM (2014)
13. Kalnay, E., Kanamitsu, M., Kistler, R., et al.: The NCEP/NCAR 40-year reanalysis project. Bull. Amer. Meteor. Soc. **77**(3), 437–471 (1996)
14. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simmulated annealing. Science **220**(4598), 671–680 (1983)
15. Lancichinetti, A., Fortunato, S.: Community detection algorithms: a comparative analysis. Phys. Rev. E **80**(5), 056117 (2009)
16. Newman, M.E.J.: Analysis of weighted networks. Phys. Rev. E **70**(5), 056131 (2004)
17. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
18. Omondi, P., Ogallo, L.A., Anyah, R., et al.: Linkages between global sea surface temperatures and decadal rainfall variability over Eastern Africa region. Int. J. of Climatol. **33**(8), 2082–2104 (2013)
19. Pons, P., Latapy, M.: Computing communities in large networks using random walks. J. Graph Algorithms Appl. **10**(2), 191–218 (2006)
20. Schiermeier, Q.: The real holes in climate science. Nature **463**(7279), 284–287 (2010)
21. Smith, T.M., Reynolds, R.W., Peterson, T.C., Lawrimore, J.: Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). J. Climate **21**(10), 2283–2296 (2008)
22. Steinbach, M., Tan, P.N., Kumar, V., et al.: Discovery of climate indices using clustering. In: Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 446–455. ACM (2003)
23. Steinhaeuser, K., Chawla, N.V., Ganguly, A.R.: An exploration of climate data using complex networks. ACM SIGKDD Explor. Newsl. **12**(1), 25–32 (2010)
24. Steinhaeuser, K., Chawla, N.V., Ganguly, A.R.: Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. Statistical Analysis and Data Mining **4**(5), 497–511 (2011)
25. Tsonis, A.A., Roebber, P.J.: The architecture of the climate network. Phys. A **333**, 497–504 (2004)