Nonlin. Processes Geophys., 22, 33–46, 2015
www.nonlin-processes-geophys.net/22/33/2015/
doi:10.5194/npg-22-33-2015

Nonlinear Processes in Geophysics
Open Access

On the data-driven inference of modulatory networks in climate science: an application to West African rainfall

D. L. González II[1,2], M. P. Angus[1], I. K. Tetteh[1], G. A. Bello[1,2], K. Padmanabhan[1,2], S. V. Pendse[1,2], S. Srinivas[1,2], J. Yu[1], F. Semazzi[1], V. Kumar[3], and N. F. Samatova[1,2]

[1]North Carolina State University, Raleigh, NC 27695-8206, USA
[2]Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831, USA
[3]University of Minnesota, Minneapolis, MN 55455, USA

Correspondence to: N. F. Samatova (samatova@csc.ncsu.edu)

Received: 1 February 2014 – Published in Nonlin. Processes Geophys. Discuss.: 4 April 2014
Revised: 10 July 2014 – Accepted: 21 November 2014 – Published: 13 January 2015

Abstract. Decades of hypothesis-driven and/or first-principles research have been applied towards the discovery and explanation of the mechanisms that drive climate phenomena, such as western African Sahel summer rainfall variability. Although connections between various climate factors have been theorized, not all of the key relationships are fully understood. We propose a data-driven approach to identify candidate players in this climate system, which can help explain underlying mechanisms and/or even suggest new relationships, to facilitate building a more comprehensive and predictive model of the modulatory relationships influencing a climate phenomenon of interest. We applied coupled heterogeneous association rule mining (CHARM), Lasso multivariate regression, and dynamic Bayesian networks to find relationships within a complex system, and explored means with which to obtain a consensus result from the application of such varied methodologies. Using this fusion of approaches, we identified relationships among climate factors that modulate Sahel rainfall. These relationships fall into two categories: well-known associations from prior climate knowledge, such as the relationship with the El Niño–Southern Oscillation (ENSO) and putative links, such as North Atlantic Oscillation, that invite further research.

1 Introduction

The climate system is inherently complex, due to the existence of nonlinear interactions, or *couplings*, between its subsystems (e.g., the ocean and the atmosphere), global-scale temperature anomalies (e.g., El Niño–Southern Oscillation), and other climate behaviors. Such a system exhibits hierarchical modularity of its organization and function (Havlin et al., 2012): each constituent subsystem performs a similar function and does not act in isolation; instead, they interact or cross-talk. The challenge is to discover the key subsystems and their cross-talk mechanisms, that is, the positive and negative feedbacks that collectively modulate the dynamic behavior of the system through a sophisticated network of modulatory pathways that ultimately define the system's functional response.

For example, the rainfall anomaly in the Sahel region of western Africa, which is the focus of this study, represents a "functional response" for the climate system, which is in actuality the predictant of a model (such as Lasso, DBN, etc.). Rainfall in the Sahel is dependent on global sea surface temperature (SST) patterns, as well as on local climate variability. There is a multitude of complex associations between various subsystems that drive the Sahel's climate response mechanisms. Some of these associations have been discovered throughout more than two decades of hypothesis-driven and/or first-principles-based research. These associations include a diverse range of climate mechanisms. For example, warmer temperatures in the Mediterranean Sea re-

gion lead to increased evaporation, and southward moisture advection in the lower troposphere toward the Sahel (Rowell, 2003). On a more global scale, the Atlantic Multidecadal Oscillation (AMO) displaces the intertropical convergence zone (ITCZ) further northward, bringing more moisture to the Sahel region (Zhang and Delworth, 2006). The North Atlantic Oscillation (NAO) has been linked to the moisture budget in northern Africa (Hurrell, 1995) through a direct influence on the sea level pressure (SLP), although this mechanism remains underexplored. In the Pacific, a warm ENSO event is associated with enhanced trade winds over the tropical Atlantic and weaker moisture advection over West Africa, consistent with a weaker monsoon system strength (Janicot et al., 2001). Figure 1 illustrates an overview of the climate modulatory network, which is a collection of modulatory pathways, with some mechanisms driving rainfall in the Sahel known to be directly/indirectly associated, and some not fully understood. Comprehending these mechanisms is particularly important due to the influence of rainfall variability in the region. Severe drought occurred throughout the 1970s and 1980s, leading to severe disruption of agriculture and major food shortages (Mortimore and Adams, 2001). Dry conditions (low rainfall anomaly) also lead to the spread of meningitis as, under wet conditions, higher humidity during both the spring and summer seasons strongly reduces disease risk by decreasing the transmission capacity of the bacteria (Sultan et al., 2005). These issues make the Sahel particularly vulnerable to fluctuations in rainfall, and provide motivation to improve domain scientists' knowledge of the contributing factors (Tetteh, 2012).

For a mechanistic understanding of functional responses such as African Sahel rainfall, we posit that a data-driven approach may facilitate the discovery of key players that might cross-talk by identifying candidate modulatory pathways and/or suggesting new factors and relationships with the proper characterization of their inductive or suppressive roles. The goal of our approach is to elucidate the putative modulatory pathways that suggest cross-talking mechanisms controlling a system's functional response. More specifically, given the key climate drivers and their modulatory directions on the response, we must infer (a) the putative pathways of modulatory events (e.g., Pacific ENSO $\rightarrow$ AMO $\rightarrow$ Sahel rainfall in Fig. 1) and (b) the modulatory signs (e.g., induction vs. suppression, such as a positive anomaly sign of Atlantic ENSO (EATL), $\text{EATL}_{\text{HIGH}}$ being related to the negative anomaly sign of Sahel rainfall, $\text{Rainfall}_{\text{LOW}}$) that collectively define the network of modulatory pathways for the response. Furthermore, given that there is a variety of methodologies that can be used to find such modulatory relationships, we must provide a consensus result that accounts for all evidence of a given relationship. To the best of our knowledge, this is a novel proposition in the field of knowledge discovery in the physical science domain, in general, and climate extremes (e.g., droughts), in particular. Moreover, this data-driven approach could contribute, in the long run, to the identification and characterization of more comprehensive and predictive models of the physical phenomenon under study.

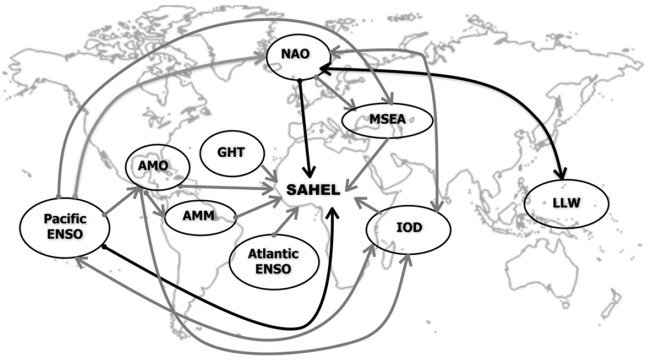

Figure 1. Complex relationships between climate indices and Sahelian rainfall, with some direct and indirect relationships well defined in the literature (light arrows) and others not fully understood (dark arrows).

2 Methods

In our previous work, we proposed an approach for the aforementioned data-driven, semi-automatic inference of phenomenological physical models based on Lasso multivariate regression (Pendse et al., 2012). This approach was applied to quantify the influence of key factors on the Sahel rainfall anomaly. The results obtained enabled the formulation of the North Atlantic Oscillation (NAO)-driven hypothesis, among others, which theorizes that the NAO modulates the drivers of West African climate, the Atlantic dipole and the EATL, via the low-level westerly (LLW) jet.

We extended this work by developing coupled heterogeneous association rule mining (CHARM), which allowed us to mine higher-order couplings of climate relationships and to capture the anomaly phases with which each climate factor is related to each other (e.g., a negative anomaly of LLW may be related to a positive anomaly of EATL, and the presence of both factors may be associated with a negative Sahel rainfall anomaly) (Gonzalez et al., 2013) (Sect. 2.1). Such relationships are not typically captured from modulatory inference frameworks, let alone traditional association rule mining (ARM) methodologies.

Here, we propose to extend CHARM by incorporating other existing methodologies, namely Lasso multivariate regression (Tibshirani, 1994) (Sect. 2.2) and dynamic Bayesian networks (Murphy, 2002) (Sect. 2.3), as complementary approaches to increase the confidence of the inferred modulatory relationships. Moreover, in order to obtain a consensus as to which of the relationships identified have the most evidence of being present, we treat the results of each methodology as individual pieces of evidence in an information fusion approach, and combine them into a unified, coherent re-

sult. This unified result can provide us with a means with which to increase the confidence of the relationships identified throughout the different methodologies. This should allow us to contrast the methodologies by studying how each of their results differ, and to correlate these results with known relationships found in the literature. Furthermore, the application of this unified result to the climate network may allow the identification of previously undiscovered relationships, which can then be analyzed from a traditional climate perspective. In Sect. 3 (specifically Table 3), we present the application of such a method to the climate indices affecting Sahel rainfall.

2.1 CHARM: coupled heterogeneous association rule mining

CHARM is an extension of ARM that enables the discovery of climatologically relevant modulatory pathways from spatio-temporal climate data. The traditional ARM methodology CHARM is based on is presented in Sect. 2.1.1, and the limitations of ARM that CHARM aims to address are described in Sect. 2.1.2.

2.1.1 Traditional association rule mining (ARM)

Traditional ARM was pioneered by Agrawal et al. (1993) as a methodology for capturing the frequency with which two items are present within transactions in market basket data. For instance, Fig. 2 presents a set of transactions that indicate whether or not an item was purchased. ARM takes this information and organizes each transaction as a combinatorial set of items. For example, Customer_2 has one 3-item set (i.e., an item set with 3 items) {Bread, Diapers, Beer} as its largest possible item set, and three 2-item sets: {Bread, Diapers}, {Bread, Beer}, and {Diapers, Beer}. By studying the frequency with which each item set occurs across transactions, one can potentially conclude that said items are typically purchased together, and are possibly related. In Fig. 2, we see that {Bread, Milk} occurs in 60 % of transactions, and thus we can say that if we were to see bread in a transaction, 60 % of the time we should see milk in that transaction as well. This measure is known as the *support* of an association rule.

In our work, we capture a climate relationship as being such item sets of climate variables that co-occur at least twice in our data, where, in our case, climate variables are climate indices, as discussed in Sect. 2.1.2. Literature support for such relationships is available in Sect. 3, and captured in Table 3.

The aforementioned relationship can be equally represented as {Bread → Milk} or {Milk → Bread}, utilizing an arrow to capture the fact that the presence of the antecedent (i.e., the items on the left side) implies that the consequent (i.e., the items on the right side) will be present with the given support. However, if we use a metric such as *confidence*, which captures the conditional probability of the consequent being present given the presence of the antecedent, the direction of the rule carries more weight. For example, {Bread → Milk} has a confidence of 0.75 (of the four times bread is present, milk is only present thrice), while {Milk → Bread} has a confidence of 1 (bread is present every time milk is present) (Tan et al., 2006). As such, the metric used to measure the interest of mined rules affects their overall interpretation, and should be selected carefully (Sect. 2.1.4) (Tan et al., 2001).

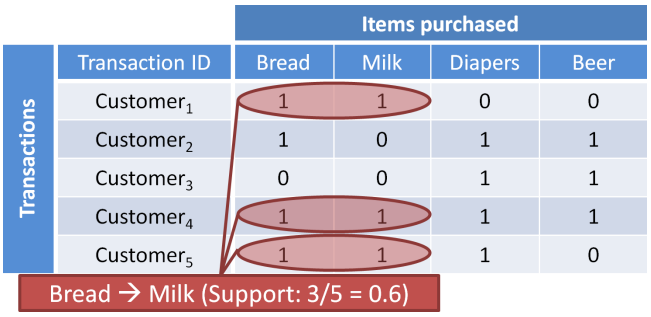

Figure 2. Simplistic traditional representation of market basket data in the form of transactions. The Bread → Milk rule has a support of 0.6, meaning it appears in 60 % of the transactions.

ARM is an increasing area of interest for domain sciences, because of the growing need to mine data to identify the co-occurrence of important events (Agrawal and Srikant, 1994; Tan et al., 2001). ARM, unlike other methodologies for inference of phenomenological models, takes into account the latent but vital signals embedded in the intermediary pathways associated with the system's functional response. However, the application of ARM to spatio-temporal climate data puts forth a series of challenges. In Sect. 2.1.2, we outline these challenges and describe CHARM as a means of addressing them.

2.1.2 Coupling of climate indices

Due to the complexity of the climate system, building comprehensive models over climate data is not trivial, in part because of the interactions between its subsystems, the dimensionality and structure of its underlying data, and the quality of such data.

The key drivers of a climate system are spatially distributed and active at different temporal phases in the modulatory network of the system's functional response. State-of-the-art mining methodologies are not well equipped to handle such diversity of spatio-temporal alignment between the system's features. For example, due to the transactional nature of the ARM methods, each spatial grid point (specified by latitude, longitude, and/or altitude) at a given point in time (e.g., month and year) defines a transaction ID, or a row in the transaction matrix. This requires that all features be aligned with respect to their transaction IDs, complicating the use of

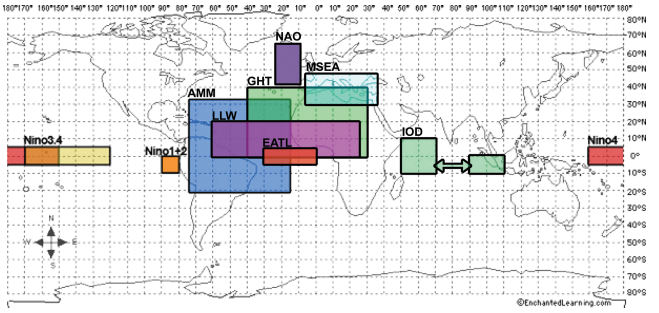

Figure 3. Climate indices can be distant, or many times partially co-located, complicating spatial alignment.

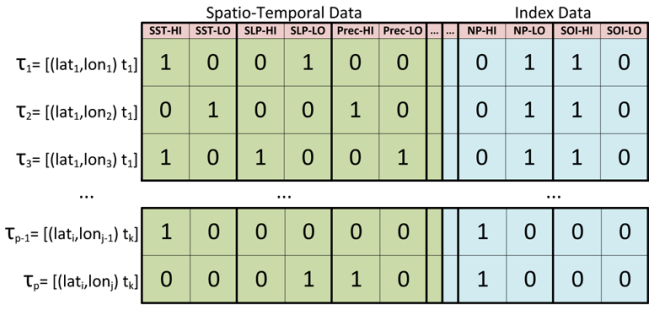

Figure 4. Spatially defined variables' anomaly presence or absence is affixed to climate index anomalies expanded to represent a global effect.

multi-resolution, multivariate, spatio-temporal climate data by these methods. For this reason, we leverage climate indices, known to be a valid abstraction of the underlying subsystem's zonal climate behavior (Hallett et al., 2004), thus significantly reducing the number of features needed to capture spatial data. However, these climate indices that capture data for different subsystems are located in different parts of the globe (see Fig. 3).

Some climate variables from observations or simulations (e.g., SST) are defined only over the ocean, yet others (e.g., rainfall) are defined over land. Hence, considering both features as columns in a transactional matrix is impossible, given that they have no common grid points. Even if they share some spatial region, they are often still not perfectly aligned, due to variation of their grid resolutions. While mathematical methods (e.g., interpolation or extrapolation) exist to facilitate data alignment, they introduce uncertainty and instability, affecting the interpretability of the results (Gonçalves, 2006). Subsequently, when a new feature is integrated into the study, realignment and the aforementioned mathematical operations must be performed again.

An ARM-based approach for the discovery of relationships among climate variables was proposed by Tan et al. (2001). This approach studies only spatially aligned data sets, and affixes climate indices alongside them. However, due to its inherently grid-based nature, this approach assigns each climate index's locally observed anomalies to all grid points. That is, it assumes that the anomaly affects the entire globe equally. While such an assumption may hold for some climate drivers, it can increase the number of false positives due to its inherent amplification bias. For example, in the representative case shown in Fig. 4, the high anomaly of the SOP index (SOI-HI) and the low anomaly of the NP index (NP-LO) would occur in so many transactions (see Sect. 2.1.1) that it would be present in every item set for time t_1, which complicates the understanding of the information gained from any resulting rules that include them.

Climate scientists often study climate factors in a coupled manner and relate certain variables to others. Traditional lagged climate techniques employed for coupled pattern analysis include singular value decomposition and grid point correlations, among others (Polo et al., 2008). For example, principal component analysis (PCA) has been used to determine the relationship between the Indian Ocean dipole and East African rainfall (Schreck and Semazzi, 2004; Manatsa et al., 2012). Hence, we adopt a similar approach by coupling climate indices. We take a quotient of these relationships before identifying any anomaly, to capture the anomaly in the relationships between these variables.

For each climate index λ to be used as a *coupling listener*, we iterate through all other climate indices δ as *coupling inciters*, and calculate their ratio, δ/λ, as a *data coupling* that intends to capture the behaviors of the logical sentence "how abundant is δ given the presence of λ?". An issue in calculating these ratios is the potential emergence of large values due to the denominator possibly being orders of magnitude smaller than the numerator. To handle this, we normalize the resulting data couplings such that they range between -1 and 1, allowing us to avoid wide-ranging quotients that could affect the abstraction of anomalous events (Sect. 2.1.3).

We note that each tuple (row in the database) now represents a specific coupling inciter λ and time, while each column represents a particular coupling listener δ, and each cell contains the relevant data coupling value. For ARM, these data must be binned, after which the resultant dataset cells indicate the presence or absence of anomalies for each previously calculated coupling, described further in Sect. 2.1.3. We address the increase in dimensions this leads to in Sect. 2.2.1, by using only the most prominent temporal phases in order to reduce the search space.

2.1.3 Identifying anomalous events

As suggested by NOAA (2014), and based on our interactions with climate scientists, we identify anomalies as any set of values below the 33.33rd percentile or above the 66.67th percentile for any given variable. Given that the data were normalized before calculating ratios, we identify the anomalies using the aforementioned norm, based on the phase-wise groupings. We take each tuple correspond-

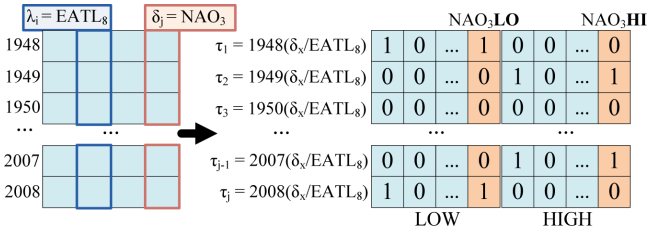

Figure 5. Data coupling for $\lambda_i = \mathrm{EATL}_8$ and $\delta_j = \mathrm{NAO}_3$.

ing to each unique combination of δ and λ, and identify high anomalies as being those ratios in the upper 66.67th percentile, and low anomalies as being ratios in the lower 33.33rd percentile.

Since we are trying to identify the presence or absence of anomalous events, we divide each column into two separate high and low cases, and assign a binary 1 when either anomaly occurs, and a 0 otherwise. This results in a very sparse matrix, as no particular year can fit in both high and low categories, and it is likely that the majority of years have most variables falling into a non-anomalous category.

Figure 5 represents a particular (i, j)th iteration. In this example, for the coupling of $\lambda_i = \mathrm{EATL}_8$ and $\delta_j = \mathrm{NAO}_3$, we identify high and low anomalies and assign transaction IDs that indicate that the cells pertain to the coupling of that year's data for the listener (shown in the column header) against the stated inciter. This transaction ID shows that each row in the matrix consists of the anomaly of the ratios calculated over each possible coupling δ_x/λ_i, where x indicates that all values in this row were divided by the same λ_i.

2.1.4 CHARM pathway significance assessment

As mentioned before, rule interestingness in ARM is estimated using metrics that quantify the importance of each rule. Selecting which metric to use depends on the information to be obtained (Tan et al., 2001). Support and confidence are commonly used to measure rule quality, and although there are other possible metrics to measure interestingness, none is regarded a "catch-all" for high-quality rules (Tan et al., 2001). Such metrics also require predetermined thresholds to cut off rules deemed "uninteresting", which reduces the accuracy in retention of significant rules. Hence, we first prune based on bare-minimum thresholds for support and confidence (the rule appearing in a single transaction), and if any given rule needs to be further analyzed (based on domain knowledge or otherwise), we perform a Monte Carlo simulation to test against the null hypothesis of observing the rule at random. Thus, we define a rule as significant and interesting if it meets the following criteria: p value ≤ 0.01, support $\geq 6\,\%$, and confidence $\geq 75\,\%$.

This criteria constrains the search space and trims the result space, pruning unimportant rules. Once a set of possibly

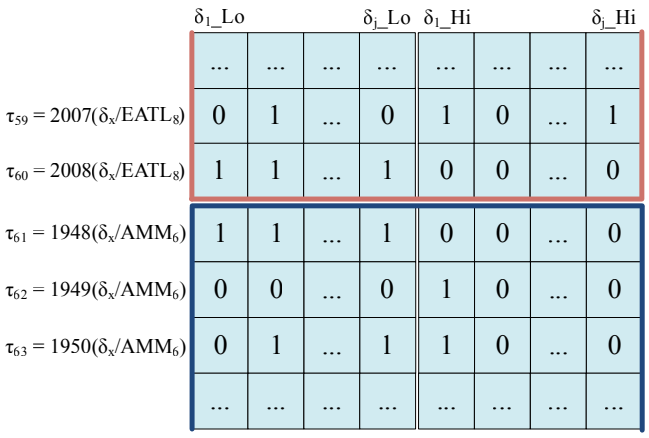

Figure 6. Heterogeneous rule sets are generated for each coupling inciter individually, preserving the independence of their anomalies.

interesting rules is identified, the computationally more demanding, but embarrassingly parallel, statistical significance test is applied to prune insignificant rules further. On average, this removed 20–30 % of the generated rules from the result set.

2.1.5 Coupling heterogeneity

Data coupling creates a large set of transactions, covering each year studied for each possible coupling inciter. Rules that only have sufficient support when counted over multiple inciters would be difficult to interpret; thus, we must heterogeneously generate rules for each coupling inciter separately, as shown in Fig. 6. This allows us to identify a preferential bias towards a particular coupling inciter, and preserves information relating to each data coupling individually.

2.1.6 CHARM computational complexity

Finding all frequent item sets for ARM is an *NP*-complete problem that, when bounding transaction length, becomes linear with complexity $\mathcal{O}(r \cdot n \cdot 2^l)$, where n is the transaction count, l is the maximum item-set length, and r is the number of maximal frequent item sets (Zaki, 2000). Rules are generated such that a user-specified minimum confidence and/or a minimum support is satisfied. Thus, for an item set of length k, there are $2^k - 2$ potentially confident rules, making the complexity $\mathcal{O}(c \cdot 2^q)$, where c is the number of frequent item sets, and q is the length of the longest frequent item set (Tan et al., 2001; Zaki, 2000).

CHARM leverages the *Apriori* algorithm, ensuring that only maximal frequent item sets are considered for rule generation (Agrawal et al., 1993), while leveraging sequential ARM to identify such relationships across different temporal instances (Huang et al., 2008). As mentioned in Sect. 2.1.5, each inciter is studied heterogeneously. Thus, the

Table 1. Prominent season selections for climate variables.

No.	Climate variable	Abbr.[3]	Seasons chosen (top 3)[1,2]											
			1	2	3	4	5	6	7	8	9	10	11	12
1	Nino1+2	$Nino12_m$					−	+	+					
2	Nino3	$Nino3_m$				+	−						+	
3	Nino4	$Nino4_m$					+	+					−	
4	Nino3.4	$Nino34_m$				−	+						+	
5	Multivariate ENSO	MEI_m				+	−						+	
6	North Atlantic Oscillation	NAO_m	+	+	−									
7	Atlantic Multidecadal Oscillation	AMO_m				−					+		+	
8	Atlantic Meridional Mode	AMM_m						−	+				+	
9	Lower-level westerly jets EOF1	$LLW1_m$					−	+					+	
10	Lower-level westerly jets EOF2	$LLW2_m$		−	+	+								
11	Lower-level westerly jets EOF3	$LLW3_m$			+	+					−			
12	Mediterranean Sea EOF1	$MSEA1_m$					−		+		+			
13	Mediterranean Sea EOF2	$MSEA2_m$			−		+			+				
14	Mediterranean Sea EOF3	$MSEA3_m$	+	+										−
15	850hPa geo-potential height EOF1	$GHT1_m$								+	−	+		
16	850hPa geo-potential height EOF2	$GHT2_m$								−	+	+		
17	850hPa geo-potential height EOF3	$GHT3_m$	−					+						+
18	Indian Ocean dipole	IOD_m				−	+	+						
19	Atlantic ENSO	$EATL_m$						+	+	−				

[1] The topmost influential season for each variable is marked with a −. [2] 1 = Jan-Feb-Mar, 2 = Feb-Mar-Apr, …, 12 = Dec-Jan-Feb. [3] Subscript m represents the chosen season (i.e., NAO_3: season 3 chosen for NAO).

method operates in smaller parallel executions with low overhead.

2.2 Lasso multivariate regression

Least absolute shrinkage and selection operator (Lasso) multivariate regression is an approach pioneered by Tibshirani (1994) that takes a set of inputs and an outcome measurement and fits a linear model, seeking to shrink the regression and sparsify the predictor feature space. This is achieved by constraining the L^1 norm of the β parameter vector $\boldsymbol{B} = \{\beta_1, \beta_2, \ldots, \beta_n\}$, calculated as in Eq. (1), such that it is no greater than a given s value to be minimized (Tibshirani, 1994).

$$L^1\text{norm} = |\boldsymbol{B}|_1 = \sum_{r=1}^{n} |\beta_r| \,. \tag{1}$$

In the context of this study, this process highlights the prominent phases of the features (Sect. 2.2.1). It derives the temporal phases of predictors lagged behind a response of interest, generating predictor coefficients indicating the magnitude and type of the modulatory relationships with said response (Pendse et al., 2012).

Recent work on inference of modulatory relationships based on Lasso multivariate regression of temporal and spatio-temporal data includes means of improving upon the Lasso methodology. We apply the method proposed by Pendse et al. (2012), given that it incorporates prominent phase detection and significance assessment. Pendse et al. (2012) present an approach toward a data-driven, semi-automatic inference of phenomenological physical models based on the Lasso multivariate regression model, and quantify the influence of key "players" on the response of interest (e.g., Sahel rainfall anomaly) through use of the expected causality impact (ECI) score. The work presented in Pendse et al. (2012) also proposes methods for search space pruning, significance estimation and impact analysis that provide quantifiable metrics in terms of predictors' contributions to the rainfall variability and their probability of detections (PODs).

2.2.1 Prominent phase detection

We employ the methodology suggested by Pendse et al. (2012) to identify the most prominent phases (i.e., seasons) in the data. For the benefit of reproducibility, we utilized the supplemental material provided therein (Pendse et al., 2012). The results obtained by Pendse et al. (2012) were consistent with many well-known modulatory relationships from prior climate knowledge (Chang et al., 2006; Marshall et al., 2001; Sutton et al., 2000). These results complement the ex-

isting physical models and may help climate scientists categorize the correct season for the response of interest (e.g., Sahel rainfall variability). Hence, by leveraging these prominent phases (shown in Table 1), we can focus on features that should have a stronger influence over the response.

2.2.2 Lasso pathway significance assessment

To assess pathway significance, we follow the method described in Pendse et al. (2012). That is, we apply the Monte Carlo method to estimate the statistical significance of the relationships found between the input features and the response in terms of the null hypothesis, by iteratively permuting the response and performing Lasso multivariate regression for these permuted data. This method allows us to prune insignificant edges in the Lasso network, represented by higher p values.

2.2.3 Lasso computational complexity

Given Lasso's iterative nature in finding appropriate λ and β values, the computational complexity of Lasso is reliant on the q parameters and n observations provided by the source data, as it would need to attain solutions for all subsets $\mathcal{M}_k, k \in 1, \ldots, m$ (Meinshausen, 2007). Hence, the computational complexity of this methodology is $\mathcal{O}(n \cdot q \cdot \min\{n, q\})$ (Meinshausen, 2007). Furthermore, since all variables must at some point be evaluated as the Lasso response $r \in q$, this is multiplied by a factor q. However, the q value that affects the actual Lasso execution would also grow smaller, as considerations are made to remove q parameters that temporally cannot modulate r.

2.3 Dynamic Bayesian networks

DBNs expand upon hidden Markov models (HMMs) and Kalman filter models (KFMs), indexing instances of arbitrary variables. DBNs are represented as a structure similar to that of Bayesian networks, with the added benefit of incorporating the temporal space (Dean and Kanazawa, 1989; Murphy, 2002). DBNs are a very popular means with which to mine and represent modulatory relationships in spatial and temporal data, given that the conditional probability distribution of each node can be estimated independently (Friedman et al., 1998; Murphy, 2002; de Kock et al., 2008). The model's dynamicity is obtained by combining a traditional Bayesian network with a temporal Bayesian network that allows for capturing behaviors of the Bayesian network over the temporal space, and is not to be confused with the idea that the model changes over time (Murphy, 2002).

2.3.1 DBN pathway significance assessment

To assess pathway significance, we again apply the Monte Carlo method to estimate the statistical significance of edges representing modulatory relationships. This affects the computational complexity of the methodology, as each random combination must be mined individually. Hence, to alleviate this matter somewhat, we verify only columns for which relationships were found by the base method, and omit the features for which no relationships were found.

2.3.2 DBN computational complexity

Several different implementations of DBN inference exist, each with varying degrees of complexity. To mine the DBNs for our problem, we leverage the toolkit provided by Zou and Feng (2009), built upon the design proposed in Murphy (2002). This toolkit allows us to infer the network structure of the DBNs in $\mathcal{O}(T)$, where T is the length of the sequence to be mined, which could be exponentially large, depending on the number of possible feature combinations a sequence could contain (Murphy, 2002). Given our utilization of the toolkit, we abide by this complexity for our estimation, only restricting the execution by disallowing temporally infeasible edges (i.e., edges are only allowed between two nodes if the originating node occurs temporally before the destination node). In doing so, we ensure that the directionality of the network is temporally sound and fits proper modulatory relationships.

2.4 Construction of modulatory networks

Each of the aforementioned methodologies presents results in a different manner, affecting the interpretability of the information they provide. Hence, the resulting relationships between climate factors should be structured such that all possible modulatory pathways are captured in a comparable context, while preserving the information given by each method.

The results provided by DBN capture a network of relationships between climate factors as a directed acyclic graph (DAG). Such a graph includes a set of vertices and directed edges, which in the context of this study represent the climate factors and the relationships between them, respectively. Furthermore, there are no cycles in the graph (i.e., following a path originating at a given node will never lead back to that node). This structure provides an intuitive visualization of the behaviors in the system, as each edge represents the existence of a relationship between the climate factors it connects. Therefore, we adapt the results provided by CHARM to a similar structure, by building a network where the edges represent each possible combination of high/low anomalies, directed from antecedent to consequent.

However, given that CHARM uses coupled climate indices, we must ensure the the networks generated for the three methods can be interpreted equally. Hence, each Lasso and DBN experiment will also use such coupled data and will also be executed heterogeneously, as described in Sect. 2.1.5. This allows us to use the results provided by DBN directly, given that it already adheres to the proper network structure.

Table 2. Regions for generated indices.

Index	Region
IOD	10° S–10° N, 50–70° E 10° S–0°, 90–110° E
LLW (EOF 1,2,3)	0°–20° N, 60° W–25° E
MSEA (EOF 1,2,3)	30–46° N, 6° W–36° E
GHT (EOF 1,2,3)	0°–40° N, 40° W–30° E
EATL	3° S–3° N, 30° W–0°

As for the results from the Lasso experiments, we generate the network of modulatory pathways by drawing a directed edge from vertex A to B when a β coefficient was found for an execution where B was the response and A was a predictand. Given the temporal window constraints set upon this problem, we can follow the graph backwards from our desired response to study all relationships, both direct and indirect.

2.5 Consensus modulatory network inference via information fusion

To infer a consensus modulatory network for a functional system response, we must combine the modulatory networks inferred by CHARM, Lasso multivariate regression, and DBN into a single unified network that captures the consensus of the results. The field of evolutionary biology has leveraged methods related to information fusion to combine evidence found for specific gene classifications in collected field data (Bailey and Gribskov, 1998; Li et al., 2008). Of the methods in this field, we chose to combine the resulting p values of each edge for each modulatory inference methodology by overlaying the resulting graphs from each methodology upon one another, and performing Fisher's combined probability test, shown in Eq. (2):

$$\chi^2 = -2\sum_{i=1}^{k} \log_e(p_i), \tag{2}$$

where p_i represents the p value for the ith independent test. This score presents a large χ^2 test statistic when p_i values are smaller, suggesting that the null hypotheses are not true for every test. In contrast, when all the null hypotheses are true, and the p_i are independent, χ^2 has a chi-squared distribution with $2k$ degrees of freedom, where k is the number of tests being combined. This can then be used to determine the p value for χ^2 (Fisher, 1932).

After obtaining the combined p value, we compute an ARM-inspired support count to quantify the number of methods providing evidence of this result. With this, we determine which edges are worthwhile of inclusion, opening the realm for climate scientists to determine which amount of evidence constitutes a satisfiable minimum for which an edge is acceptable, and additional information can be obtained

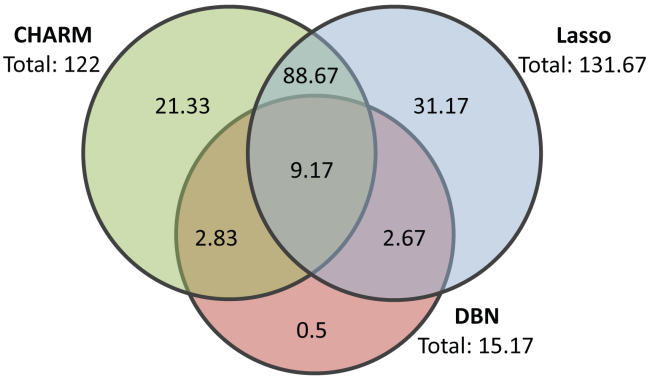

Figure 7. Number of relationships found in network, averaged across coupling inciters.

from the underlying individual results. For example, if ARM found some $A_{HIGH} \rightarrow B_{LOW}$ relationship between features A and B, Lasso or DBN also found evidence of some $A \rightarrow B$ relationship, and we obtain a significant Fisher statistic, we can state that this relationship is founded, since three out of the six possible method results have evidence of such a relationship. Furthermore, given the information provided by ARM's result highlighting specific phases, domain scientists can investigate the $A_{HIGH} \rightarrow B_{LOW}$ relationship in further detail.

We use these statistics to determine a consensus in the relationships found by the methods employed and build a network to capture said consensus. Note that the number of relationships in the consensus result will not be restricted by the methodology that identifies the fewest relationships. Instead, each methodology serves as evidence of the consensus result and affects the strength of the evidence provided for a particular relationship. Hence, each method contributes to the consensus result, with no specific methodology acting as a determining factor that would bias the result towards that specific methodology. Determining a bound at which to remove rules from the consensus result based on the amount of evidence provided is a topic of future work.

3 Results and discussion

Setting the Sahel rainfall anomaly as the system's response presents an ideal model for assessing the climatological relevance of modulatory pathways identified by CHARM. We evaluate the computational validity of CHARM, using the criteria defined in Sect. 2.1.4, by studying the mined rules for the connections it identifies, and compare the results to those of the other approaches and study the effect of combining the rules as described in Sect. 2.5.

3.1 Data

Table 1 presents the data used for this study, which were obtained from the NCEP/NCAR (2014), along with rainfall data obtained from the UDel_AirT_Precip data set provided by NOAA/OAR/ESRL (2014). Along with these, our study uses new indices created by climate scientists (items 8–9) using empirical orthogonal function (EOF) techniques (Wilks, 2006) to isolate the dominant mode(s) in reanalysis data (NOAA/PSD, 2014). The inclusion of the new indices is based on the fundamental knowledge that Sahel climate is modulated by different climatic drivers (Hurrell, 1995; Rowell, 2003).

These drivers originate from the ocean, atmosphere, land surface, and vegetation, where they interact intricately, and ultimately exert a strong influence over the Sahel region. However, the tropical Pacific, Atlantic, and Indian oceans, as well as the Mediterranean Sea and the overlying atmosphere, are key drivers of Sahel climate, so the creation of such indices ensures they are given an equal chance to participate in the experiment, represented in Table 2. Hence, where the climate literature suggested a teleconnection between a given climate variable and Sahel rainfall, but a representative climate index for it was not readily available from NCAR, an EOF analysis of the 850 mb height field was created instead, using reanalysis data. Each mode is represented as feature ⟨#⟩ (i.e., the first mode of variance over the Mediterranean Sea is referred to as MSEA1).

We select the most prominent seasons for these indices, as described in Sect. 2.2.1, and utilize eastern Sahel rainfall over the season July-August-September (JAS) as the desired response. These variable-phase combinations are denoted as feature $_{\langle \text{phase} \rangle}$, where the subscript corresponds to temporal phases (i.e., 1 = Jan-Feb-Mar, 2 = Feb-Mar-Apr, …, 12 = Dec-Jan-Feb). Thus, we can contrast these to provide a description of the sub-region's climate variability in association with this response for the period of 1950–2008.

3.2 Network interpretations

We will discuss specific use cases of our experiment herein. Images for the generated networks for each method, their associated DAG matrices and combination metrics can be obtained via the Supplement[1].

Rainfall interconnections

Table 3 captures the relationships known from reference material, in contrast to the findings of the evaluated methods for the $EATL_8$ coupling inciter, which implies that, for the given data couplings, the coupling inciter used was Atlantic ENSO in the temporal phase of August-September-October. This table serves to present length of pathway from the variable in question to the expected rainfall response, so as to verify the findings of this experiment in terms of known relationships gathered over two decades of research (Tetteh, 2012). The dynamical substance in the processes involved and teleconnections in these mining techniques is highlighted.

[1] http://freescience.org/cs/cni_combined

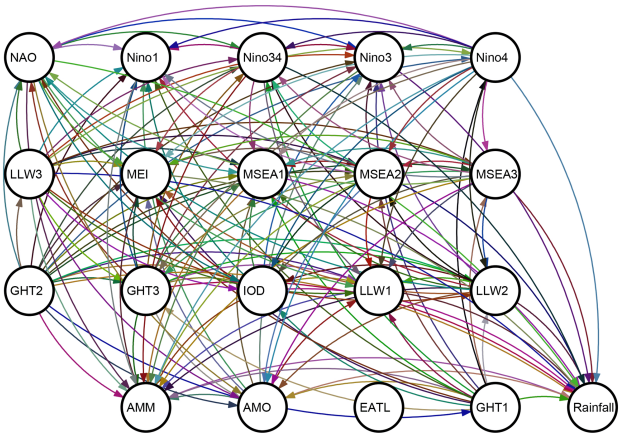

Figure 8. Resulting combined network for coupling inciter $EATL_8$.

Lasso reveals that four oceanic modes, the Pacific (represented by MEI, Nino 3.4 and Nino 4), IOD, MSEA3, and AMO, influence the eastern Sahel rainfall (Rowell, 2003). The dynamical processes inferred from warm ocean surface anomalies associated with the IOD (Lu, 2009), MSEA3 (Rowell, 2003) and AMO (Zhang and Delworth, 2006) are related to an increase in the magnitude of the main rainfall season in the Sahel. The IOD and MSEA3 specifically facilitate positive moisture advection, whereas the AMO displaces the intertropical convergence zone (ITCZ) to its climatological position over the Sahel. These mechanisms are tied to moisture transport from the tropical Atlantic by LLW2 and LLW3. On the contrary, warming of the Pacific is generally associated with rainfall diminution over the Sahel (Janicot et al., 1996).

3.3 Process evaluation

We find that Lasso and CHARM coincide in capturing AMM, the most important oceanic mode governing decadal climate variability of the Sahel, and which primarily determines moisture availability (Grossman and Klotzbach, 2009). The positive (negative) phase of the AMM is associated with rainfall enhancement (suppression). However, its role or impact is modulated directly or indirectly by distinct phases of Nino3, MSEA1 and MSEA2. While the warm (cold) phase of Nino3 suppresses (enhances) moisture flux over the Sahel, MSEA1 and MSEA2 have a competing effect, with positive and negative moisture transport over the Mediterranean Sea, respectively, and are involved in negative and positive moisture transport over the Mediterranean Sea. The model also reveals that the high (low) phase of LLW1 over the Atlantic is associated with strengthening (weakening) of westerly moisture flow. This co-occurs with GHT 1, 2,

Table 3. Comparison of known relationships of climate features with rainfall response with mined network proximity[1,2,3].

				Method								
				CHARM			Lasso			DBN		
No.	Climate variable	Abbr.	⇒	1	2	3	1	2	3	1	2	3
1	Atlantic Meridional Mode	AMM	D	✓			✓					
2	Atlantic Multidecadal Osc.	AMO	D		✓		✓					
3	Atlantic ENSO	EATL	I									
4	Geo-potential height EOF1	GHT1	D	✓			✓					
5	Geo-potential height EOF2	GHT2	D	✓			✓					✓
6	Geo-potential height EOF3	GHT3	D	✓			✓				✓	
7	Indian Ocean dipole	IOD	D		✓		✓					
8	Lower-level w. jets EOF1	LLW1	D	✓			✓					
9	Lower-level w. jets EOF2	LLW2	D		✓		✓					
10	Lower-level w. jets EOF3	LLW3	D		✓		✓					✓
11	Mediterranean Sea EOF1	MSEA1	D	✓			✓					
12	Mediterranean Sea EOF2	MSEA2	D	✓			✓					✓
13	Mediterranean Sea EOF3	MSEA3	D		✓		✓					
14	Multivariate ENSO	MEI	I				✓					
15	Niño1+2	Nino12	I			✓						
16	Niño3	Nino3	I	✓			✓					
17	Niño4	Nino4	I		✓		✓				✓	
18	Niño3.4	Nino34	I		✓		✓					
19	North Atlantic Oscillation	NAO	I		✓		✓			✓		

[1] ⇒: Known relationship, I: indirect, D: direct. [2] References for known relationships by row: 1: Grossman and Klotzbach (2009), 2: Zhang and Delworth (2006), 3: Zebiak (1993), 4–6: Kidson and Newell (1977), 7: Saji et al. (1999), 8–10: Nicholson (2009), 11–13: Rowell (2003), 14–18: Nicholson (1997), and 19: Hurrell (1995). [3] Relationships with EOF modes are unknown, but labels apply for actual climate phenomena.

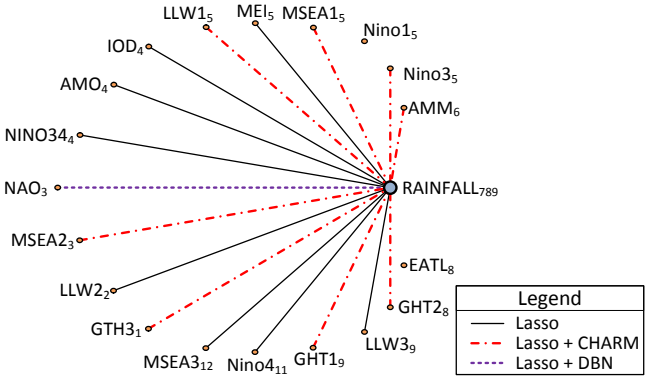

Figure 9. Relationships directly associated with rainfall for $\lambda = \mathrm{EATL}_8$.

and 3, which determine troughs and ridges that govern high and low rainfall anomalies.

Lasso and DBN coincide in capturing extratropical NAO forcing. Although the NAO is known to impact Sahel rainfall (Hurrell, 1995), the mechanism by which this occurs is unclear. A link to the tropical Atlantic, particularly through the LLWs, is suggested by the results here. It is possible that the moisture flux from the tropical Atlantic is dependent on the phase of the NAO. On a finer scale, the model also predicts a direct link to the NAO. The association between the NAO and Sahel rainfall may be multifaceted, and our results are being investigated further by the authors, based on an NAO-driven hypothesis over the entire West African Sahel (Tetteh, 2012).

Figure 7 captures the aggregate number of relationships found by each method, averaged across all coupling inciters studied. We find that, per each coupling inciter, Lasso is the more sensitive methodology, finding edges for most possible feature combinations. Given the design of our CHARM experiment capturing phase-specific relationships (i.e., $A_{\mathrm{HIGH}} \rightarrow B_{\mathrm{LOW}}$), as described in Sect. 2.1.3, we group all possible high/low combinations merely to visualize how many relationships CHARM found as a whole. We note that these highly coincide with the findings of Lasso, but find their share of unique relationships that contribute to the final result. Lastly, we find that DBN produces very few relationships, but the majority of these contribute to the Fisher

Table 4. Network vertex statistics for coupling inciter $EATL_8$.

		Method							
		Betweenness centrality				Clustering coefficient			
No.	Var.	CHARM	Lasso	DBN	Fused	CHARM	Lasso	DBN	Fused
1	AMM	3.7	0	0	1	0.392	0.405	0	0.448
2	AMO	1.367	7.983	1	0	0.414	0.321	0.167	0.446
3	EATL	0	0	0	0	0	0	0	0
4	GHT1	0	0.374	0	0	0.424	0.374	0	0.445
5	GHT2	0	0.338	0	0	0.392	0.338	0.167	0.448
6	GHT3	0.917	0.583	6.5	0	0.413	0.364	0.143	0.448
7	IOD	0	2.233	0	0	0.442	0.346	0	0.446
8	LLW1	1.7	0	0	0	0.412	0.379	0	0.473
9	LLW2	0.417	4.4	8	0	0.438	0.352	0.133	0.448
10	LLW3	0	0.833	0	0	0.419	0.402	0.2	0.445
11	MSEA1	1.7	0	0	0	0.412	0.4	0.333	0.473
12	MSEA2	1.733	5.4	0.5	0	0.393	0.343	0.333	0.445
13	MSEA3	0.617	0.75	1	0	0.415	0.382	0.167	0.448
14	MEI	0	1.983	0	0	0.482	0.352	0	0.473
15	Nino12	0	0.433	0	0	0.449	0.35	0.167	0.468
16	Nino3	1.7	1	0	0	0.412	0.417	0	0.473
17	Nino4	0.617	0.583	2	0	0.415	0.346	0.167	0.448
18	Nino34	1.367	0.167	4	0	0.414	0.429	0	0.446
19	NAO	0.167	0.25	4	0	0.44	0.433	0	0.445
20	Rainfall	0	0	0	0	0.446	0.338	0	0.46

statistic for the three methods, as 97 % of the found relationships coincide with either CHARM or LASSO, while 60 % coincide with both. The central area of Fig. 7 would lead to further study, as it indicates that all three methods provided evidence of relationships in this area.

Given the intent to find drivers for the rainfall feature, Fig. 10 captures the average number of direct relationships to the rainfall response found by each method. This again highlights the sensitivity of Lasso as, of a maximum of 20 features, an average of 17.67 were selected, whilst CHARM and DBN find fewer such relationships. When evaluating coupling inciter $EATL_8$ (see Fig. 9), we see this in further detail as, while Lasso captures 17 direct relationships to the rainfall response, CHARM and DBN capture 8 and 1, respectively. Hence, Lasso appears to detract from discovering indirect relationships, unless β values are inspected directly. This especially affects the fused network, as most features are marked as directly associated with the response (see Fig. 8).

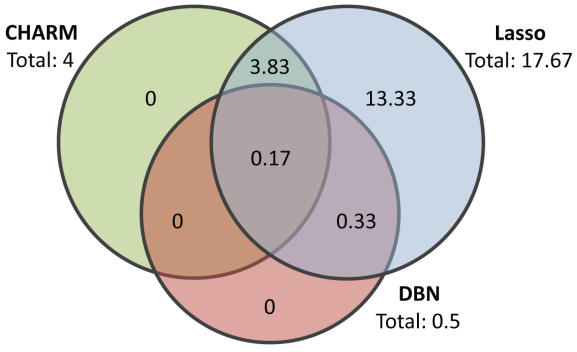

Figure 10. Number of relationships found directly related to rainfall, averaged across coupling inciters.

Fused network relationships

Figure 8 captures the resulting network after the different models are fused into a consensus result, and presents the final set of edges provided by the model. The vast number of edges presented is mostly driven by the high sensitivity of the Lasso methodology and, as mentioned in Sect. 3.3, such a number of direct connections can detract from understanding indirect rainfall relationships. The network metrics captured in Table 4 highlight the fact that AMM has the highest betweenness centrality in the case of CHARM and the fused network, meaning it is found on all the shortest paths in the network, marking AMM's importance in the network, as stated in Sect. 3.2. The clustering coefficients for CHARM and DBN have standard deviations $\sigma = 0.02$ and $\sigma = 0.11$, respectively, but σ equals 0.01 when fused together, indicating that Lasso's sensitivity plays a key role in the fused result; as in the individual Lasso graph, most nodes were hubs for many inbound connections. Furthermore, we find that the fused network performed best at creating a cluster with the desired rainfall response, seemingly best influenced by the CHARM network, although, given Fig. 9, we know the influence of Lasso again came into play. Hence, exploring a

means of limiting Lasso's influence may be a beneficial next step for future work.

4 Conclusions

We evaluated three different methods for finding modulatory relationships in spatio-temporal climate data and validated the results obtained against known relationships modulating rainfall in the Sahel region of western Africa. These results show that each method has its benefits and drawbacks. Noting that significant changes had to take place for utilizing CHARM for this purpose given spatial alignment issues, we devised data coupling as a means with which to study the relationships in the underlying data. These changes served to make CHARM an efficient methodology for addressing the data-driven discovery of predictive, climatologically relevant, and statistically significant modulatory pathways in the physical model of the Sahel rainfall anomaly.

We also evaluated the consensus network obtained after combining the results of these methods via information fusion. In any case, this study served to validate these methods against known relationships from over two decades of hypothesis-driven and first-principles research. The IOD, ENSO, MSEA and AMO were confirmed as important SST anomalies modulating rainfall in the region, as previously discussed in the climate literature. The relationship with the NAO is found to have both direct and indirect components, and is particularly related to equatorial westerlies (LLWs) in the Atlantic, known to influence the region (Nicholson, 2009). It is hypothesized that the NAO modulates the position and strength of the equatorial westerlies, impacting the Tropical Easterly Jet and therefore Sahel rainfall. This hypothesis is currently under investigation by climate domain scientists (i.e., Tetteh, 2012), based on the results of this study, and serves as an example of a relationship that is not fully understood being highlighted by the framework presented here.

Future work

In this work, we used the same set of climate indices for all the individual methodologies employed, to facilitate the comparison and fusion of results. It is possible that using different data sets or different time/spatial series for each methodology can improve their individual results, and in turn the overall outcome of the models. However, additional considerations would be needed for interpreting which models provide evidence of particular relationships between individual climate indices.

Furthermore, the presented CHARM approach studies rules heterogeneously, which is for the benefit of understanding rules with significant support within particular coupling inciters. However, future work would include addressing rules across multiple inciters, given the geophysical nature of the underlying data and understanding that climate relationships are in reality affected by multiple inciters.

Additionally, given our findings regarding Lasso's sensitivity to finding relationships at varying beta magnitudes, future work will be directed towards limiting such sensitivity and/or its influence on the fused network.

Finally, some of the modulatory relationships identified by these methods may represent underlying causal pathways in the climate system. Future work will also focus on inferring these causal pathways by leveraging causal modeling frameworks, such as causal Bayesian networks. Under this framework, inferring causal relationships becomes a problem of network structure learning. Several score-based and constraint-based algorithms have been proposed to this end (Spirtes, 2010). However, due to the inherent complexity of the climate system, learning this causal network structure is not a simple task. Future work should include identifying an appropriate causal inference algorithm for the problem at hand, by determining which underlying assumptions must hold to infer causal models for the climate domain. This causal inference algorithm should not assume causal sufficiency or acyclicity of the causal structure (Hyttinen et al., 2013), since latent variables (i.e., confounders) and feedback loops are ubiquitous in the climate system. This algorithm should also be able to handle the high dimensionality and small sample size of climate data (Bühlmann, 2013). Furthermore, an algorithm that allows one to incorporate prior knowledge (i.e., known causal relationships from domain knowledge) would also be desirable (Borboudakis et al., 2011).

Acknowledgements. This work was supported in part by the US Department of Energy, Office of Science, the Office of Advanced Scientific Computing Research (SDAVI Institute), and the US National Science Foundation (Expeditions in Computing program). Oak Ridge National Laboratory is managed by UT-Battelle for the LLC US D.O.E. under contract no. DEAC05-00OR22725. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Edited by: V. Mishra
Reviewed by: D. Erickson and one anonymous referee

References

Agrawal, R. and Srikant, R.: Fast algorithms for mining association rules in large databases, in: VLDB 1994, edited by: Bocca, J., Jarke, M., and Zaniolo, C., 487–499, 1994.

Agrawal, R., Imieliński, T., and Swami, A.: Mining association rules between sets of items in large databases, Sigmod Record, 22, 207–216, 1993.

Bailey, T. L. and Gribskov, M.: Combining evidence using p-values: application to sequence homology searches, Bioinformatics, 14, 48–54, 1998.

Borboudakis, G., Triantafilou, S., Lagani, V., and Tsamardinos, I.: A constraint-based approach to incorporate prior knowledge in causal models, in: ESANN'2011, 2011.

Bühlmann, P.: Causal statistical inference in high dimensions, Math. Meth. of OR, 77, 357–370, 2013.

Chang, P., Yamagata, T., Schopf, P., Behera, S. K., Carton, J., Kessler, W. S., Meyers, G., Qu, T., Schott, F., Shetye, S., and Xie, S.-P.: Climate Fluctuations of Tropical Coupled Systems–The Role of Ocean Dynamics, J. Climate, 19, 5122–5174, 2006.

de Kock, M., Le, H., Tadross, M., and Potgeiter, A.: Weather Forecasting Using Dynamic Bayesian Networks, Tech. Rep., University of Cape Town, 2008.

Dean, T. and Kanazawa, K.: A Model for Reasoning About Persistence and Causation, Tech. Rep., Brown University, Providence, RI, USA, 1989.

Fisher, R. A.: Statistical Methods for Research Workers, Oliver & Boyd, Edinburgh, 1932.

Friedman, N., Murphy, K., and Russell, S.: Learning the structure of dynamic probabilistic networks, in: UAI'98, edited by:P Cooper, G. and Moral, S., 139–147, 1998.

Gonçalves, G.: Analysis of interpolation errors in urban digital surface models created from Lidar data, 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, edited by: Caetano, M. and Painho, M., 160–168, 2006.

Gonzalez, D. L., Pendse, S. V., Padmanabhan, K., Angus, M. P., Tetteh, I. K., Srinivas, S., Villanes, A., Semazzi, F., Kumar, V., and Samatova, N. F.: Coupled Heterogeneous Association Rule Mining (CHARM): Application toward Inference of Modulatory Climate Relationships, in: 2013 IEEE 13th International Conference on Data Mining (ICDM'13), 1055–1060, IEEE, 2013.

Grossman, I. and Klotzbach, P.: A review of North Atlantic modes of natural variability and their driving mechanisms, J. Geophys. Res.-Atmos., 114, D24107, doi:10.1029/2009JD012728, 2009.

Hallett, T., Coulson, T., Pilkington, J., Clutton, T., Pemberton, J., and Grenfell, B.: Why large-scale climate indices seem to predict ecological processes better than local weather, Nature, 430, 71–75, 2004.

Havlin, S., Kenett, D., Ben-Jacob, E., Bunde, A., Cohen, R., Hermann, H., Kantelhardt, J., Kertész, J., Kirkpatrick, S., Kurths, J., Portugali, J., and Solomon, S.: Challenges in network science: Applications to infrastructures, climate, social systems and economics, The Eur. Phys. J. Special Top., 214, 273–293, 2012.

Huang, Y., Zhang, L., and Zhang, P.: A Framework for Mining Sequential Patterns from Spatio-Temporal Event Data Sets, IEEE T. Knowl. Data En., 20, 433–448, 2008.

Hurrell, J.: Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation, Science, 269, 676–679, 1995.

Hyttinen, A., Hoyer, P. O., Eberhardt, F., and Järvisalo, M.: Discovering Cyclic Causal Models with Latent Variables: A General SAT-Based Procedure, CoRR, abs/1309.6836, 2013.

Janicot, S., Moron, V., and Fontaine, B.: Sahel droughts and ENSO dynamics, Geophys. Res. Lett., 23, 515–518, 1996.

Janicot, S., Trzaska, S., and Poccard, I.: Summer Sahel-ENSO teleconnection and decadal time scale SST variations, Clim. Dynam., 18, 303–320, 2001.

Kidson, J. W. and Newell, R. E.: African rainfall and its relation to the upper air circulation, Q. J. Roy. Meteor. Soc., 103, 441–456, 1977.

Li, X., Ren, Q., Weng, Y., Cai, H., Zhu, Y., and Zhang, Y.: SCG-Pred: A Score-based Method for Gene Structure Prediction by Combining Multiple Sources of Evidence, Genomics, Proteom. Bioinform., 6, 175–185, 2008.

Lu, J.: The dynamics of the Indian Ocean sea surface temperature forcing of Sahel drought, Clim. Dynam., 33, 445–460, 2009.

Manatsa, D., Chipindu, B., and K., B. S.: Shifts in IOD and their impacts East Africa rainfall, Theor. Appl. Climatol., 110, 115–128, 2012.

Marshall, J., Kushnir, Y., Battisti, D., Chang, P., Czaja, A., Dickson, R., Hurrell, J., McCartney, M., Saravanan, R., and Visbeck, M.: North Atlantic climate variability: phenomena, impacts and mechanisms, Int. J. Climatol., 21, 1863–1898, 2001.

Meinshausen, N.: Relaxed Lasso, Comput. Stat. Data An., 52, 374–393, 2007.

Mortimore, M. J. and Adams, W. M.: Farmer adaptation, change and 'crisis' in the Sahel, Global Environ. Change, 11, 49–57, 2001.

Murphy, K. P.: Dynamic Bayesian Networks: Representation, Inference and Learning, PhD thesis, University Of California, Berkeley, 2002.

NCEP/NCAR: NCEP/NCAR Climate Index TimeSeries Data, available at: http://esrl.noaa.gov/psd/data/climateindices/list, last access: 1 February 2014.

Nicholson, S. E.: An Analysis of the ENSO Signal in the Tropical Atlantic and Western Indian Oceans, Int. J. Climatol., 17, 345–375, 1997.

Nicholson, S. E.: On the factors modulating the intensity of the tropical rainbelt over West Africa, Int. J. Climatol., 29, 673–689, 2009.

NOAA: NOAA Glossary, available at: http://nws.noaa.gov/climate/help/glossary.php, last access: 1 February 2014.

NOAA/OAR/ESRL: NOAA/OAR/ESRL PSD, available at: http://esrl.noaa.gov/psd/data/gridded/data.UDel_AirT_Precip.html, last access: 1 February 2014

Pendse, S., Tetteh, I., Semazzi, F., Kumar, V., and Samatova, N.: Toward Data-driven, Semi-automatic Inference of Phenomenological Physical Models: Application to Eastern Sahel Rainfall, in: SDM '12, edited by: Ghosh, J., Liu, H., Davidson, I., Domencioni, C., and Kamath, C., 2012.

Polo, I., Rodríguez, B., Losada, T., and García, J.: Tropical Atlantic Variability Modes (1979-2002): Time-Evolving SST Modes Related to West African Rainfall, J. Climate, 21, 6457–6475, 2008.

NOAA/PSD: NOAA PSD Reanalysis data, available at: http://esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml, last access: 1 February 2014.

Rowell, D. P.: The Impact of Mediterranean SSTs on the Sahelian Rainfall Season., J. Climate, 16, 849–862, 2003.

Saji, N. H., Goswami, B. N., Vinayachandran, P. N., and Yamagata, T.: A dipole mode in the tropical Indian Ocean, Nature, 401, 360–363, 1999.

Schreck, C. J. and Semazzi, F. H.: Variability of the recent climate of eastern Africa, Int. J. Climatol., 24, 681–701, 2004.

Spirtes, P.: Introduction to Causal Inference, J. Mach. Learn. Res., 11, 1643–1662, available at: http://dl.acm.org/citation.cfm?id=1756006.1859905 (last access: 1 February 2014), 2010.

Sultan, B., Labadi, K., Guégan, J.-F., and Janicot, S.: Climate drives the meningitis epidemics onset in West Africa, PLoS medicine, 2, E6–E6, 2005.

Sutton, R. T., Jewson, S. P., and Rowell, D. P.: The Elements of Climate Variability in the Tropical Atlantic Region., J. Climate, 13, 3261–3284, 2000.

Tan, P., Steinbach, M., Kumar, V., Potter, C., Klooster, C., and Torregrosa, A.: Finding Spatio-Temporal Patterns in Earth Science Data, KDD 2001 Workshop on Temporal Data Mining, 26 August 2001, San Francisco, CA, 2001.

Tan, P., Steinback, M., and Kumar, V.: Association Analysis: Basic Concepts and Algorithms, Introduction to Data Mining, Pearson, 327–396, 2006.

Tetteh, I. K.: West African Seasonal Climate Variability and Predictability, PhD thesis, North Carolina State University, 2012.

Tibshirani, R.: Regression shrinkage and selection via the Lasso, J. Roy. Stat. Soc. B, 58, 267–288, 1994.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences: An Introduction, Elsevier, 2006.

Zaki, M.: Generating non-redundant association rules, in: KDD '00, edited by: Simoff, S. and Zaïane, O., 34–43, ACM, 2000.

Zebiak, S. E.: Air-Sea Interaction in the Equatorial Atlantic Region, J. Climate, 6, 1567–1586, 1993.

Zhang, R. and Delworth, T. L.: Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes, J. Geophys. Res., 33, L17712, doi:10.1029/2006GL026267, 2006.

Zou, C. and Feng, J.: Granger causality vs. dynamic Bayesian network inference: a comparative study, BMC Bioinform., 10, 1–17, 2009.