Data-driven discovery of modulatory factors for African rainfall variability

Michael Angus **Gonzalo Bello** **Mandar Chaudhary**

North Carolina State University

Fifth Workshop on Understanding Climate Change from Data, August 4th, 2015

Response-Guided Community Detection:
Application to Climate Index Discovery

Toward Discovery of Key Factors Casually Affecting Climate Extremes:
Application to African Sahel Rainfall Anomaly Forecasts

Response-Guided Community Detection:
Application to Climate Index Discovery

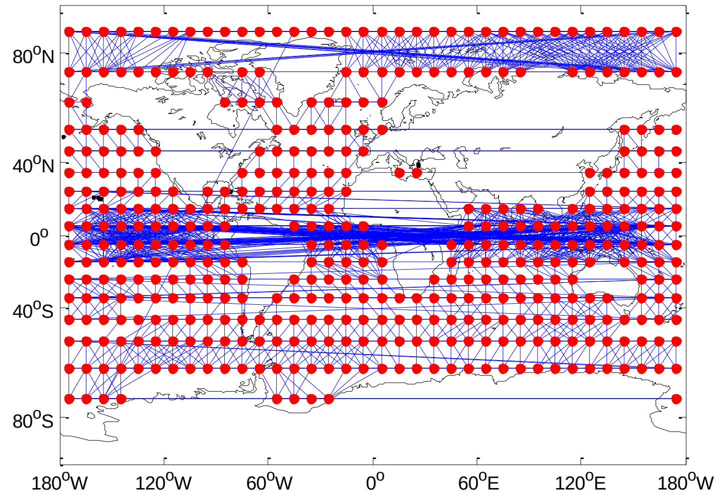
Gonzalo A. Bello[1]

Michael Angus[1] **Navya Pedemane**[1] **Jitendra K. Harlalka**[1]
Fredrick H. M. Semazzi[1] **Vipin Kumar**[2] **Nagiza F. Samatova**[1,3]

[1] North Carolina State University, Raleigh, NC, USA
[2] University of Minnesota Twin Cities, Minneapolis, MN, USA
[3] Oak Ridge National Laboratory, Oak Ridge, TN, USA

Introduction

- **Climate networks** have been adopted to model climate data [3,5,6]. **Communities** in these networks represent potential **climate indices** [5].
- **Community detection** techniques are traditionally unsupervised learning methods, and thus the communities identified may not be associated with the response variable of interest.



Example of a climate network

Problem Statement

- Introduced the problem of **response-guided community detection**:

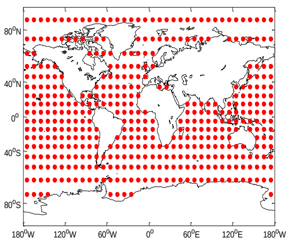
Identifying communities in a graph associated with a response variable of interest by explicitly incorporating information of this response variable during the community detection process.

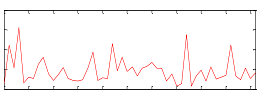
- Studied the application of response-guided community detection to the task of **climate index discovery**—specifically, to the discovery of climate indices associated with **seasonal rainfall variability in the Greater Horn of Africa** (GHA).

Proposed Methodology for Climate Index Discovery

Input

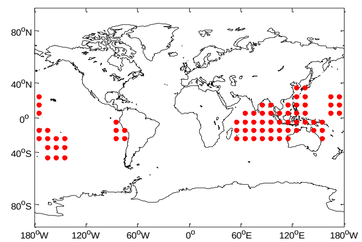
Multivariate spatiotemporal data set



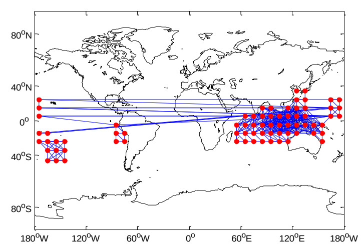


Response variable of interest

Response-guided network construction

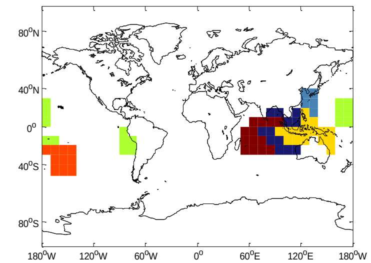


Select spatial points

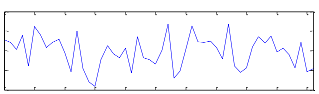


Construct network

Response-guided community detection



Identify comm.

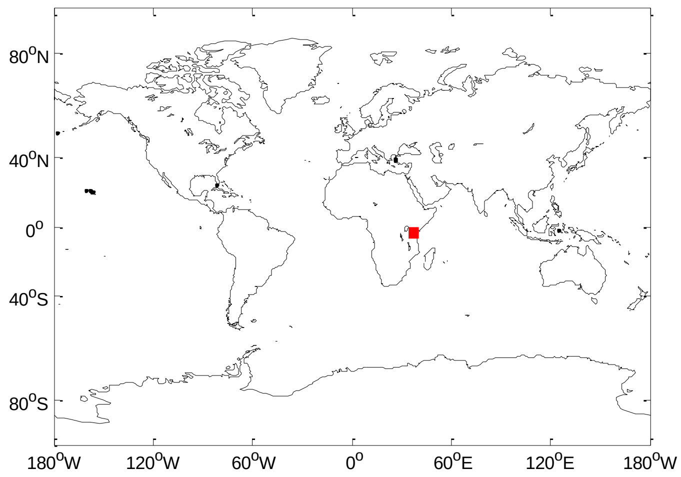


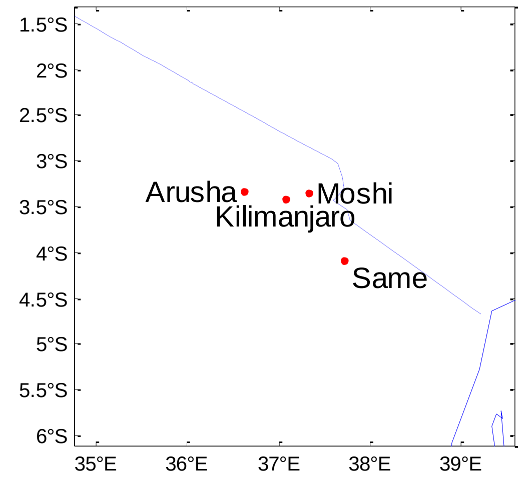
Climate indices

Output

Experimental Evaluation

- Applied the proposed methodology to the discovery of climate indices associated with **October to December** (OND) seasonal rainfall in the GHA.
- Used data from four highly correlated stations in the North Eastern Highlands of Tanzania: **Arusha**, **Kilimanjaro**, **Moshi**, and **Same**.





Location of GHA stations used in the experimental study

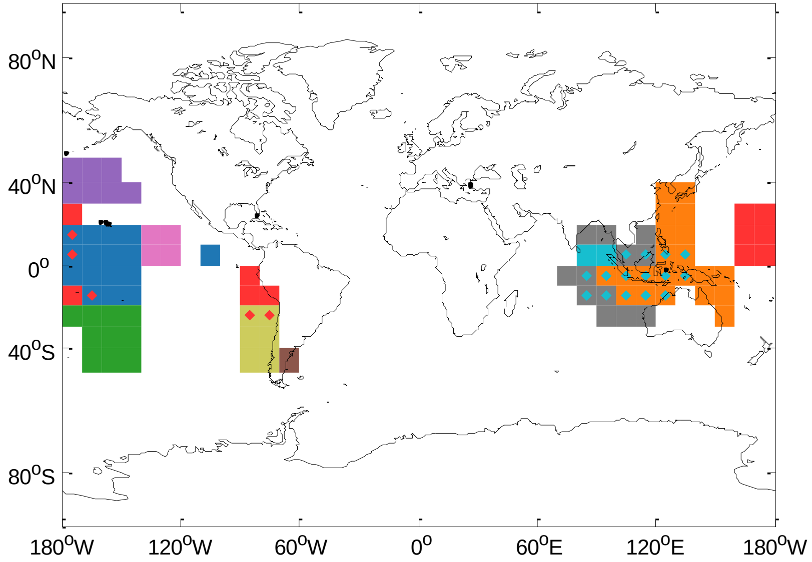
Data Description

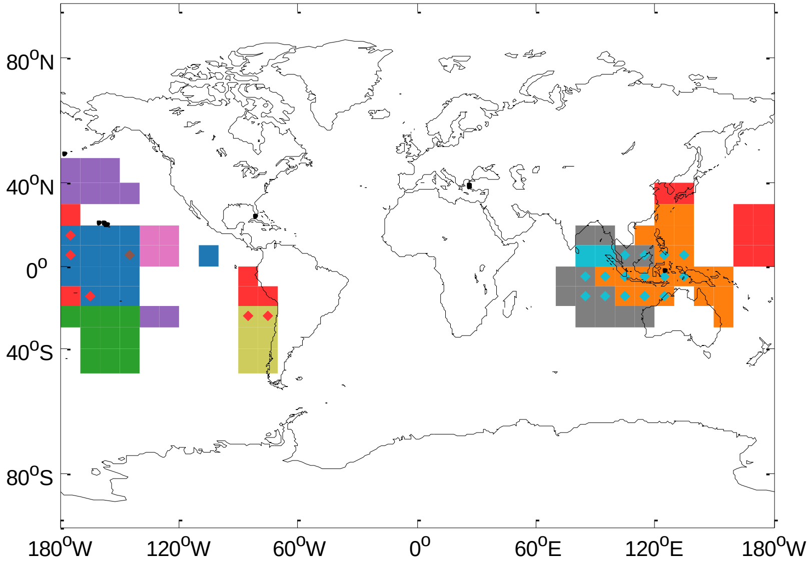
- Monthly precipitation data for stations in Tanzania from 1960 to 2011 (52 years) provided by the Tanzania Meteorological Agency.
  - Data was divided into training set (1960 to 1997) and test set (1998 to 2011).
- Monthly gridded ocean data for the following climate variables:
  - Sea Surface Temperature (SST) — NOAA ERSST V3 data set
  - Sea Level Pressure (SLP)
  - Geopotential Height at 500 mb (GH)
  - Relative Humidity at 850 mb (RH)
  - Precipitable Water (PW)

  NCEP/NCAR Reanalysis 1 data set
- Data was **normalized** using monthly z-scores transformations and **linearly detrended**.

Climate Indices Discovered

- Discovered climate indices using the proposed methodology with OND seasonal rainfall in the GHA as the response variable of interest.



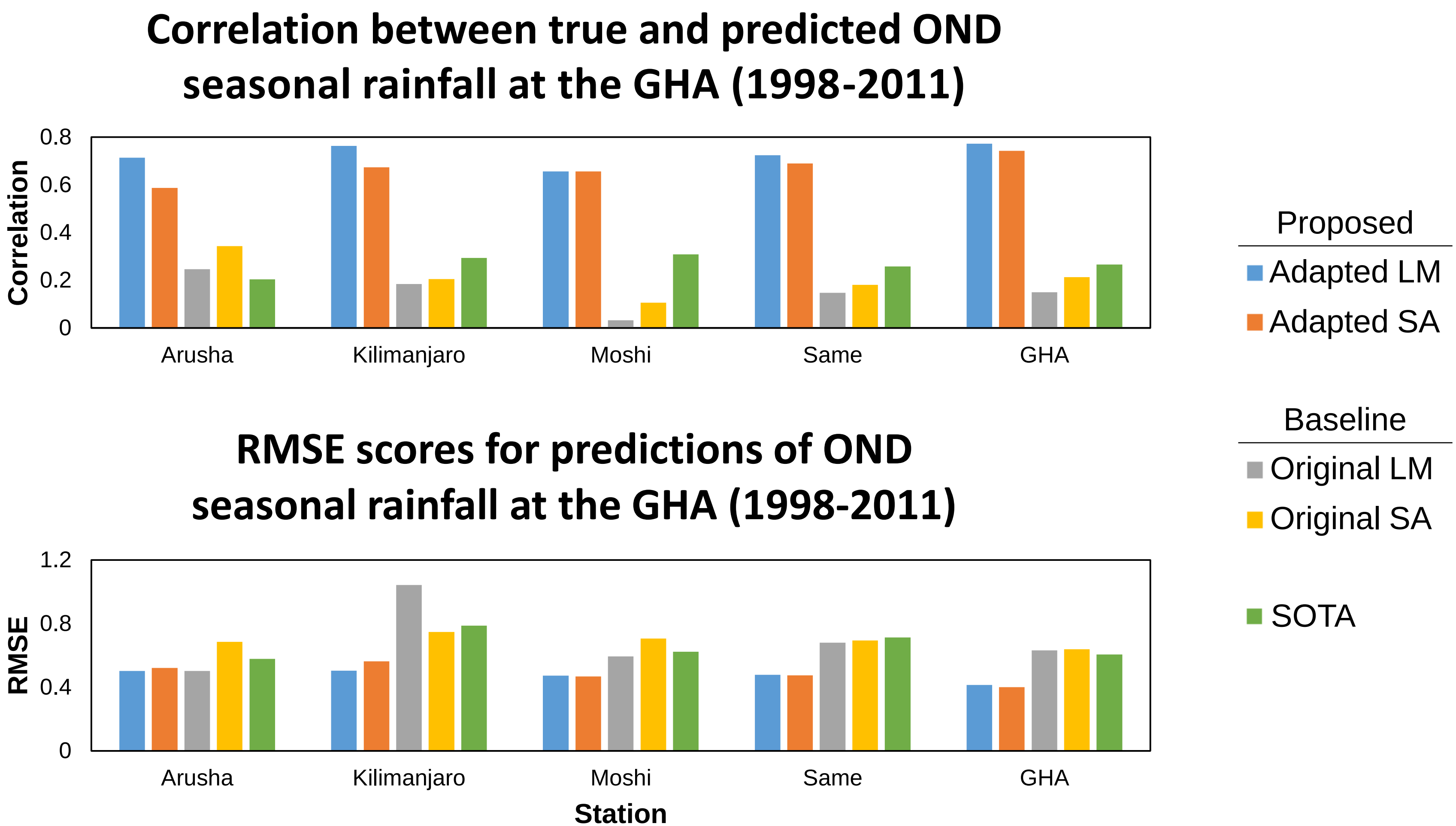


Climate indices discovered using the proposed methodology with two well-known community detection algorithms adapted for response-guided community detection: the Louvain method (LM) (left) [1] and simulated annealing (SA) (right) [4].

Prediction of OND Seasonal Rainfall in the GHA

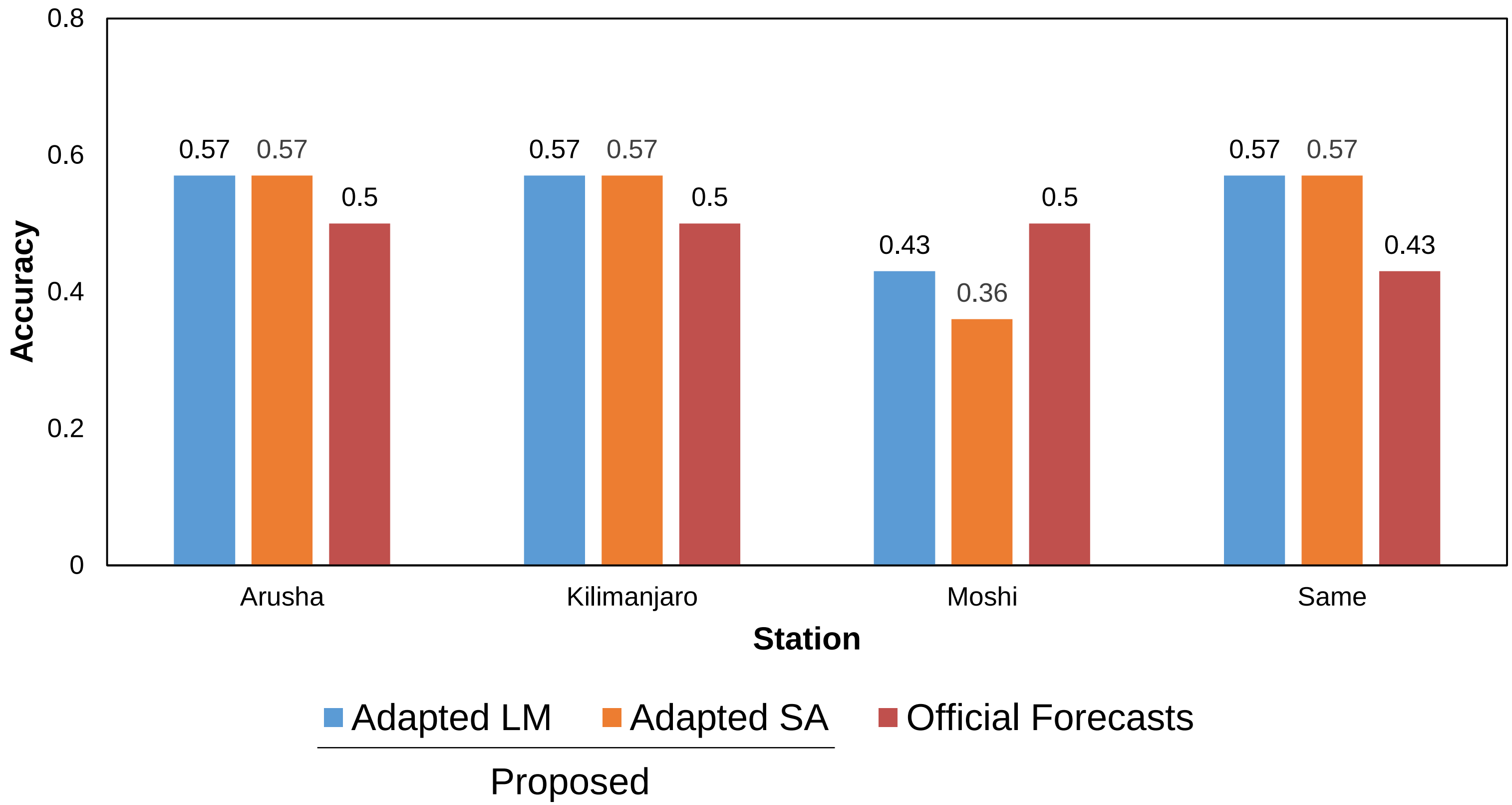
- Trained **linear regression** models to predict OND rainfall at each station and average OND rainfall at the GHA.
- Trained **decision trees** to classify the OND rainfall season at each station as *below normal*, *normal*, or *above normal*.
- The **top climate indices** with the highest absolute correlation with OND rainfall at the GHA over the training set were used as predictors.
- Experiments were performed using data up to August.
- Compared results with those obtained using:
  - Baseline methodology.
  - State-of-the-art methodology [5].
  - **Official forecasts** issued by the Tanzania Meteorological Agency.

Results of Predictions of OND Seasonal Rainfall in the GHA from 1998 to 2011



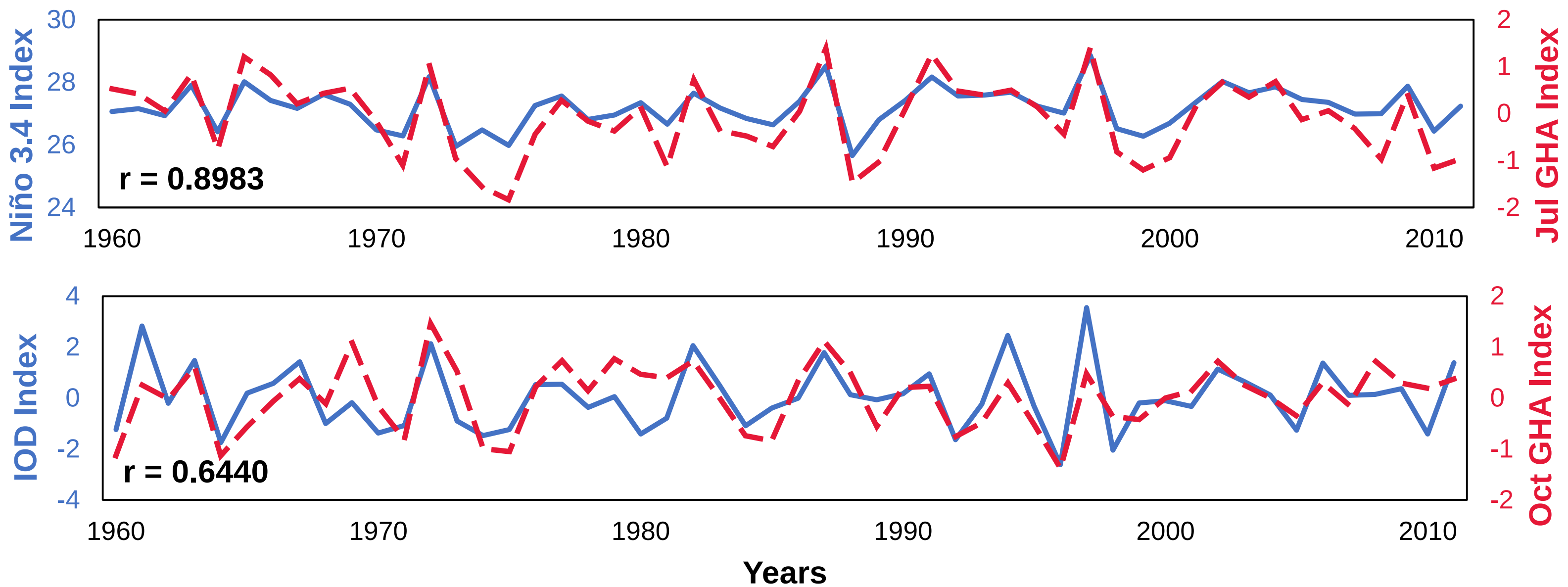
Results of Predictions of the OND Rainfall Season in the GHA from 1998 to 2011

Classification accuracy of predictions of the OND rainfall season at the GHA (1998-2011)



Physical Interpretation of Climate Indices Discovered

- Discovered climate indices significantly correlated ($p < 0.01$) with El Niño Southern Oscillation (**ENSO**) and the Indian Ocean Dipole (**IOD**), which are known to be associated with OND rainfall variability in the GHA [2].



Time series and linear correlation of the Niño 3.4 index (upper) and the IOD (lower) with climate indices discovered using the proposed methodology

Conclusions

- Introduced the problem of **response-guided community detection**.
- Proposed a methodology for the **discovery of climate indices** from multivariate spatiotemporal data using response-guided community detection.
- The predictions obtained using the climate indices discovered show that the proposed methodology **improves the forecast skill for the response variable of interest** with respect to existing methodologies and official forecasts.
- The climatological relevance of the climate indices discovered is supported by domain knowledge, which suggests that the proposed methodology is able to **capture the underlying patterns known to be associated with the response variable of interest**.

References

1. V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
2. J. H. Bowden and F. H. M. Semazzi. Empirical Analysis of Intraseasonal Climate Variability over the Greater Horn of Africa. *Journal of Climate*, 20(23): 5715-5731, 2007.
3. J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal-Special Topics*, 174(1):157–179, 2009.
4. R. Guimerà and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028): 895-900, 2005.
5. K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly. 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.
6. A. A. Tsonis and P. J. Roebber. The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications*, 333:497–504, 2004.

Toward Discovery of Key Factors Casually Affecting Climate Extremes:

Application to African Sahel Rainfall Anomaly Forecasts

Mandar S. Chaudhary[1]

Michael Angus[1]
Fredrick H. M. Semazzi[1]

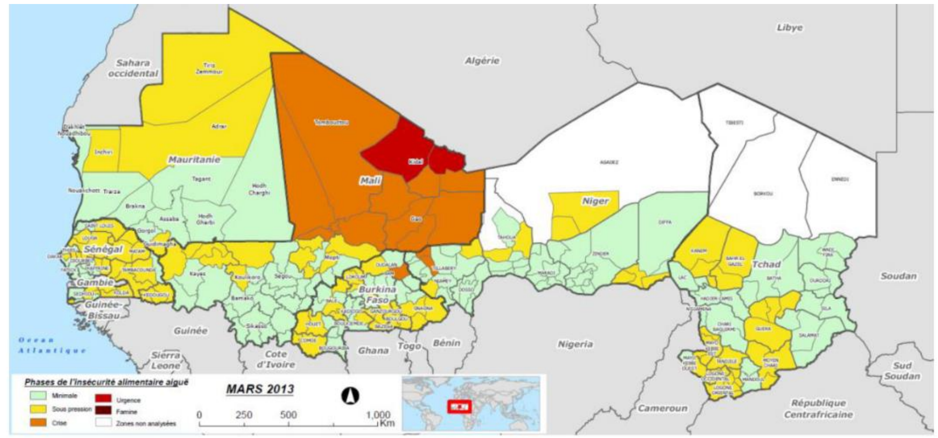
Doel González[1]
Vipin Kumar[2]

Gonzalo A. Bello[1]
Nagiza F. Samatova[1,3]

[1] North Carolina State University, Raleigh, NC, USA
[2] University of Minnesota Twin Cities, Minneapolis, MN, USA
[3] Oak Ridge National Laboratory, Oak Ridge, TN, USA

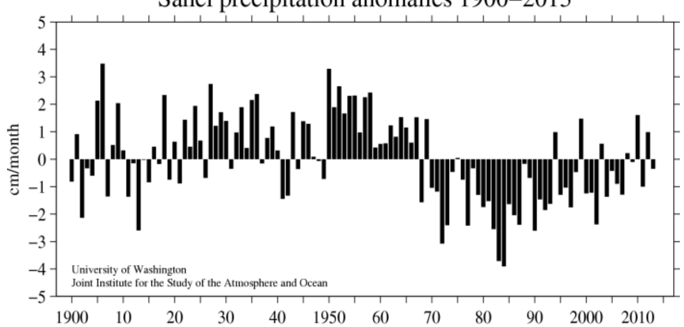
Motivation: Sahel drought

- Drought in the Sahel region during 2013, shown right, despite a modest recovery in rainfall since the persistent 1980 droughts.



Food and Agricultural Organization of the United Nations "Food Security and Humanitarian Implications in West Africa and the Sahel" available from: http://www.fao.org/fileadmin/userupload/emergencies/docs/FAO-WFP%20Joint%20Note%20-%20July%202013.pdf

Sahel precipitation anomalies 1900–2013



June through October averages over 20–10N, 20W–10E. 1900–2013 climatology
NOAA NCDC Global Historical Climatology Network data

http://research.jisao.washington.edu/data_sets/sahel/

- High level of interest in the climate community, as forcing by large scale climate patterns hints at possible predictability.

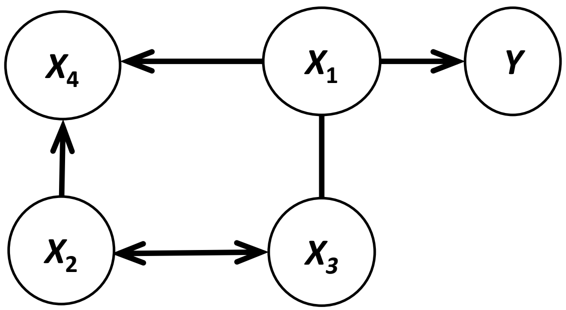
Problem Statement

- The problem of **feature discovery** is defined as,

  Given a causal graph $G=(V, E)$, select potential causal relations D', and estimate a set of stable causal effects Θ, such that,

  - $V = \{X_1, X_2, \ldots, X_p, Y\}$ and E is a collection of directed edges, undirected edges and bi-directed edges.
  - $X_i \rightarrow X_j \in D'$ such that $X_i \in V \backslash \{Y\}$, $X_j \in V$ and $D' \subseteq E$.

  $\forall\ X_i \rightarrow X_j$ (or $X_i \rightarrow Y$) estimate stable causal effect of X_i and X_j (or X_i) on Y, denoted by θ_i and θ_j respectively. If they exist, then construct a new feature space f_{new}, that improves the forecast performance.



A causal graph G

Given: $E = \{X_1 \rightarrow X_4, X_2 \rightarrow X_4, X_1 \rightarrow Y, X_1 - X_3, X_2 \leftrightarrow X_3\}$

Step 1: $D' = \{X_1 \rightarrow X_4, X_2 \rightarrow X_4, X_1 \rightarrow Y\}$,

Step 2: $\Theta = \{(\theta_1, \theta_4), (\theta_1)\}$

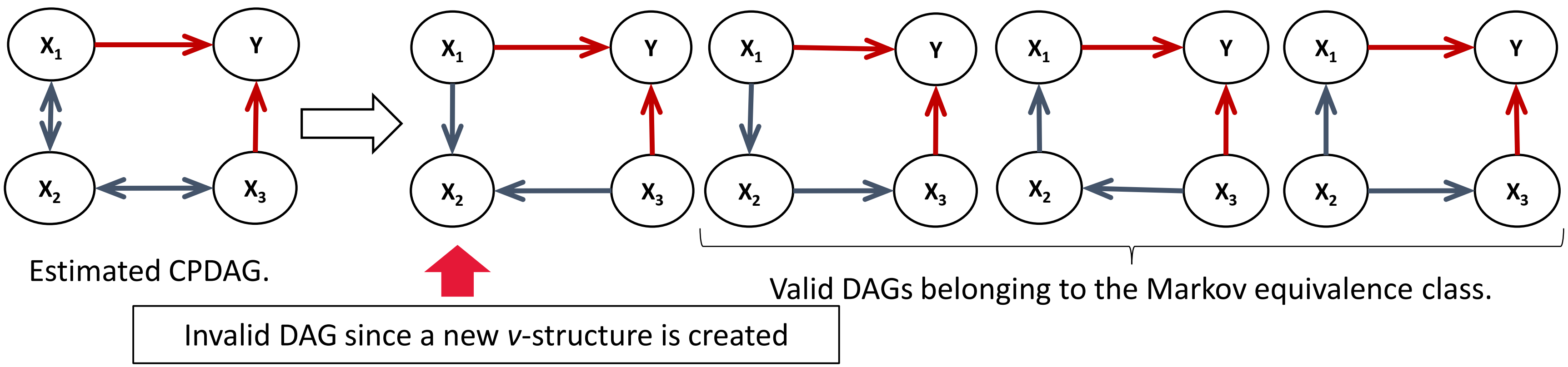
Step 3: $f_{new} = \{f_{X1 \rightarrow X4}, f_{X1 \rightarrow Y}\}$

Data Preparation

- Multivariate Time Series (MTS) Data (1951-2007):
  - 30 climate indices from NOAA ESRL,
  - 2 sea surface temperature indices constructed using COBE SST data provided by NOAA ESRL, and
  - 2 climate indices created using NCEP/NCAR Reanalysis data [5].
  - All indices are collected over 6-month period (Jan-Jun).
- Sahel rainfall data (Jul-Aug-Sep)
  - GPCC Precipitation Full V6 (0.5x0.5) data made available by NOAA Earth System Research Laboratory.
  - Latitude: 10N-20N, Longitude: 20W - 35 E.
- All the indices are detrended and normalized using their *z*-scores, while the monthly values of Sahel rainfall season are linearly detrended, aggregated by summing the monthly detrended values into a vector and then normalized using its *z*-score.

Select Potential Causal Relations

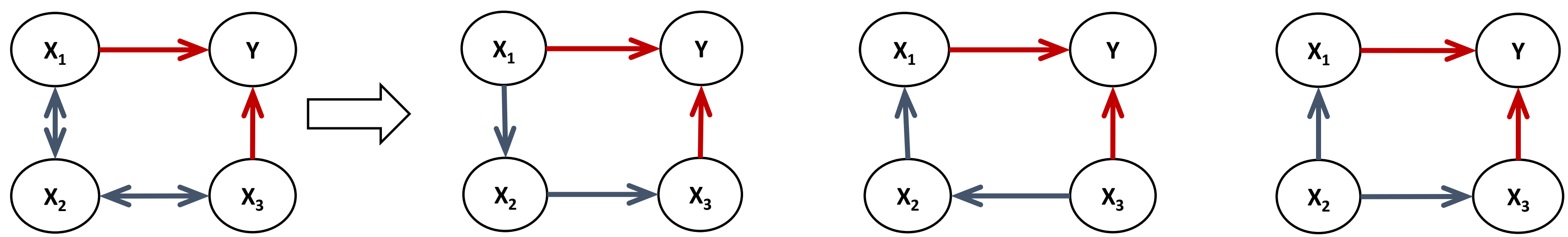
- The **PC-stable** algorithm [2] incorporated with constraints to prevent temporally incoherent causal relations, was used to estimate a **completed partially directed acyclic graph (CPDAG)**, which represents the **Markov equivalence class** of the true causal graph.



- The directed edges, $\boldsymbol{D}$ $=\{X_1 \rightarrow Y, X_2 \rightarrow Y\}$ represent **persistent potential causal relations** across all the valid DAGs, we consider only the edges in $\boldsymbol{D}$ for feature discovery.

Estimate Causal Effects

- For each edge $X_i \rightarrow X_j$ (or $X_i \rightarrow Y$) ∈ $\boldsymbol{D}$, estimate the causal effect of the variables, $\{X_i, X_j\}$ (or X_i) on the response variable, Y across all the Markov equivalent DAGs using the **IDA** (**I**ntervention calculus when **D**AG is **A**bsent) method [1].
  - For example, given $X_1 \rightarrow Y$, estimate the causal effect of X_1 on Y.



Estimated CPDAG

DAG 1: $pa_1 = \{\}$

DAG 2: $pa_1 = \{X_2\}$

DAG 3: $pa_1 = \{X_2\}$

$$Y = \beta_{11}X_1 + \epsilon_1$$

$$Y = \beta_{12}X_1 + \beta_{22}X_2 + \epsilon_2$$

$$Y = \beta_{13}X_1 + \beta_{23}X_2 + \epsilon_2$$

$$B = \{\beta_{11}, \beta_{12}, \beta_{13}\}$$

Multi-set of causal effects

Estimate Causal Effects by Addressing Multicollinearity

- Due to the presence of **multicollinearity** in the dataset, the estimated causal effects from the linear regression models are not reliable.
- To address this issue, we replace linear regression with **Principal Component Regression (*PCR*)** [3] for each DAG,
  - Performs Singular Value Decomposition (SVD) on the predictor set $\boldsymbol{X'}$=[X_1 pa_1] and regress Y on score matrix obtained as follows,

$$\boldsymbol{SVD([X_1\ pa_1]) = UD}\mathbf{V}^\top \boldsymbol{= T}\mathbf{P}^\top$$

$$\boldsymbol{Y = \beta_T T + \epsilon_0}$$

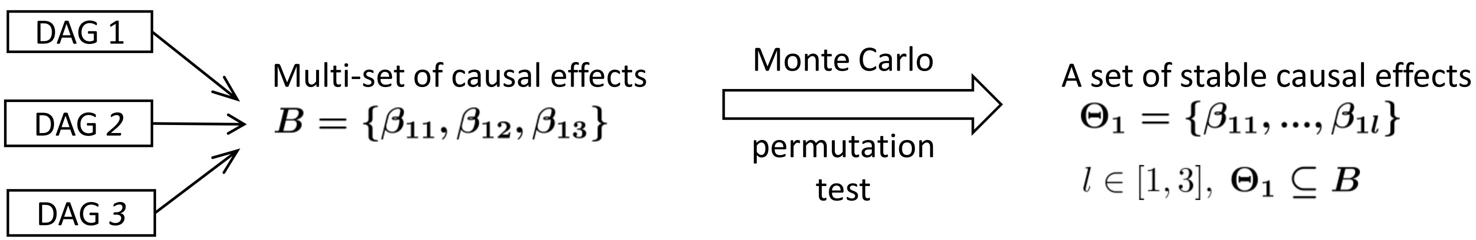
    - β_T={$\beta^1{}_T$, $\beta^2{}_T$, …, $\beta^n{}_T$} contains the regression coefficients of n principal components.
  - Select the regression coefficient of the principal component which captures the maximum variance of $\boldsymbol{X'}$.

$$\boldsymbol{\beta_{X'} = P^m \cdot \beta_T^m}$$

$$\boldsymbol{\beta_{X'} = \{\beta_{11}, \beta_{pa_{11}}\}}$$

Assess Stability of Estimated Causal Effect

- Assess the stability of an estimated causal effect



  - *p*-value: measure the number of times the absolute value of the randomized causal effect is greater than or equal to the estimated causal effect.
  - The estimated causal effect is **stable**, if *p*-value ≤ 0.05.

– Select the causal effect $\beta_1 \in \Theta_1$ such that, $|\beta_1| = min|\Theta_1|$.

– For any $X_k \in \{X_k \rightarrow X_j, X_j \rightarrow X_k$, or $X_k \rightarrow Y\}$, if Θ_k is empty, then we discard the directed edge.

– Otherwise, we store the directed edge in D' and the stable causal effects of indices in the edge in Θ'.

Construct a New Feature Space

- Weighted linear combination of directed edges
  - There are two kinds of directed edges observed in D', (i) $X_i \rightarrow X_j$ and (ii) $X_k \rightarrow Y$, we construct new features from these edges as follows,
    - $X_i \rightarrow X_j$, and $(\beta_i, \beta_j) \in \Theta'$, a new feature is constructed as,

$$\boldsymbol{f_{X_i \rightarrow X_j} = \beta_i \cdot X_i + \beta_j \cdot X_j}$$

    - $X_k \rightarrow Y$, and $\beta_k \in \Theta'$, the corresponding new feature is,

$$\boldsymbol{f_{X_k \rightarrow Y} = \beta_k \cdot X_k}$$

  - Thus, the new feature set consists of the union of features obtained from the two kinds of directed edges.

$$\boldsymbol{f_{new} = f_{X_i \rightarrow X_j} \cup f_{X_k \rightarrow Y}\ , \text{ where } X_i \rightarrow X_j, X_k \rightarrow Y \in D'}$$

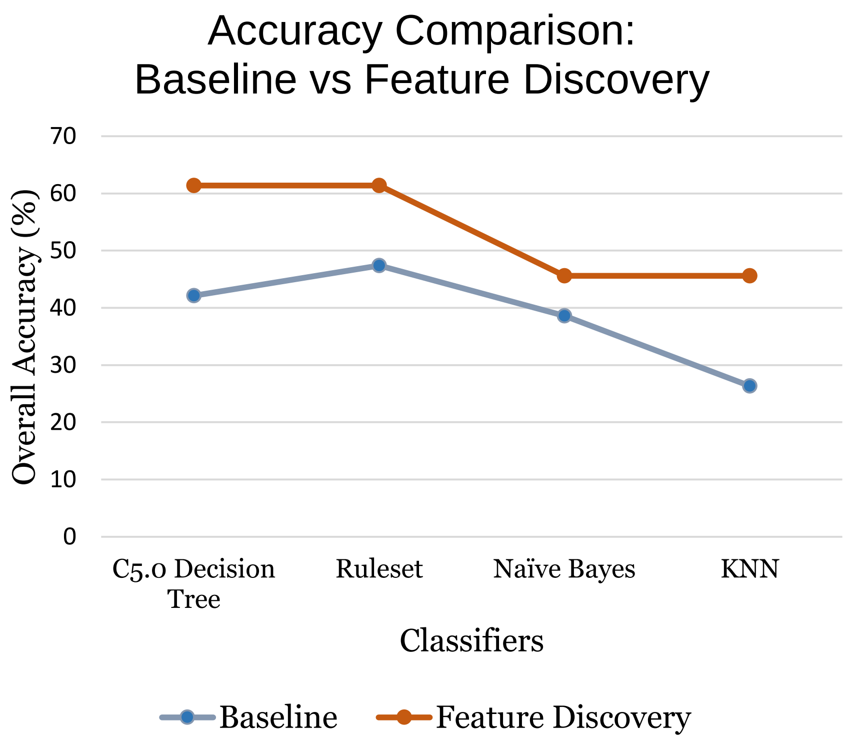
- The training dataset and testing dataset are transformed into the new feature space, f_{new}.

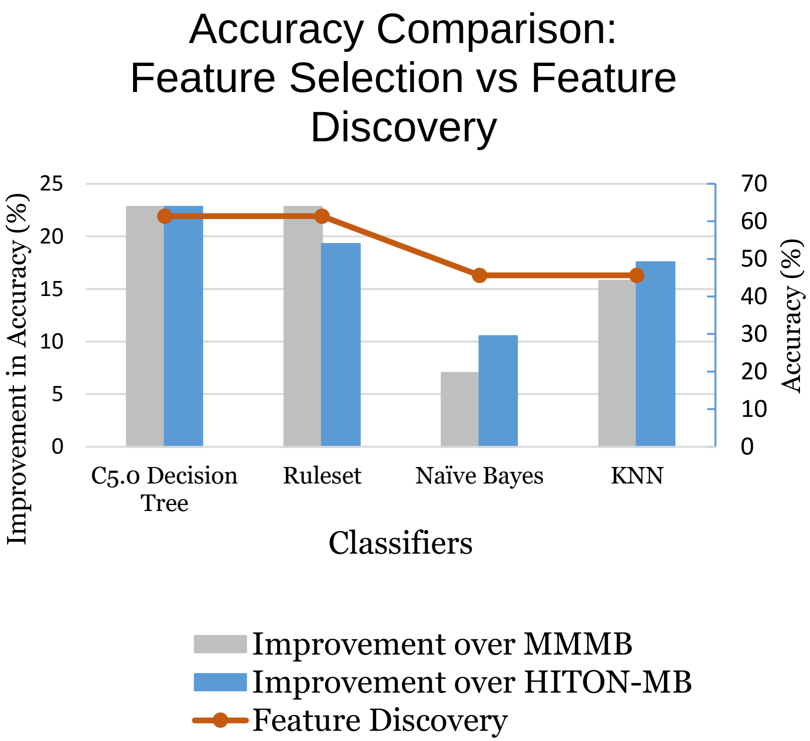
Comparison with Other Methods

- We built 57 classification models to evaluate the performance of our proposed method and compared it with,
  - Baseline methodology
    - All the climate indices in the multivariate training dataset are used to forecast rainfall.
  - Local causal discovery-based feature selection methods.
    - Max-Min Markov Blanket (MMMB) [4]
    - HITON-Markov Blanket (HITON-MB) [4]
  - We validated our method against the traditional climate analysis baseline methods, such as Principal Component Analysis and Climatology with a 10 year window, and found that our method had an improved performance.

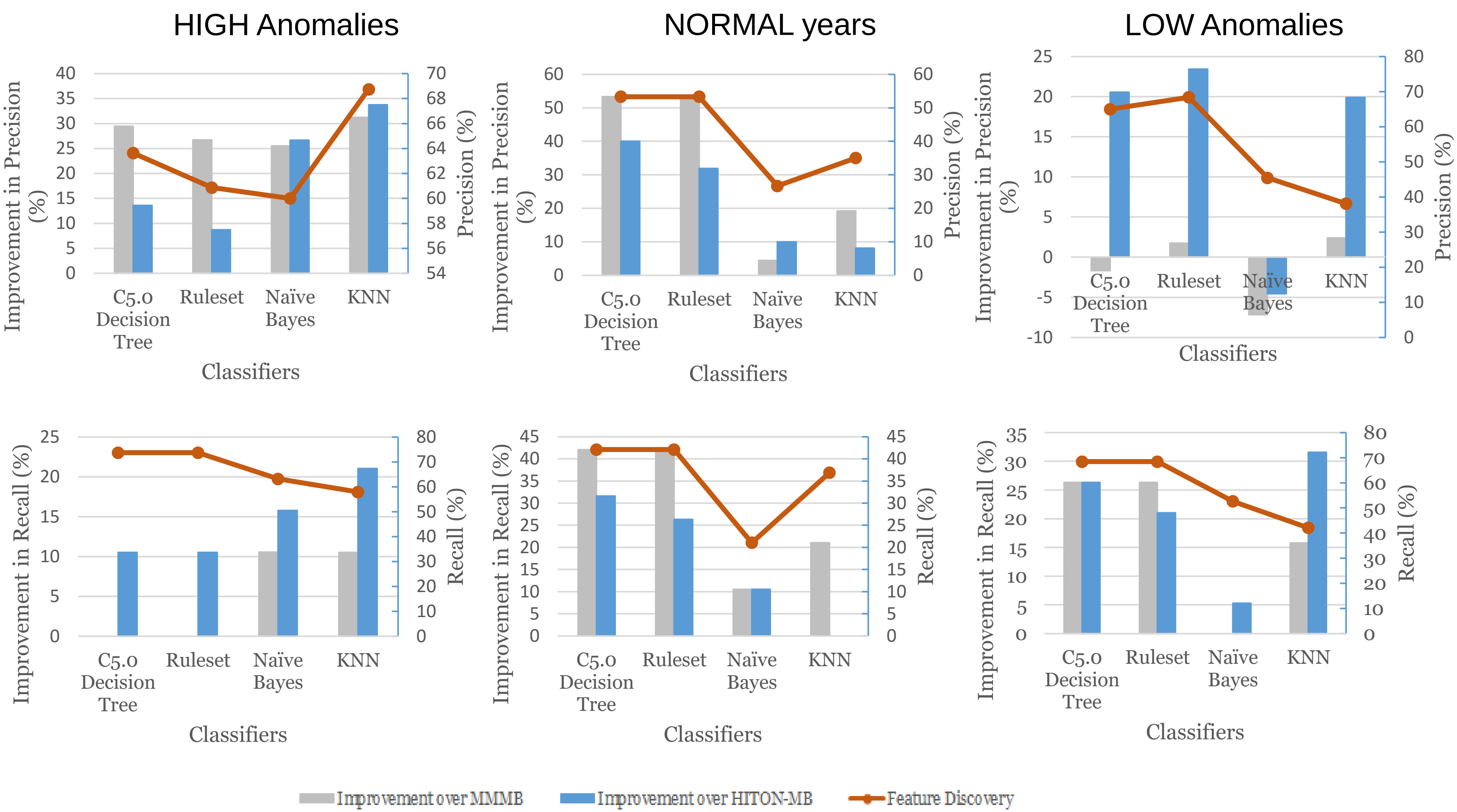
NC STATE UNIVERSITY

Performance Comparison: Accuracy

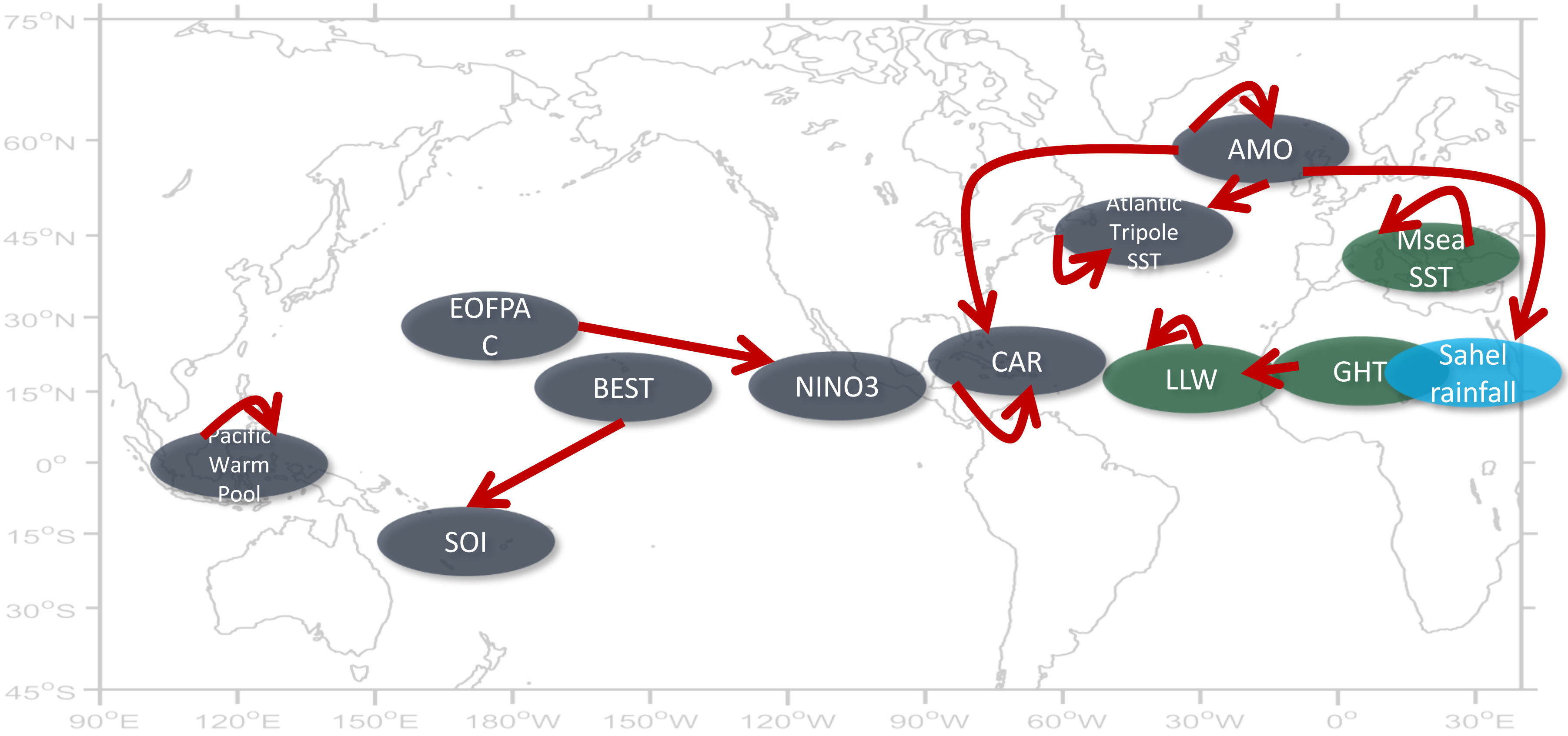




Performance Comparison: Precision and Recall



Frequently Selected
Potential Causal Relations



Conclusions

- In this work, we proposed a **feature discovery** methodology from causal graphs,
  - Formulated selection of potential causal relations to explore a new feature space.
  - Estimate causal effects by **addressing multicollinearity**.
  - Performed **stability assessment** of estimated causal effects.
  - Proposed a method to construct new features by weighted linear combination of causal relations and the stable causal effects.
- The methodology was applied to a multivariate time series dataset to **forecast seasonal rainfall anomalies** in the African Sahel region.
- The features discovered from causal models are **physically interpretable** and **relevant** to the behavior of seasonal rainfall and the newly constructed features improve the forecasting performance.

References

1. Maathuis, Marloes H., Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics* 37.6A (2009): 3133-3164.
2. D. Colombo and M.H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15:3741{3782, 2014.
3. Jolliffe, Ian T. A note on the use of principal components in regression. *Applied Statistics* (1982): 300-303.
4. Aliferis, Constantin F., et al. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *The Journal of Machine Learning Research* 11 (2010): 171-234.
5. Eugenia, et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* 77.3 (1996): 437-471.

Acknowledgements

- Thanks to: Dr. F. Semazzi, Dr. N. Samatova, D. González, J. Harlalka, N. Pedemane, Dr. V. Kumar.
- Support for this research was provided by the NSF grant #1029711.
- Data for this research was obtained from the NCEP/NCAR Reanalysis and the NOAA ERSST datasets.
- Climate indices available from:

  http://www.esrl.noaa.gov/psd/data/climateindices/list/