Toward Data-driven, Semi-automatic Inference of Phenomenological Physical Models: Application to Eastern Sahel Rainfall

Saurabh V. Pendse[1,2,*], Isaac K. Tetteh[1,*], Fredrick Semazzi[1], Vipin Kumar[3], and Nagiza F. Samatova[1,2,+]

[1] North Carolina State University, NC 27695, USA
[2] Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA
[3] University of Minnesota, Minneapolis, MN 55455, USA
[*] Authors contributed equally
[+] Corresponding author: samatova@csc.ncsu.edu

Abstract

First-principles based predictive understanding of complex, dynamic physical phenomena, such as regional precipitation or hurricane intensity and frequency, is quite limited due to the lack of complete phenomenological models underlying their physics. To address this gap, *hypothesis-driven*, *manually*-constructed, conceptual hurricane models and models for regional-scale precipitation extremes have been emerging. To complement both approaches, we propose a methodology for *data-driven*, *semi-automatic* inference of plausible phenomenological models and apply it to derive the model for eastern Sahel rainfall, an important factor for socioeconomic growth and development of this region. At its core, our methodology derives cause-effect relationships using the Lasso multivariate regression model and quantifies compound affect that the complex interplay among the key predictors at their prominent temporal phases plays on the response (rainfall). Specifically, we propose methods for (a) detecting and ranking predictors' prominent temporal phases, (b) optimizing the regularization penalty, (c) assessing predictor statistical significance, (d) performing impact analysis of data normalization on model inference, and (e) calculating the Expected Causality Impact (ECI) score to quantify impact analysis. The culmination of this study is the plausible phenomenological model of the eastern Sahel seasonal rainfall and quantified key climate drivers involved in the rainfall variability at different time lags. To the best of our knowledge, this is the first phenomenological model of this phenomenon; several of its components are consistent with the known evidence from literature.

1 Introduction

First-principles based physical models of Earth systems have been the primary contributors to our fundamental understanding of the key physical processes relevant to climate change assessment and impact. They have been providing relatively reliable predictions at global scale for *ancillary* variables, such as Sea Surface Temperature (SST) or wind speed at different heights. Yet, for *variables crucial for impact assessment*, such as regional precipitation, hurricane intensity and frequency, droughts and floods, physics-based models have been found the least reliable. As Quirin Schiermeier said, “The sad truth of climate science is that the most crucial information is the least reliable” [20].

To address this gap, *hypothesis-driven*, *manually*-constructed, conceptual models such as hurricane models [7] and regional-scale precipitation extremes [17, 23] have been emerging. Given the fact that climate and earth sciences have recently experienced a rapid transformation from a data-poor to a data-rich environment, the next logical question is whether and to what extent a *data-driven*, *semi-automatic* inference of plausible phenomenological models could complement powerful, yet limiting, physics-based model abstractions.

While highly noble, this goal of *data-driven*, *semi-automatic* inference of plausible phenomenological models is ambitious and technically challenging. Climate variability is influenced by a set of interacting processes that directly or indirectly contribute to the phenomenon of interest. For example, in case of rainfall and surface relative humidity variability over West Africa, moisture supply over West Africa primarily emanates from the eastern equatorial and South Atlantic, determined from the strength of the meridional and the zonal modes.

Moreover, other teleconnection patterns such as El Niño Southern Oscillation (ENSO), North Atlantic Oscillation (NAO), and Indian Ocean Dipole (IOD) are competitively engaged to dictate the rainfall and surface relative humidity variability at different scales [9, 19, 24].

Toward this goal, methods for inferring causal relationships between climate variables and the response (rainfall) could provide unique insights into the mechanics of a complex system, and thus complement statistical correlations that suggest associative relations between variables in the system. While promising, it is non-trivial to devise methods for accurate causality inference, especially for chaotic systems with a large number of complex, cascading interactions between factors, each having a local or a global effect on the system state .

A cardinal method for estimating causality is known as Granger causality [10]. Namely, a feature "causes" another feature, if the regression model built using the temporally lagged values of the feature along with the lagged values of the response feature is statistically more significant than the auto-regression model. Inspired by the Granger causality, multi-variate generalizations [1, 15] along with the methods based on partial correlation between time-series data [5] have been reported. They addressed one of the major drawbacks of Granger causality—testing for causality between a variable pair, rather than a variable group, which may yield high false positives or false negatives.

From causality perspectives,, the Lasso multivariate regression model [26], also known as the ℓ_1-penalty regression, derives (a) temporal phases of variables coined as *lagged* predictors, (b) a *penalty* term that sparcifies the predictor feature space, and (c) *predictor coefficients* of the magnitude and type of causal relationships shared with the response.

Recent work on causal inference based on the Lasso regression model on temporal and spatio-temporal data includes the time-series Lasso and the Lasso lambda [1] that offer means for calculating the parameters of regularized linear regression and best subset selection, the group elastic net [15] and the sparse group Lasso [2] that address the grouping of lagged variables corresponding to a feature and enforce spatial smoothness via an additional spatial penalty term.

1.1 Research Issues and Contributions Although the aforementioned methods yield better results than their traditional counterparts based on pairwise Granger causality, they have not addressed a number of important issues affecting correctness and predictability of inferred phenomenological models for the target phenomenon. Below, we highlight some of them and articulate our contribution toward resolving these issues.

1.1.1 Detection of Prominent Temporal Phases Naïvely including all the lagged predictors in the multivariate regression model may lead to an under-determined system. An interesting aspect is the fact that there is a higher probability of a causality between a response and a predictor, when both are in their most prominent temporal phases. We propose a temporal phase ranking methodology to detect these prominent phases for both predictor and response variables, thus reasonably addressing this under-determined problem and, at the same time, providing a more sensitive variable causality measure (Section 4.2).

1.1.2 Regularization Parameter (ℓ_1) Optimization The Lasso regression model is sensitive to the penalty parameter λ. Specifically, too high a value of λ leads to a large number of coefficients converging to zero, while too low a value of λ leads to a sub-optimal sparsity among the coefficients. Optimal selection of λ is considered a *black art.* We present an iterative method for λ-optimization based on optimization of the average mean-squared error (Section 3).

1.1.3 Statistical Significance Estimation The Lasso method produces values for the predictor coefficients whose sign and magnitude may provide clues about the strength and directionality of the inferred causality relationship between the predictor and the response. However, it does not assess the statistical significance of the inferred causality, namely whether such a relationship could be expected simply by chance. We propose a robust $\boldsymbol{\varphi}$-method to determine statistical significance of regression coefficients with respect to their causality influence on the response (both in terms of the magnitude and the sign) and show its superior performance compared to traditional methods based on assumptions about specific distributions of random variables [4, 27] (Section 5).

1.1.4 Impact Analysis and Phenomenological Model Can methods adopted to infer causal relations be assumed to be independent of the nature of the data? Inherent variability in the data sources and heterogeneity in how this data is preprocessed stimulate a study of what impact different forms of data normalization can have on the inferred causality models. We present a systematic analysis of five data normalization techniques. We augment the Lasso method with our modified Expected Causality Impact Analysis Model (ECIAM) using Probability Tree Diagram (PTD) method [25] to perform a systematic analysis of the five data normalization techniques and their impact on the inferred causal relationships. We finally propose a draft of the phenomenological model of eastern Sahel rainfall and show that its core components are consistent with the known evidence from literature (Section 6).

1.2 Motivating Example: The Sahel Use Case

Our primary motivation hinges on the important roles that surface relative humidity (RH) and rainfall play in the socio-economic sector of West Africa. Specifically, RH affects the occurrence of meningococcal meningitis, a highly fatal and morbid disease, affecting nearly 250,000 people annually, with an average of 25,000 deaths [16]. Knowledge of RH-meningitis relationship is currently being exploited as a strategy to control the disease, by channeling the limited vaccines and logistics to the most vulnerable areas, where RH is well under a certain threshold, typically 40%. Likewise, rain-fed agriculture is the main driver of the national economies in this region. Therefore, an understanding of RH and rainfall variability and predictability will contribute to improvements in modeling studies as well as enhancements in formulating policies, plans, and programs for sustainable socio-economic growth of this region.

1.3 Data Table 1 summarizes spatio-temporal data used for this study, namely, monthly values of 32 climate indices over the 61 year period from 1950 to 2010 obtained from NCEP/NCAR reanalysis [14] and NOAA ERSST [22] along with 11 explicitly generated indices.

2 Background: The Lasso Method

The Lasso method [26] is a penalized regression model, where a penalty term, called the ℓ_1-penalty or the ℓ_1 regularization parameter, is added to encourage sparsity among the coefficients of the regression model. The penalty term regularizes the parameters $\boldsymbol{\beta}$ and prevents model over-fitting by reducing the coefficients of insignificant predictors to 0.0; this results in automatic feature selection.

Let $\boldsymbol{X} = [\boldsymbol{x}^1\ \boldsymbol{x}^2 \ldots \boldsymbol{x}^p]$ be an $n \times p$ design matrix of n p-dimensional feature vectors $\boldsymbol{x}^i$ $(n > p)$, $\boldsymbol{\beta} = [\beta_1 \beta_2 \ldots \beta_p]^T$ be a $p \times 1$ vector of regression coefficients, and $\boldsymbol{y}$ be the $n \times 1$ response vector. Given λ, the Lasso solves the following optimization problem:

$$\min_{\boldsymbol{\beta}} h(\boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1, \tag{2.1}$$

where $\lambda \geq 0$ and $\lambda||\boldsymbol{\beta}||_1$ is the penalty term with $||\boldsymbol{\beta}||_1 = \sum_{1=1}^{p} |\boldsymbol{\beta}_i|$.

Methods to approximately fit ℓ_1 regularized models include the Forward Stagewise Regression (FSR) and Least Angle Regression (LAR) [6]. The three major computational issues to consider are speed, memory usage, and accuracy obtained using these methods. FSR does not necessarily produce the best model if there are redundant predictors and is known to give regression coefficients that need shrinkage [26]. LAR gives a rapid convergence and is efficient for an overdetermined system $(n > p)$, but its implementation involves cross-product matrices, which are known to be inaccurate in case of multi-colinearity within the predictors.

We solve the Lasso optimization problem using the shooting algorithm [8], a coordinate descent method that cycles through the coordinates, optimizing the current one and keeping the remaining coefficients fixed. It is an iterative method that computes the coefficient paths efficiently, especially, in very high-dimensional settings, by utilizing the preceding solution as the starting point for the next iteration.

Let $\boldsymbol{X}^{(-i)} = [\boldsymbol{x}^1 \ldots \boldsymbol{x}^{i-1}\ \boldsymbol{x}^{i+1} \ldots \boldsymbol{x}^p]$ be the matrix of all $\boldsymbol{x}$ vectors, excluding the vector $\boldsymbol{x}^i$. Similarly, let $\boldsymbol{\beta}^{(-i)} = [\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_p]^T$ be the coefficient vector $\boldsymbol{\beta}$ excluding the coefficient β_i. Algorithm 1 presents a pseudo-code of the Lasso method using the shooting algorithm.

Algorithm 1 Multi-variate Lasso optimization

```
Require: $\boldsymbol{X}$–design matrix, $\boldsymbol{y}$–response vector, $\lambda$–penalty parameter, $\delta$–convergence threshold
Initialize $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ (random or using Ordinary Least Squares)
Initialize $\delta$ (e.g, $\delta \le 1e^{-4}$)
while $(\max|\boldsymbol{\beta}_{cur} - \boldsymbol{\beta}_{prev}| \ge \delta)$ do
    Compute $f = h(\boldsymbol{\beta})$
    for $i = 1, 2, \ldots, p$ do
        Given $\boldsymbol{\beta}^{(-i)}$ and $\boldsymbol{y}^i = \boldsymbol{y} - \boldsymbol{X}^{(-i)}\boldsymbol{\beta}^{(-i)}$, find $\beta_i^*$ via the shooting algorithm:
        $\min_{\beta_i} h(\beta_i) = \frac{1}{2}||\boldsymbol{y}^i - \boldsymbol{x}^i\beta_i||_2^2 + \lambda|\beta_i| + \lambda||\boldsymbol{\beta}^{(-i)}||_1$
        $\boldsymbol{\beta}_{prev} = \boldsymbol{\beta}_{cur}$
        Set the $i^{th}$ element of $\boldsymbol{\beta}_{cur}$ to $\beta_i^*$
    end for
end while
return $\beta_{cur}$
```

3 Regularization Parameter (ℓ_1) Optimization

It is of utmost importance to select λ in equation 2.1 so that it leads to the most accurate sparsity among the regression coefficients and gives a balanced feature selection. Too low a value of λ leads to a very few coefficients converging to 0.0, thus, compromising the over-fitting criterion. Too high a value of λ causes most of the coefficients to converge to 0.0, thus producing a sub-optimal prediction.

Table 1: Data sources for explicitly generated indices[1]

No.	Variable	Abbr.	Region	Preprocessing
1	Lower Level Westerly Jets EOF 1,2,3	LLW	0°-20°N, 60°W-25°E	EOF Anomaly Index
2	Mediterranean Sea EOF 1,2,3	MSEA	30°N-46°N,6°W-36°E	EOF Anomaly Index
3	850 Hpa Geo-potential Height EOF 1,2,3	GHT	0°-40°N,40°W-30°E	EOF Anomaly Index
4	Indian Ocean Dipole :	IOD		
	Western Tropical Indian Ocean		10°S-10°N, 50°E-70°E	EOF Anomaly Index
	Southeastern Tropical Indian Ocean		10°S-0°, 90°E-110°E	EOF Anomaly Index
5	Atlantic ENSO	EATL	3°S-3°N, 30°W-0°	EOF Anomaly Index

[1] Indices generated using data obtained from NCEP reanalysis[14] and NOAA[22]

Generalized cross-validation, minimization of Stein's unbiased risk estimate [26], and Bayesian Information Criterion (BIC) [15] have been used for tuning the ℓ_1-penalty. Optimization via generalized cross-validation outperforms the unbiased risk estimate in terms of the median mean-squared error across different size simulation models [26]. The BIC-based model relies on a proper estimation of degrees of freedom and depends on finding the correct likelihood function.

We next discuss our adaptation for choosing λ that minimizes the average mean squared error in a K-fold cross validation of the data (see Algorithm 2). The rows of $\boldsymbol{X}$ and $\boldsymbol{y}$ are first partitioned to generate K predictor-response pairs, i.e. $(\boldsymbol{X}_1, \boldsymbol{y}_1)$, $(\boldsymbol{X}_2, \boldsymbol{y}_2)$, ..., $(\boldsymbol{X}_K, \boldsymbol{y}_K)$. Let $(\boldsymbol{X}_{(-j)}, \boldsymbol{y}_{(-j)})$ denote the predictor-response pair obtained by deleting the j^{th} part $(\boldsymbol{X}_j, \boldsymbol{y}_j)$ from $(\boldsymbol{X}, \boldsymbol{y})$. Suppose that $\boldsymbol{\beta}^L_{(-j)}$ be the Lasso solution for the $(\boldsymbol{X}_{(-j)}, \boldsymbol{y}_{(-j)})$ predictor-response pair. If n_j be the number of data points in the j^{th} predictor response pair $(\boldsymbol{X}_j, \boldsymbol{y}_j)$, then the average mean-squared error (AMSE) for a given value of λ is given by:

$$(3.2) \qquad \overline{\epsilon_K}(\lambda) = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{n_j}\left|\left|\left(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}^L_{(-j)}\right)\right|\right|_2^2$$

We first calculate AMSE for selected values of K and non-negative $\lambda \in (\lambda_{\min}, \lambda_{\max})$. In order to select $\lambda_{\max}$, we determine the value of λ that results in all (or an unreasonable majority) of the coefficients converging to 0.0. We choose K that minimizes the Residual Sum of Squares (RSS) error in the resulting regression model upon cross-validation. Finally, we choose a value of λ that minimizes AMSE:

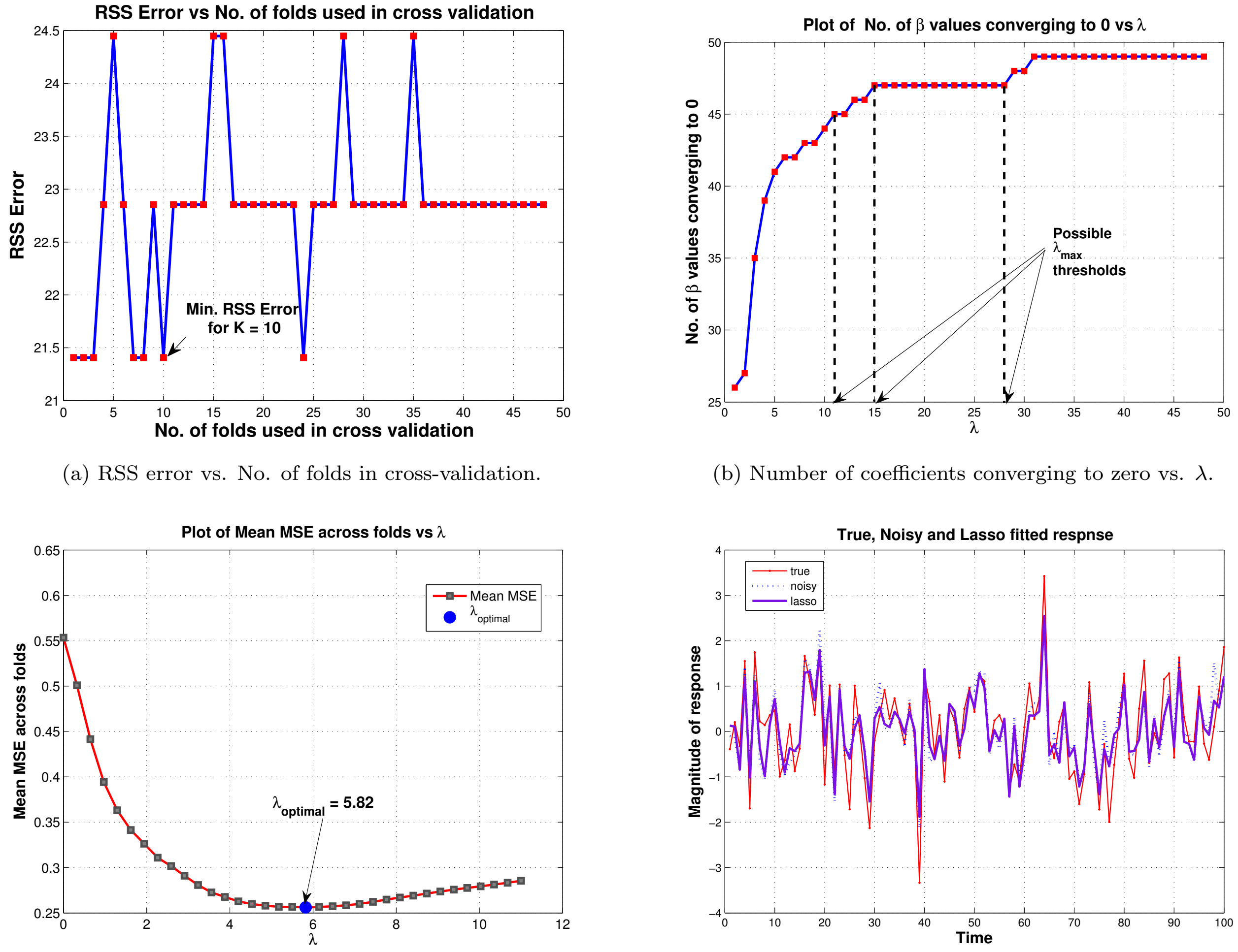
$$(3.3) \qquad \lambda_{opt}(K) = \min_{\lambda\in(\lambda_{\min},\lambda_{\max})}\overline{\epsilon_K}(\lambda)$$

We demonstrate the working of the Lasso algorithm (using the shooting method) through a simple example. We randomly generate 100 data points for a set of 50 predictor variables, i.e. the design matrix $\boldsymbol{X}$ by using pseudo-random numbers from a standard normal distribution, and a coefficient vector $\boldsymbol{\beta}$ with some of the values randomly taken to be non-zero, while the others are set to zero. Using this $\boldsymbol{X}$ and $\boldsymbol{\beta}$, we generate the response vector $\boldsymbol{y}$. Now, we use $\boldsymbol{X}$ and $\boldsymbol{y}$ as inputs to the Lasso algorithm.

For this experiment, we first choose an optimal λ value and then perform the actual Lasso regression on the synthetic data. Specifically, we choose λ in the range $(0, 11]$ and $K = 10$ based on the analysis shown in the Fig. 1a and Fig. 1b. It is evident from Fig. 1a that the frequently used 10 fold cross-validation results in a minimum RSS error. Also, from Fig. 1b, we define several possible thresholds (11,15, 28) for $\lambda_{\max}$ depending on the accuracy desired. We know that increasing the value of λ results in a larger proportion of the coefficients converging to zero. Hence, $\lambda_{\max}$ is chosen so that the number of coefficients converging to zero becomes constant for values greater than the threshold value. We obtain a local minimum for AMSE over all the folds for a given value of λ which is deemed as optimal for the Lasso algorithm. This process is defined in algorithm 2. Fig. 1c provides a visualization of the same for our test example. Fig. 1d shows the true, noisy and Lasso-fitted output for the synthetically generated data.

4 Detection of Prominent Temporal Phases

4.1 Lagged Data for the Lasso Regression Our main objective is to determine causal relationships between climate variables over specific months from a given set of multi-variate, temporal data. Thus, the predictors are time-lagged climate variables specific to the months under consideration for causality. By causality, we mean that the occurence of a given response $\boldsymbol{y}_{t_0}$ at time t_0 can be attributed to the occurrence of predictor variables $\boldsymbol{x}^i_{t'}$, i.e. the i^{th} predictor variable at time t', where $t' = t_0 - \triangle t_i$. Here, $\triangle t_i$ is the time-lag associated with the i^{th} predictor. We use monthly lags for our experiments.



(a) RSS error vs. No. of folds in cross-validation.

(b) Number of coefficients converging to zero vs. λ.

(c) AMSE error vs. penalty term (λ).

(d) A comparison of the ground truth without noise, with noise, and with the Lasso-fitted output.

Figure 1: Optimization of the regularization parameter (ℓ_1).

Algorithm 2 ℓ_1-penalty optimization

```
Require: $\boldsymbol{X}$–an $n \times p$ design matrix; $\boldsymbol{y}$–an $n \times 1$ response vector
1: Find $K$ for fold number in cross-validation that minimizes the RSS error between the actual and predicted response.
2: Partition the data into $K$ predictor-response pairs $(\boldsymbol{X}_1,\boldsymbol{y}_1)$, …, $(\boldsymbol{X}_K,\boldsymbol{y}_k)$.
3: $\lambda_{\min} = 0$. Choose $\lambda_{\max}$, for which the number of coefficients converging to zero (obtained from the Lasso method) becomes constant for $\lambda \geq \lambda_{\max}$.
4: For $\lambda \in (\lambda_{\min}, \lambda_{\max})$, compute AMSE, see Eq. 3.2.
5: $\lambda_{\text{opt}} = \min_{\lambda \in (\lambda_{\min},\lambda_{\max})} \overline{\epsilon_K}(\lambda)$
6: return $(\lambda_{\text{opt}}, K)$
```

In essence, we are trying to establish a relationship between the past behavior of a predictor variable and the present behavior of the response variable, and based upon it, possibly infer existent causal relationships between the predictors and the response.

Let $\boldsymbol{x^i}$ represent the i^{th} predictor, $i = 1, \ldots, p$, defined over the set $\boldsymbol{Q}$ of T years, $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_T\}$. Let $x_j^{i,q}$ denote the value of the i^{th} predictor at the j^{th} time lag for the q^{th} year. Let $\boldsymbol{M}_i$ be the set of specific months taken into account for the i^{th} predictor variable across all the years. Let $\boldsymbol{y}$, the response variable, be similarly defined.

We seek to determine the causal relationship between the response $\boldsymbol{y}_{t_0}$ and the predictors $\boldsymbol{x}^i_{t'}$ during their corresponding months, $t_0 \in \boldsymbol{M}_y$ and $t' \in \boldsymbol{M}_i$, using the data from all the years in $\boldsymbol{Q}$. Mathematically, this relationship can be described as the following system of equations:

$$y^q_{t_0} \quad = \quad \sum_{i=1}^{p} \sum_{j \in \boldsymbol{M}_i} \beta_{i,j} x_j^{i,(q-\delta_j)}, \forall q \in \boldsymbol{Q}, \forall t_0 \in \boldsymbol{M}_y, \tag{4.4}$$

where δ_j determines whether the previous year or the current year values are to be used for the j^{th} month of the predictor variable $\boldsymbol{x}^i$, i.e. for x^i_j. The need for this month arises when some of the months from the response set $\boldsymbol{M}_y$ precede the months for the predictor $\boldsymbol{x}^i$, i.e. for x^i_j when considered in the same year. For example, if the months considered for the response are (Jan, Feb, Mar) and those for a predictor are (July, Aug, Sep), then the values for July, Aug, Sep for that predictor must be drawn from the previous year and not the same year. Hence, in this case $\delta_{\text{July}} = \delta_{\text{Aug}} = \delta_{\text{Sep}} = 1$. We use the resulting response vector $\boldsymbol{y}$ and the design matrix $\boldsymbol{X}$ as inputs to the Algorithm 1.

4.2 Ranking of Temporal Phases Although literature in the climate community [9, 19, 24] may indicate the phases to be chosen for each variable, it would be valuable to automatically extract similar results using data mining techniques. Let $\boldsymbol{D}^i$ be a $T \times 12$ matrix, where T is the number of years taken into consideration for the i^{th} predictor, $i = 1, \ldots, p$. Each matrix column $D^i(:, j)$ would represent the behavior of the i^{th} predictor for the j^{th} phase. Thus, each data matrix $\boldsymbol{D}^i$ would represent 12 time series, each corresponding to a phase (i.e. month or season). We use a weighted voting technique, described next, to rank the months based on their prominence.

Let $\boldsymbol{m}^i_j$ denote the time series of the i^{th} predictor for the j^{th} phase (month or season) across all the years. The crux of the method is that every temporal phase for the i^{th} predictor $\boldsymbol{m}^i_a$ votes for every other phase $\boldsymbol{m}^i_b$, $\forall a, b \in \{1, \ldots, 12\}$ and $i \in \{1, \ldots, p\}$. The vote V_{ab} for $\boldsymbol{m}^i_a$ by $\boldsymbol{m}^i_b$ is weighted by the *relative importance measure* of $\boldsymbol{m}^i_b$, defined below. Thus, the effective contribution of $\boldsymbol{m}^i_b$ to the total vote count of $\boldsymbol{m}^i_a$ is $V_{ab}r_b$, where r_b is the measure of relative importance or *rank* of $\boldsymbol{m}^i_b$. Hence, the cumulative relative importance of r_a of $\boldsymbol{m}^i_a$ can be written as:

$$r_a = \sum_{b=1}^{p} V_{ab} r_b \tag{4.5}$$

This equation reduces to the eigenvalue decomposition problem for the vote matrix $\boldsymbol{V}$, where the elements of the eigenvector corresponding to the largest eigenvalue represent the ranks of the corresponding temporal phases (months or seasons) [18].

Mathematically, we measure the votes cast by $\boldsymbol{m}^i_a$ for $\boldsymbol{m}^i_b$ based on the similarity, such as Pearson correlation, between their corresponding time series:

$$V_{ab} = 1 + \text{sim}(\boldsymbol{m}^i_a, \boldsymbol{m}^i_b) \tag{4.6}$$

By the Perron-Frobenius theorem [12], eigendecomposition for Eq. 4.5, exists for a real eigenvalue $\omega > 0$ with a real eigenvector $\boldsymbol{r}$ with positive elements, such as those generated by the power iteration method [12].

In order to make the ranking process more robust and sensitive to the influences that different temporal phases may have on different parts of the time series for a given variable, we compute the ranks repeatedly, every time selecting a different part of the time series for a feature, i.e. we randomly choose a subset of the time series so that it approximates to about 75% of the entire time series. This way, we ensure that (a) the major part of the time series is preserved and (b) local prominence as well as prominence over the entire time series is taken into account, and (c) the ranking procedure is not biased towards a specific part.

In our experiments, we computed the average seasonal rank using data restandardization by method E defined in table 3 using $1,000$ sub-samples from our data. We then selected the top 4 ranked seasons corresponding to each predictor as the ones belonging to its prominent temporal phase. Table 2 shows the results obtained for some of our candidate predictors which are found to be consistent with [9, 19, 24]. We also computed ranks using other preprocessed data (section 6). The complete results thereof are available at `http://www.freescience.org/cs/prm_causality/ranks.zip`.

5 Statistical Significance Estimation

A solution to the Lasso optimization problem (see Eq. 2.1) produces the corresponding $\boldsymbol{\beta}$ coefficients. The magnitude of each coefficient $\beta_{i,j}$ indicates the strength of the putative causal relationship between the response $\boldsymbol{y}$ and the predictor variable $\boldsymbol{x}^i$ for the j^{th} month. A positive sign of the coefficient indicates that an increase in the predictor value likely causes the increase in the the response, and vice versa for the negative sign.

However, it is very important to adopt a proper method for deciding on the significance of the obtained coefficients. A naïve approach may pre-define a threshold value for the coefficient vector $\boldsymbol{\beta}$ and deem every coefficient as significant, if it exceeds this threshold value. This approach does not assign the true significance to the coefficients, because the threshold may be different for different datasets and even for the coefficients corresponding to different predictors in the same dataset.

Table 2: Average prominent season selection for predictor variables over 1,000 iterations

No.	Predictor	Top 4 Seasons[1]											
		1	2	3	4	5	6	7	8	9	10	11	12
1	Nino3				✓	✓	✓					✓	
2	North Atlantic Oscillation (NAO)	✓	✓	✓									✓
3	Atlantic Meridional Mode (AMM)					✓	✓	✓				✓	
4	Eastern Sahel Rainfall Index						✓	✓	✓	✓			
5	Mediterranean Sea EOF 1					✓	✓	✓		✓			

[1] 1 = Jan-Feb-Mar, 2 = Feb-Mar-Apr, 3 = Mar-Apr-May......, 12 = Dec-Jan-Feb

One way of addressing this problem is to statistically assess the significance of the coefficient values, namely, to estimate the p-values for all the coefficients. Traditional methods focus on the significance of the coefficients based on their magnitude. These methods obtain a p-value for each of the coefficients against the null hypothesis that the obtained value is an outlier in the distribution of the coefficients obtained by sub-sampling the time series data over a large number of iterations. Methods, such as bagging, are normally employed in this case to generate sub-samples and then the Student's t-test is subsequently used to test the null hypothesis. However, this metric is more relevant from a robustness standpoint; it indicates whether the magnitudes of the coefficients are robust, and whether the value of the coefficient has a low enough variance.

In the case of causality inference, however, we are interested in two types of significance measures:

1. The significance with respect to the robustness of the coefficients in terms of their magnitudes: To measure the robustness, we calculate the significance based upon: (a) the Student's t-test (the $\boldsymbol{t}$ approach) and (b) the χ^2-test (the $\boldsymbol{\chi^2}$ approach).

2. The significance in terms of the coefficient's sensitivity to different degrees of causal relationships between the response and the predictors: To measure this kind of significance, we simulate the variation in the degree of causality by randomizing the response over several iterations, while keeping the predictors unchanged. We name this approach as $\boldsymbol{\varphi}$ and detail it next.

For the $\boldsymbol{\varphi}$ approach, we randomly permute the response vector $\boldsymbol{y}$ to obtain $\tilde{\boldsymbol{y}}$, while keeping the design matrix $\boldsymbol{X}$ constant. Such a permutation destroys most of the causal relationships that existed between the real response and the predictors. Running the Lasso algorithm for $\tilde{\boldsymbol{y}}$ as the response produces the coefficient vector $\tilde{\boldsymbol{\beta}}$ that, in theory, represents random causal relationships between the response $\boldsymbol{y}$ and the predictors $\boldsymbol{X}$. The large ensemble of such permutations (e.g., $\nu \geq 1,000$) results in the distribution of the $\boldsymbol{\beta}$ coefficients that are expected at random. Based on this distribution, we may calculate the statistical signicance (p-values) of the true $\boldsymbol{\beta}$ coefficients for a given false discovery rate, as depicted in Algorithm 3. [Note that $\tilde{\boldsymbol{\beta}}_h$ denotes the coefficient vector produced by the Lasso algorithm on the h^{th} permutation $\tilde{\boldsymbol{y}}_{\boldsymbol{h}}$ of $\boldsymbol{y}$ and $count(expr)$ stands for the number of times the logical expression $expr$ is true.]

Algorithm 3 Statistical significance of $\boldsymbol{\beta}$

```
Require: $\tilde{\boldsymbol{\beta}}$–a $\nu \times p$ ensemble matrix from running the
    Lasso algorithm on $\nu$ permutations of $\boldsymbol{y}$
for $i$=1 to $p$ do
    if $\beta(i) > 0$ then
        $\text{pval}_{\beta_i} = \dfrac{\text{count}(\tilde{\beta}_h(i) > \beta(i)) + 1}{\nu + 1}$
    else if $\beta(i) < 0$ then
        $\text{pval}_{\beta_i} = \dfrac{\text{count}(\tilde{\beta}_h(i) < \beta(i)) + 1}{\nu + 1}$
    else
        $\text{pval}_{\beta_i} = 1$
    end if
end for
$\boldsymbol{pval} = (\text{pval}_{\beta_1}, \ldots, \text{pval}_{\beta_p})$
return $\boldsymbol{pval}$
```

Table 4 compares the selection of significant predictor variables using the three approaches: $\boldsymbol{t}$, $\boldsymbol{\chi^2}$, and $\boldsymbol{\varphi}$. Each of the 19 predictors for the month of October are pre-processed using the method D, see Table 3. The response, the rainfall index, is for its prominent phase (Jul-Aug-Sep). The $\boldsymbol{t}$ approach assesses almost all the coefficients as statistically significant and the $\boldsymbol{\chi^2}$ approach provides a slightly higher filtration. The $\boldsymbol{\varphi}$ method selects 9 out of the 19 predictors as statistically significant.

Table 5 summarizes the implication of each predictor selection on the performance of the linear regression model built from the significant predictors that are pre-processed by each of the five methods in Table 3.

Specifically, we report the percentage of retention of R^2 correlation from the baseline, when all the predictors are used for the regression model. Since the approach $\boldsymbol{t}$ hardly selects any predictors, its expected retention rate is 100%. The approach $\boldsymbol{\chi^2}$ results in a strong model degeneracy; its retention rate is around 60%. In contrast, the $\boldsymbol{\varphi}$ method retains 98% predictive skill while reducing the number of significant predictors by half. [Note that the goal here is not to maximize the predictive skill of the baseline model. In fact, the better performing models ($R^2 = 0.64$) have been reported [28] but for a different set of predictors than the ones of interest to our study.]

Table 3: Data normalization types and formulas

Type	Formula
A (control)	$\boldsymbol{m}_\eta^{i'} = \boldsymbol{m}_\eta^i$
B (test)	$\boldsymbol{m}_\eta^{i'} = \dfrac{\boldsymbol{m}_\eta^i - \overline{\boldsymbol{m}}_\eta^i}{\sigma(\boldsymbol{m}_\eta^i)}$
C (test)	$\boldsymbol{m}_\eta^{i'} = \dfrac{\boldsymbol{m}_\eta^i - \text{mean}(\boldsymbol{m}_\eta^i, \boldsymbol{m}_{\eta+1}^i, \boldsymbol{m}_{\eta+2}^i)}{\sigma(\boldsymbol{m}_\eta^i, \boldsymbol{m}_{\eta+1}^i, \boldsymbol{m}_{\eta+2}^i)}$
D (control)	$\boldsymbol{m}_\eta^{i'} = \text{mean}(\boldsymbol{m}_\eta^i, \boldsymbol{m}_{\eta+1}^i, \boldsymbol{m}_{\eta+2}^i)$
E (test)	$\boldsymbol{m}_\eta^{i'} = \dfrac{\text{mean}(\boldsymbol{m}_\eta^i, \boldsymbol{m}_{\eta+1}^i, \boldsymbol{m}_{\eta+2}^i)}{\sigma(\boldsymbol{m}_\eta^i, \boldsymbol{m}_{\eta+1}^i, \boldsymbol{m}_{\eta+2}^i)}$

6 Impact Analysis and Phenomenological Model

We now present a study of the impact of source data characteristics on the performance of the Lasso causal inference model. We describe the data preprocessing methods used and complement the Lasso method with the ECIAM to obtain representative and accurately interpretable results. We then assess the climatological relevance of the results by stating plausible and observed physical phenomena that are evident from the causality analysis.

6.1 Data Normalization The data in Table 1 is accumulated from different sources, and each variable may have undergone different types of preprocessing. Hence, to deal with an inherent heterogeneity in the data, we apply five types of data normalization and study normalization affects on the causality models. Table 3 summarizes these normalization techniques. [Note that normalization A, B, and C is monthly-based, while D and E is season-based.]

For each type, we apply the same normalization to each predictor, keeping the response unaltered. Specifically, consider the i^{th} predictor, $\boldsymbol{x}^i = (\boldsymbol{m}_{\text{Jan}}^i, \ldots, \boldsymbol{m}_{\text{Dec}}^i)$, where $\boldsymbol{m}_\eta^i$ ($\eta \in$ (Jan, Feb, …, Dec)) represents the monthly time series corresponding to the η^{th} month for the i^{th} predictor over all the years (i.e., $1950 - 2010$) (Section 4.2). For methods C, D and E, when we reach Dec, we roll over to Jan for the next month in the season. Then, for each type, the normalized $\boldsymbol{m}_\eta^{i'}$ form of $\boldsymbol{m}_\eta^i$ is calculated based on the corresponding equation in Table 3.

6.2 ECIAM-based Causality Analysis For each normalization A-E, to generate a set of β coefficients, we run the Lasso algorithm using the predictor climate indices for a given month $\eta \in$ (Jan, Feb, …, Dec) and the rainfall index in its known prominent phase (July-Aug-Sep), also estimated in Table 2 using our ranking method in Section 4.2. Thus, for each normalization, we get a $12 \times p$ matrix $\boldsymbol{B}$ of β coefficients. We augment this matrix with the corresponding statistical significance measure (see $\boldsymbol{\varphi}$ method in Section 5) assigned to each coefficient in $\boldsymbol{B}$, thus producing a tensor $\boldsymbol{\Omega}$.

Analyzing the 3-dimensional tensor $\boldsymbol{\Omega}$ and deducing consistency with physically observed phenomena is nontrivial. Hence, we adapt our ECIAM [25] to analyze the compound effect of the predictors in each month on the response. Through this process, we convert the information represented in $\boldsymbol{\Omega}$ into a set of monthly *ECI* scores that capture the degree to which the significant predictors collectively cause the rainfall.

To introduce the ECI scores, let's first consider the logical form of $\boldsymbol{\Omega}$ as a Probability Tree Distribution (PTD). The root of the tree corresponds to the response variable and captures the cumulative impact of its month-specific child nodes across all months (i.e., 12 children). Each month-specific child node η will, in turn, have b_η children, where b_η is the number of *significant* predictor variables from the Lasso regression model built for month $\eta \in$ (Jan, Feb, …, Dec). Figure 3a illustrates a synthetic PTD.

Given the PTD, let ECI_η denote the ECI score for the month $\eta \in$ {Jan, Feb, …, Dec} defined as follows:

$$ECI_\eta = \sum_{i=1}^{b_\eta} P_i^\eta \cdot S_i, \tag{6.7}$$

where P_i^η is the probability of the occurrence of predictor i and S_i is the impact score (or the causality impact score) for the predictor i, as formally defined below.

More formally, P_i^η is the joint probability of observing the causal effect of the i^{th} predictor at month η on the response. It is calculated as the probability of the path from the root of the tree to the leaf predictor i passing through the root's child η. Each month η has an equal probability of occurrence, namely $\dfrac{1}{12}$,

Table 4: p-values of predictor coefficients selected by the $\boldsymbol{t}$, $\boldsymbol{\chi^2}$ and $\boldsymbol{\varphi}$ methods. Significance level $\alpha = 0.05$

Name	Abbr.	$\boldsymbol{t}$	$\boldsymbol{\chi}^2$	$\boldsymbol{\varphi}$	$\beta \pm \sigma^*$
North Atlantic Oscillation	NAO_{10}	0.0	0.45	0.0	1.82 ± 0.18
Atlantic Meridional Mode	AMM_{10}	0.0	0.0	0.0	0.13 ± 0.01
Atlantic Multidecadal Oscillation	AMO_{10}	0.0	0.15	0.33	-0.02 ± 0.02
Lower Level Westerly Jets EOF 1	LLW1_{10}	0.0	0.04	0.06	-0.09 ± 0.03
Lower Level Westerly Jets EOF 2	LLW2_{10}	0.0	0.74	0.17	0.0 ± 0.04
Lower Level Westerly Jets EOF 3	LLW3_{10}	0.0	0.02	0.38	0.01 ± 0.03
Mediterranean Sea EOF 1	MSEA1_{10}	0.0	0.33	0.0	-0.48 ± 0.20
Mediterranean Sea EOF 2	MSEA2_{10}	0.0	0.12	0.0	0.59 ± 0.11
Mediterranean Sea EOF 3	MSEA3_{10}	0.0	0.0	0.03	-0.14 ± 0.05
850 Hpa Geo-potential Height EOF 1	GHT1_{10}	0.01	0.0	0.13	-0.12 ± 0.04
850 Hpa Geo-potential Height EOF 1	GHT2_{10}	0.0	0.02	0.0	-0.36 ± 0.11
850 Hpa Geo-potential Height EOF 1	GHT3_{10}	0.0	0.0	0.19	-0.13 ± 0.11
Indian Ocean Dipole	IOD_{10}	0.0	0.0	0.18	0.1 ± 0.08
Atlantic ENSO	EATL_{10}	0.27	0.04	0.83	0 ± 0.25
Nino3	Nino3_{10}	0.25	0.03	0.0	0 ± 0.04
Nino1+2	Nino1_{10}	0.0	0.11	0.03	0.39 ± 0.49
Nino3.4	Nino34_{10}	0.0	0.43	0.14	0.0 ± 0.0
Nino4	Nino4_{10}	0.0	0.04	0.31	0.02 ± 0.02
Multivariate ENSO	MEI_{10}	0.0	0.0	0.03	-0.13 ± 0.03

$^*-$ based on the bagging sampling

Table 5: R^2 between the true (rainfall) and the predicted response using (a) all the predictors and (b) only the significant predictors as per methods $\boldsymbol{t}$, $\boldsymbol{\chi^2}$, and $\boldsymbol{\varphi}$.

Method	R^2	$\boldsymbol{t}$		$\boldsymbol{\chi^2}$		$\boldsymbol{\varphi}$		
	(All 19)	R^2	% retained	R^2	% retained	R^2	% retained	Predictors
A (control)	0.5456	0.5456	100	0.3237	59.33	0.5147	94.33	7
B (test)	0.5393	0.5393	100	0.3449	63.95	0.5320	98.64	10
C (test)	0.5410	0.5410	100	0.3452	64	0.3690	68.42	12
D (control)	0.5422	0.5422	100	0.3326	61.34	0.4733	87.29	9
E (test)	0.5444	0.5444	100	0.3327	61.11	0.4861	89.29	5

and each significant predictor i for that month has an equal probablity of occurence, namely $\frac{1}{b_\eta}$. Therefore, $P_i^\eta = \frac{1}{12} \cdot \frac{1}{b_\eta}$.

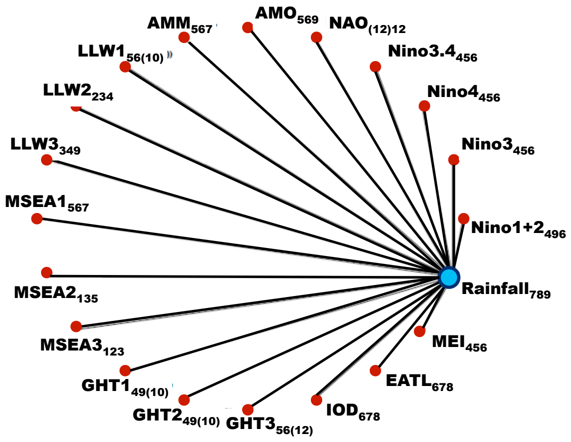
The causality impact score S_i for the predictor i is formally defined as follows: $S_i = \sum_{j=1}^{p} (\psi_j \beta_j \cdot \gamma(\boldsymbol{x}^j) \cdot A(\boldsymbol{x}^j))$, where $\psi_j = 1$, if the coefficient corresponding to the j^{th} predictor is significant, else $\psi_j = 0$. Also, $\gamma(\boldsymbol{x}^j)$ is the decay/growth factor that accounts for the attenuation/amplification in the effect of a predictor with distance, and $A(\boldsymbol{x}^j)$ is the area of geographical region of the j^{th} predictor. For the sake of simplicity, in this study, we assume that $A(\boldsymbol{x}^j) = \gamma(\boldsymbol{x}^j) = 1 \, \forall j \in \{1, \dots, p\}$. ECIAM analysis with these factors is outside the scope of this paper and is a part of future work.

Given the formulas for P_i^η and S_i, ECI_η in Eq. 6.7 can thus be defined as follows:

$$ECI_\eta = \sum_{i=1}^{b_\eta} \left(P_i^\eta \cdot \sum_{j=1}^{p} (\psi_j \beta_j) \right) \tag{6.8}$$

All the beta coefficient values with their corresponding ψ values are available at `http://www.freescience.org/cs/prm_causality/beta_plots.zip`.

6.3 A Putative Phenomenological Model This study has been spurred by evidence of interactions between complex but competing dynamical processes



Numerical subscripts are rolling seasons (as in Table 2)

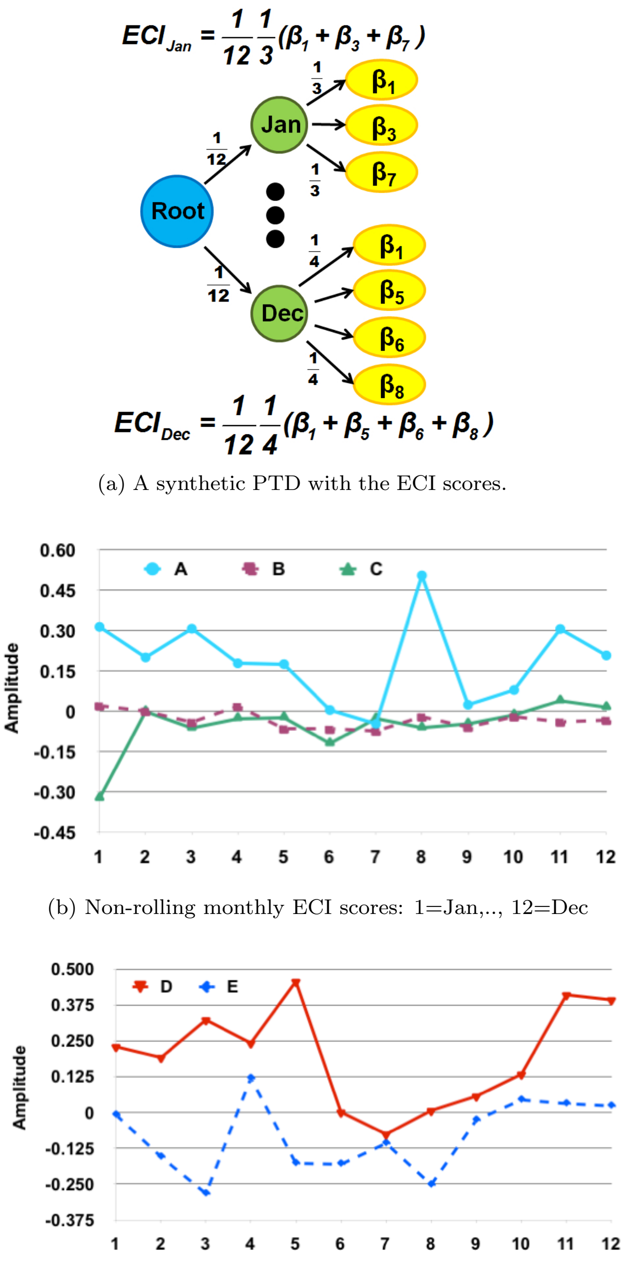
Figure 2: A putative phenomenological model for eastern Sahel rainfall variability. Prominent phases for features chosen using the ranking method in Section 4.2

operating on different time scales over the tropical Atlantic [24], as well as the quasi-stationary orographic control of the northern African climate [21], which are closely linked to the sub-Saharan African Sahel climate variability. Specifically, the overarching goal of the study is (a) to develop a plausible phenomenological model centered on eastern Sahel rainfall and (b) to investigate and quantify time-evolution of the interplay of influences of key teleconnection patterns as well as dynamic factors involved in the modulation of rainfall.

Toward this goal, we developed a plausible phenomenological model (Fig. 2) based on five different experimental set ups described in Section 6. The outcomes may suggest clues to numerical climate modelers on how to improve model physics, data assimilation, and parameterization schemes and to resolve the Sahel climate change rainfall issue.

The proposed phenomenological model contributes to predictive understanding of the sub-region's climate in a number of ways. Our analysis of plausible physical mechanisms revealed by each experiment suggests that depending on ambient climatic conditions and forcing factors, some climate drivers may switch their roles, *from enhancement to antagonism* and vice versa, or *to total dissipation*, depending on their phases [24].

Also, the functional significance of the predictors—in terms of (a) the total number of branches in each PTD, (b) their detection probability, and (c) their magnitude and directionality of influence (i.e., beta coefficients)—quantifies pivotal differentiators of the physical processes responsible for the experimental outcomes. For instance, the cumulative number of predictors specific to experiments A, B, C, D, and E, equivalent to the total number of leaf PTD nodes, is 55, 106, 110, 53, and 74, with the corresponding monthly or seasonal averages of 4.58, 8.83, 9.17, 4.42, and 6.17, respectively.



(a) A synthetic PTD with the ECI scores.

(b) Non-rolling monthly ECI scores: 1=Jan,.., 12=Dec

(c) Rolling seasonal ECI scores: 1=JFM,.., 12=DJF

Figure 3: PTD and ECI score graphs for methods A–E.

Specifically, experiment A shows that about 92% of the fluctuations exhibit positive ECI scores, suggesting rainfall enhancements (Fig. 3b). The pattern essentially owes its existence to 8 precipitation enhancement agents—NAO, IOD, LLW3, MSEA1, AMM, Nino4, MSEA2, and MEI (the canonical ocean-atmosphere coupled ENSO), when they are the most energetic [13]. Their antagonizers [13, 24] emanate chiefly from the tropical Pacific (represented by Nino 1+2, Nino 3, Nino 4, and MEI), GHT1-3, MSEA1-3, AMM, LLW1, and EATL [3, 13], a total of 13. The suggested mechanism here is that depending on a particular month, and when they are the most energetic, the enhancers selectively co-exist to form complex interlocking or communication systems that interfere with the activities of the suppressors [11, 19, 24]. The tipping point is July that accounts for 8.33% deficit and could be due to a temporal shut-off of the enhancement processes by the dominant roles of the suppressors. However, restoration is realized when the enhancement system is re-energized.

The precipitation enhancers detected in experiment B are NAO, IOD, MSEA1, LLW3, AMO, AMM, GHT2, Nino4, GHT3, MEI, GHT1, and MSEA2, a total of 12. However, the precipitation suppressors are GHT3, Nino1+2, MSEA3, EATL, GHT1, LLW1, MSEA2, Nino3, MEI, AMM, GHT2, GHT3, Nino4, IOD, MEI,GHT1 and MSEA1, a total of 17. The ECI curve shown in Fig. 3b is in opposition to that of experiment A. The trend suggests an exhibition of outstanding influential suppressors, whose composite actions could tend to de-couple or break down any pre-existing enhancement system, leading to a prolonged drought, especially, after April.

In experiment C, the impact of the pre-processing did not significantly change the dynamical processes (Fig. 3b). The precipitation enhancers were NAO, IOD, MSEA1, LLW3, AMM, AMO, GHT2, GHT1, MSEA2, Nino4, GHT3, MEI, Nino1+2, LLW2, Nino3.4, and Nino3, a total of 16, while the suppressors were Nino1+2, GHT1, GHT3, MSEA3, LLW1, MSEA2, EATL, Nino3, IOD, AMM, MEI, GHT2, Nino4, MEI, MSEA1, AMO, and MSEA1, a total of 17. However, a transient dynamical separability of experiments B and C was detected in Jan-Feb and Nov-Dec. In Jan-Feb, it was due to the presence or absence of AMM, EATL, MSEA2, IOD, and Nino1+2, whereas in Nov-Dec, it was due to Nino3, MSEA1, AMO, and MSEA3.

The ECI scores on the seasonal timescales are shown in Fig. 3c. Though on different timescales, experiment D is reminiscent of A (Fig. 3b vs Fig. 3c), suggesting that the monthly climatic influences persist through seasons, because of longer memories, and thus are, to a large extent, dynamically inseparable. The precipitation enhancers were ascribed to NAO, MSEA1, IOD, LLW3, GHT1, NAO, AMM, MSEA2 and LLW2, a total of 9, whereas the suppressors were Nino1+2, GHT3, MSEA2, Nino3, MSEA3, EATL, LLW1, and MEI, a total of 8.

The ECI index for experiment E (Fig. 3c) is quite distinct from D, as well as from the monthly timescale experiments. In particular, experiments D and E tend to be in anti-phase over some periods. One of the major reasons for this observation is the detection of sustained, antagonistic roles between NAO and the tropical Pacific Nino1+2, Nino3, and MEI that perhaps determine the migratory roles, especially, of the Mediterranean Sea modes, leading to these out-of-phase relationships.

7 Conclusion

In this paper, we discussed the non-trivial issues pertaining to causality in temporal data, especially, in the context of semi-automatic inference of putative phenomenological models consistent with the known evidence and physics. We proposed solutions and applied them to propose the first phenomenological model of eastern Sahel rainfall. We also identified key contributors to the rainfall variability in this region. Specifically, we determined the most active predictors in various temporal phases, and quantified their individual and compound effect on the rainfall response.

From methodology perspective, we studied the Lasso multivariate regression model and proposed possible mechanisms for optimizing its regression parameters. To deal with the under-determined nature of the problem, we also proposed a method for detecting and ranking prominent temporal phases for the target predictors, which were found to be consistent with climatological knowledge. Our proposed method for estimating the statistical significance of the derived causal relationships showed a much better performance, in terms of predictive-skill retention and feature selection, compared to traditional approaches. Finally, we discussed the robustness of the synergy of the proposed methods by taking into account its sensitivity with respect to different forms of data normalization.

In summary, this study has given us the opportunity for further exploration and elaboration of our hypotheses for the plausible physical processes, especially, one of the most outstanding observations about the antagonistic role generally existing between NAO and the Pacific ENSO-related phenomena. Observations of this kind suggest some clues to dynamical climate modelers for improvement of model physics, data assimilation, and parameterization schemes, especially, those used in climate projection over West Africa and the global tropics, as a whole.

8 Acknowledgements

This work was supported in part by the U.S. Department of Energy, Office of Science and the U.S. National Science Foundation (Expeditions in Computing). Oak

Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. D.O.E. under contract no. DEAC05-00OR22725.

References

[1] A. Arnold, Y.Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 66–75. KDD, August 2007.

[2] A. Bolstad, B.D. Van Veen, and R. Nowak. Causal network inference via group sparse regularization. *Signal Processing, IEEE Transactions*, 59(6):2628–2641, 2011.

[3] P. Chang, Y. Yamagata, and et al. Climate fluctuations of the tropical coupled systems - the role of ocean dynamics. *Journal of Climate*, 19(5122-5174), 2006.

[4] E.M. Cramer. Significance tests and tests of models in multiple regression. *The American Statistician*, 26(4):26–30, October 1972.

[5] M. Drton and M.D. Perlman. A SINful approach to gaussian graphical model selection. Technical Report 457, Department of Statistics, University of Washington, 2004.

[6] B. Efron, T. Hastie, and et al. Least angle regression. *The Annals of Statistics*, 32(2):407–451, 2004.

[7] K. Emanuel. The hurricane-climate connection. *Bulletin of the American Meteorological Society*, 89(5):3–7, 5 2008.

[8] W. J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

[9] A. Giannini, R. Saravannan, and P. Chang. Dynamics of the boreal summer african monsoon in the NSIPPI atmospheric model. *Climate Dynamics*, 25:517–535, 2005.

[10] C.W.J. Granger. Investigating causal relations by econometric models and crosss-spectral methods. *Econometria*, 37:424–4381, 1969.

[11] S.M. Hagos and K.H. Cook. Dynamics of the west african monsoon jump. *Journal of Climate*, 20:5264–5284, 2007.

[12] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[13] J.W. Hurrel, M. Visbeck, and et al. Atlantic climate variability and predictability : A CLIVAR perspective. *Journal of Climate*, 19:5100–5121, 2006.

[14] E. Kalnay, M. Kanamitsu, and et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of American Meteorological Society*, 77:437–471, 1996.

[15] A. Lozano, H. Li, A. Niculescu-Mizil, and et al. Spatial-temporal causal modeling for climate change attribution. In *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 587–595. KDD, 2009.

[16] A. M. Molesworth, M. C. Thomson, and et al. Where is the meningitis belt? Defining an area at risk of epidemic meningitis in Africa. *Trans R Soc Trop Med Hyg*, 96(3):242–249, 2002.

[17] O'Gorman and T. Schneider. The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. In *Proceedings of the National Academy of Sciences*, August 2009.

[18] L. Page, S. Brin, and et al. The pagerank citation ranking: Bringing order to the web. Technical Report 66, Stanford InfoLab, November 1999.

[19] D.P. Rowell. The impact of mediterranean ssts on the sahelian rainfall season. *Journal of Climate*, 16:849–862, 2003.

[20] Q. Schiermeier. The real holes in climate science. *Nature*, 463:284–287, 2010.

[21] F.H.M. Semazzi and L. Sun. The role of orography in determining the sahelian climate. *International Journal of Climatology*, 17:581–596, 1997.

[22] T.M. Smith and R.W. Reynolds. NOAA extended reconstructed sea surface temperature (ERSST). *Journal of Climate*, 17(2466-2477), 2004.

[23] M. Sugiyama, H. Shiogama, and S. Emori. Precipitation extreme changes exceeding moisture content increases in MIROC and IPCC climate models. *Proceedings of the National Academy of Sciences of the United States of America*, 107(2):571–575, 2010.

[24] R.T. Sutton, S.P. Jewson, and D.P. Rowell. The elements of climate variability in the tropical atlantic region. *Journal of Climate*, 13:3261–3284, 2000.

[25] I. K. Tetteh, A. Awuah, and E. Frempong. Post-project analysis: The use of a network diagram for environmental evaluation of the Barakese Dam, Kumasi, Ghana. *Environmental Modeling and Assessment*, 11(3):235–242, 2006.

[26] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

[27] F. Westada and H. Martens. Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression. *Journal of Near Infrared Spectrocopy*, 8:117–124, 2000.

[28] A. Yeshanew and M.R. Jury. North african climate variability. part 3 : Resource prediction. *Theoretical and Applied Climatology*, 89:5162, 2007.