Spatial-temporal Causal Modeling for Climate Change Attribution

A. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
{aclozano, liho, anicule, liuya, perlich, hosking, nabe}@us.ibm.com

ABSTRACT

Attribution of climate change to causal factors has been based predominantly on simulations using physical climate models, which have inherent limitations in describing such a complex and chaotic system. We propose an alternative, data centric, approach that relies on actual measurements of climate observations and human and natural forcing factors. Specifically, we develop a novel method to infer causality from spatial-temporal data, as well as a procedure to incorporate extreme value modeling into our method in order to address the attribution of extreme climate events, such as heatwaves. Our experimental results on a real world dataset indicate that changes in temperature are not solely accounted for by solar radiance, but attributed more significantly to CO2 and other greenhouse gases. Combined with extreme value modeling, we also show that there has been a significant increase in the intensity of extreme temperatures, and that such changes in extreme temperature are also attributable to greenhouse gases. These preliminary results suggest that our approach can offer a useful alternative to the simulation-based approach to climate modeling and attribution, and provide valuable insights from a fresh perspective.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms

Keywords

Climate modeling, Climate change attribution, Spatio-temporal modeling, Causal modeling Granger causality

1. INTRODUCTION

Climate change is one of the most critical socio-technological issues mankind faces in the present century [1]. Though it is regarded primarily as an energy related problem, computing technology will play an important role in devising potential solutions in a variety of ways. One that particularly interests us is that of applying data modeling to the climate data in order to better understand and quantify the causal effects of various parameters involved. There is a clear need for an effective methodology of data modeling that will allow us to analyze the large amount of time series data on the climate and climate forcing agents and draw conclusions on how these factors affect each other and which parameters are to be controled for the best environmental results.

It is well recognized that climate is a chaotic system, and hence it is difficult to reliably model it as a whole. Nonetheless, there are reasons to believe we can meaningfully characterize causal or statistical relationships that exist among parameters of interest, and make assertions about the presence or absence of such relationships and quantify them. (Recently, there have been a number of articles published in prominent scientific journals that carry out studies of this type. [11, 15, 5]) Fundamentally, our goal is to focus on 'climate change detection and attribution' (i.e., identification, quantification and prioritization of the effects of controllable forcing factors on climate), rather than on 'climate projection' (i.e., prediction of the evolution of the global climate system in the next decades).

The climate system comprises complex relationships between a large number of variables. Hence, the factors of interest involve many dimensions, including measurements of climate parameters, anthropogenic factors, and regional factors [2]. Fortunately, many of these data are publicly available in forms that are well suited for data modeling – e.g. Climate Research Unit (CRU) dataset, NOAA NESDIS data set, Carbon Dioxide Info analysis Center (CDIAC).

Considerable amount of scientific investigations have been carried out to date in the community of climate change study, to address these very questions [11]. The dominant existing approach in the community, however, is based on forward simulation with climate models built using fundamental physical laws. These models are used to estimate the expected space-time pattern (fingerprints) of the response to individual anthropogenic or natural forcing factors on the observed climate. The task of detection and attribution is then performed by estimating the factors by which these model-simulated patterns have to be scaled to be consistent with the observed change (optimal fingerprinting), and by applying standard statistical significance tests for isolated hypotheses on the value of the estimated factors. As these existing approaches rely heavily on the employed climate models, they are subject to the models' shortcomings (e.g. models' uncertainties, simplifications, and discrepancies from observed data).

Given the understanding of the existing approaches and their limitations, what we propose is an alternative approach based on data modeling, with special attention paid to address unique characteristics of climate modeling. First has to do with our emphasis on attribution, rather than forecasting, of climate change, motivating

KDD'09, June 28–July 1, 2009, Paris, France.

us to look to techniques that aim at modeling causality. Secondly, the climate data are spatio-temporal in nature, where both the climate observations and forcings are associated with specific points in space and time, and these aspects will be critical for conducting informed analyses on climate change over time and across regions over the globe. Thirdly, there is a particular interest in modeling the extreme climate events, such as the frequency and severity of heatwaves and floods, beyond just the change in the mean climate behavior [15, 5].

To address the modeling challenge described above, we develop and employ methods of 'spatial temporal causal modeling,' which allow us to model causal relationships between time and space-persistent features, given spatio-temporal data. More specifically, we develop a spatio-temporal version of the so-called 'graphical Granger modeling methods,' which is an emerging collection of methods that combine the graphical modeling techniques with the notion of 'Granger causality' to derive effective methods for causal modeling based on time series data. Here 'Granger causality' is an operational definition of causality from econometrics, which is based on the premise that 'a cause necessarily precedes its effects,' and its adoption to graphical modeling allows us, to model causal relationships between a large number of time series variables.

Specifically, we develop a novel method we call "Group Elastic Net", which can address the spatio-temporal aspect of climate modeling, and use it as our primary modeling methodology. This algorithm incorporates the spatio-temporal structure in the data in the variable selection process of the regression procedure underlying graphical Granger modeling. That is, the lagged variables from different time steps for the same feature are grouped together and the penalty function used in variable selection is modified so as to enforce sparsity at the group level, rather than at the level of the individual lagged variables. Additionally, the spatial smoothness is enforced by an additional penalty term that encourages similarity between coefficients for spatial neighbors. This formulation leads to a grouped version of the so-called "elastic net" problem, for which we devise an efficient solution.

One potential weakness of a data centric approach to climate modeling is the lack of sufficient past data on extreme events, which may pose difficulties in modeling and attributing such events. Here we develop a dynamic modeling method by applying the theory of extreme event and value modeling. Extreme-value theory [3] provides a natural family of probability distributions for modeling the magnitude of the largest or smallest of a large number of events, and a canonical stochastic process model ([7], sec. 7.3) for the occurrence of rare events, those whose magnitude exceeds a very high (or very low) threshold. The stochastic process model involves three parameters, which specify the rate of occurrence of extreme events and the distribution of the magnitude of events that exceed a threshold. We treat these parameters as varying over space and time and we model their variation by means of a Bayesian hierarchical model in which the parameters are regarded as random variables. The outputs of the model are *a posteriori* estimates of the parameters at potentially all locations in space and time. From these outputs we can estimate the spatial and temporal variation of properties of the distribution of annual extremes. In particular we look for evidence of climate change in the temporal variation of our estimates of the "N-year event", the event magnitude that occurs on average once every N years.

The relationship between extreme event modeling and graphical Granger modeling has been underexplored in the literature to date. In the present work, we employ a relatively simple approach to combining the two: using our Bayesian hierarchical model we estimate the N-year event magnitudes associated with the climate metrics of interest, and we incorporate these estimated variables as additional variables in causal modeling and attribution in the spatio-temporal modeling with the grouped elastic net algorithm described above. The choice of N-year event magnitudes as a proxy of extreme temperature is, in part, motivated by the fact that they are typically approximated using normal distributions , which is consistent with our causal modeling method, using linear Gaussian models as component models of conditional distributions.

We evaluate our proposed approach with two sets of experiments: In the first set of experiments, we use simulated spatio-temporal data to demonstrate the advantage of the proposed spatial-temporal modeling method based on group elastic net, as compared to methods that do not take advantage of the spatial aspects of the data. In the second, and main, set of experiments, we use our developed methods to model real climate data, focusing on the data for the last couple of decades in the North American region. We collected and processed a wide range of climate related data for these space and time ranges, including the climatological observations, natural forcings (e.g. solar radiance), as well as greenhouse gas measurements. The results we obtained to date include: 1) Spatio-temporal causal modeling attributes the change in the temperature significantly to that of CO_2 and other greenhouse gases, even in the presence of solar radiance; 2) Extreme value modeling confirms that the intensity of extreme weather events, such as unseasonably hot summer days and warm winter days, have significantly increased between the years of 1982 and 2001; 3) The combination of the two approaches indicate that, for the N-year return level of temperature as well, CO_2 and other greenhouse gases are attributed even in the presence of, and with greater significance than, the solar radiance.

2. METHODOLOGY

2.1 Spatio-temporal Causal Modeling

2.1.1 Preliminaries: Graphical Granger Modeling

We briefly review the notion of "Granger Causality" [9], which was introduced by the Nobel prize winning economist, Clive Granger, and has proven useful as an operational notion of causality for time series analysis in econometrics. It is based on the idea that if a time series variable causally affects another, then the past values of the former should be helpful in predicting the future values of the latter, beyond what can be predicted based only on their own past values.

More specifically, a time series x is said to "Granger cause" another time series y, if regressing for y in terms of past values of y and x is more accurate with statistical significance, as compared to regressing just with past values of y. Let $\{x_t\}_{t=1}^T$ denote the time series variables for x and $\{y_t\}_{t=1}^T$ the same for y. The so-called Granger test first performs the following two regressions:

$$y_t \approx \sum_{l=1}^{L} a_j \cdot y_{t-l} + \sum_{l=1}^{L} b_j \cdot x_{t-l} \tag{1}$$

$$y_t \approx \sum_{j=1}^{L} a_j \cdot y_{t-j} \tag{2}$$

where L is the maximum "lag" allowed in past observations, and then applies a statistical test to determine whether or not (1) is more accurate than (2), with statistical significance.

The notion of Granger causality, as reviewed above, was defined for a pair of time series variables. Now in the context of climate modeling, we are actually interested in cases in which there are *many* variables present as opposed to a pair, and each one is

a *spatio-temporal* variable as opposed to a time series variable; and we wish to determine the causal relationships between them. Hence, the notion of Granger causality needs to be appropriately extended to incorporate the spatial dimension. Let us, for any measurement or feature over time and space (e.g. temperature, CO_2, etc), use variables (e.g. x) to refer to the entire spatio-temporal series, and use indexed variables (e.g. $x_{t,s}$) to denote the associated individual spatially and temporally lagged variables.

For convenience, we assume that the measurements are sampled along a regular spatial grid. Similarly to the notion of maximum temporal lag, one may consider a maximum "spatial lag" and suppose that each point is influenced by a finite neighborhood around it. Let $N(s)$ denote the set of points in the neighborhood of s. We assume that the neighborhood structure is identical for each grid point, and thus consider neighborhoods of the form $N(s) = s + \Omega$, where $\Omega = \{\omega_1, \ldots, \omega_K\}$ is a set of "relative locations".

Now, the extended Granger causality notion is defined in terms of the following two regressions:

$$y_{t,s} \approx \sum_{\omega \in \Omega} \sum_{l=1}^{L} a_{l,\omega} \cdot y_{t-l,s+\omega} + \sum_{\omega \in \Omega} \sum_{l=1}^{L} b_{l,\omega} \cdot x_{t-l,s+\omega} \quad (3)$$

$$y_{t,s} \approx \sum_{\omega \in \Omega} \sum_{l=1}^{L} a_{l,\omega} \cdot y_{t-j,s+\omega} \quad (4)$$

The above, simplified, scheme is symmetric with respect to time and space, but there is a difference between space and time that calls for a refinement of this formulation.

For applying Granger causality to many variables, or measurements, there is a collection of methods, known as graphical Granger modeling, which combines methods of graphical modeling with the notion of Granger causality. A particularly relevant approach is that of applying regression algorithms with variable selection to determine the causal links for each variable. Lasso [14] is a prime example, which trades off the minimization of the sum of squared errors and that of the sum of the absolute values of the regression coefficients in the penalty term.

Consider N measurements $x^i \ \ (i = 1, \ldots, N)$ (e.g. temperature, pressure, etc.). For each such measurement x^i, denote by $x^i_{t,s}$ its sample at time t and location s. For any given measurement x^i, one can view the variable selection process in the regression for $x^i_{t,s}$ in terms of $x^1_{t-l,s+\omega}$,...,$x^N_{t-l,s+\omega}$ $l = (1, \ldots, L)$,$\omega \in \Omega$, as an application of the Granger test on x^i against $x^1, \ldots, x^N$. By extending the pairwise Granger test to one involving an arbitrary number of spatial-temporal series, it makes sense to say that x^j Granger causes x^i if $x^i_{t-l,s+\omega}$ is selected for any time and spatial lags l, ω in the above variable selection process.

A critical aspect that is worth emphasizing, and is overlooked in most of the existing methods in the literature, is that the question we are interested in is whether an entire series $\{x^j_{t-l,s+\omega}, l \in \{1, \ldots, L\}, \omega \in \Omega\}$ provides additional information for the prediction of $x^i_{t,s}$, and not whether for specific time and spatial lags l, ω $x^j_{t-l,s+\omega}$ provides additional information for predicting $x^i_{t,s}$. Therefore, a faithful instantiation of Granger causal modeling, in the context of spatial-temporal modeling, should take into account the group structure imposed by the spatial-temporal series into the fitting criterion that is used in the variable selection process.

The foregoing discussions naturally lead to the proposal of our novel method, "group elastic net", which addresses both the issue of "grouping" the lagged variables for the same feature, and that of the smoothness desired for the spatial dimension.

Spatial-Temporal Causal Modeling

1. Input: Measurement data $\{\mathbf{x}_{t,s}\}_{t=1,..,T,\ s \in S}$ where each $\mathbf{x}_{t,s}$ is a N-dimensional vector of measurements taken at time t and location s.
   Input: A regression method with group variable selection, **REG**.
2. Initialize the adjacency matrix for the N measurements, i.e. $G = \langle V, E \rangle$ where V is the set of N measurements (e.g. by all 0's).
3. For each measurement $x^i \in V$, run **REG** on regressing for $x^i_{t,s}$ in terms of the past lagged variables, $x^j_{t-l,s+\omega}$, $j \in 1, \ldots, N$, $l \in 1, \ldots, L$, $\omega \in \Omega$. For each measurement $x^j \in V$ place an edge $x^j \rightarrow x^i$ into E, if and only if x^j was selected as a group by **REG**.

Figure 1: Generic Spatial-Temporal Causal Modeling Method

2.1.2 *Spatio-temporal Granger Modeling via Group Elastic Net*

The generic spatio-temporal causal modeling method we described in the foregoing section is given in Figure 1. We now describe the variable selection procedure which we propose to use as an instance of the **REG** procedure in Step 3 of our algorithm. We assume "spatial stationarity", i.e., that the same model applies to each point on the grid (relaxing this assumption will be the object of future work). More precisely, we consider regression coefficients of the form $\beta^k_{l,\omega}$ where k is the measurement (e.g. temperature), l is the time lag, ω is the relative location between the point considered and a point in its neighborhood.

Let S be the set of (interior) locations s such that for each $\omega \in \Omega$ $s + \omega$ is a point of the grid (and not outside the grid). Let, $t = 1, \ldots, T$ be the time points considered.

For a given measurement x^i, we propose to use the following penalized regression model to determine which spatial-temporal series x^j $(j = 1, \ldots, N)$ Granger-cause x^i.

$$\begin{aligned} \hat{\beta} \ = \ & \arg\min_{\beta} \sum_{s \in S} \sum_{t=L+1}^{T} (x^i_{t,s} - \sum_{j=1}^{N} \sum_{l=1}^{L} \sum_{\omega \in \Omega} \beta^j_{l,\omega} x^j_{t-l,s+\omega})^2 \\ & + \lambda_2 \underbrace{\sum_{j=1}^{N} \sum_{l=1}^{L} (\beta^j_{l,:})^T \tilde{\Delta}_j \beta^j_{l,:}}_{\text{SpatialPenalty}} + \lambda_1 \underbrace{\sum_{j=1}^{N} \|\beta^j_{:}\|_{\Delta_j}}_{\text{SparsityPenalty}}, \end{aligned} \quad (5)$$

where $\beta^j_{l,:} = \text{vect}(\beta^j_{l,\omega})_{\omega \in \Omega}$, $\beta^j_{:} = \text{vect}(\beta^j_{l,\omega})_{l=(1,\ldots,L),\omega \in \Omega}$,

$$\Delta_j = \begin{pmatrix} \tilde{\Delta}_j & 0 & 0 & 0 \\ 0 & \tilde{\Delta}_j & 0 & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \tilde{\Delta}_j \end{pmatrix},$$

and $\|y\|_{\Delta_j} = (y^T \Delta_j y)^{1/2}$.

The role of the "spatial penalty" is to enforce spatial regularization. Specifically the matrix $\tilde{\Delta}_j$ is meant to enforce spatial smoothness as well as some form of distance-based coefficient decay. Namely the regression coefficients are penalized more as they correspond to increasingly distant neighborhood locations. For instance, $\tilde{\Delta}_j$ could be a diagonal matrix such that the diagonal entry corresponding to $\beta^j_{l,\omega}$ equals $\|\omega\|$.

The "sparsity penalty" is a group Lasso penalty [16], which imposes sparsity across measurements. More precisely, the regression coefficients corresponding to spatial-temporal samples of the same

measurement are penalized as a group, namely through $\|\beta^j_{:}\|_{\Delta_j}$. Then l_1 norm of $\left(\|\beta^1_{:}\|_{\Delta_j}, \|\beta^2_{:}\|_{\Delta_j}, \ldots, \|\beta^N_{:}\|_{\Delta_j}\right)$ imposes that the coefficients corresponding to a given measurement are either included as a group in the model or excluded. Note that the dependence in j of $\tilde{\Delta}_j$ and Δ_j is due to the fact that we may consider different regularization matrices for different measurements.

Let Y be the vector of length $(T-L+1)|S|$ formed by $x^i_{t,s}, t = (L, \ldots, T),\ s \in S$. Consider the spatially and temporally lagged matrix X of dimension $((T-L+1)|S|) \times (NL|\Omega|)$ such that the row corresponding to the pair (t, s) is the vector formed by $x^j_{t-l,s+\omega}$, $j = (1, \ldots, N), l = (1, \ldots, L), \omega \in \Omega$. Let β be the corresponding vector of regression coefficients, i.e. β is of length length $NL|\Omega|$ formed by $\beta^j_{l,\omega}, j = (1, \ldots, N), l = (1, \ldots, L), \omega \in \Omega$. Denote by β^j the restriction of β to the elements corresponding to measurement x^j, i.e β_j is the vector formed by $\beta^j_{l,\omega}, l = (1, \ldots, L), \omega \in \Omega$. Then Eq. 5 can be rewritten as

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} L(\lambda_1, \lambda_2, \beta) \\ &= \arg\min_{\beta} \|Y - X\beta\|^2 + \lambda_2 \sum_{j=1}^{N} \|\beta^j\|^2_{\Delta_j} + \lambda_1 \sum_{j=1}^{N} \|\beta^j\|_{\Delta_j}.\end{aligned}$$

Notice that the above formulation resemble a "group version" of the Elastic net problem [17], hence we call it the Group Elastic Net.

The following proposition states that the Group Elastic Net problem can be transformed into a group Lasso problem, and hence can be efficiently solved using existing algorithms.

PROPOSITION 2.1. *Assume that Δ_j $(j = 1, \ldots, N)$ is positive definite, let $\Delta_j = S_j^T S_j$. Let $A_j = (S_j^T S_j)^{-1} S_j^T$, and*

$$C = \begin{pmatrix} A_1 & 0 & 0 & 0 \\ 0 & A_2 & 0 & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & A_N \end{pmatrix}.$$

The Group Elastic Net problem solution

$$\beta_{\mathrm{GEN}} = \arg\min_{\beta} \|Y - X\beta\|^2 + \lambda_2 \sum_{j=1}^{N} \|\beta^j\|^2_{\Delta_j} + \lambda_1 \sum_{j=1}^{N} \|\beta^j\|_{\Delta_j} \tag{6}$$

can be obtained by solving the Group Lasso problem

$$\beta_{\mathrm{GL}} = \arg\min_{\beta} \|\hat{Y} - \hat{X}\beta\|^2 + \gamma \sum_{j=1}^{N} \|\beta^j\|^2,$$

where $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$, $\hat{Y} = \begin{pmatrix} Y \\ 0_{(p)} \end{pmatrix}$, $\hat{X} = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} XC \\ \sqrt{\lambda_2} I \end{pmatrix}$ *and where p is the number of columns of X, and setting* $\beta^j_{\mathrm{GEN}} = \frac{1}{\sqrt{1+\lambda_2}} \Delta_j^{-1} S_j^T \beta^j_{\mathrm{GL}}$.

PROOF. Let $\tilde{X} = XC$ and $\tilde{\beta}^j = S_j \beta^j$. Then solving Eq. 6 is equivalent to solving

$$\begin{aligned}&\min \|Y - \tilde{X}\tilde{\beta}\|^2 + \lambda_2 \sum_{j=1}^{N} \|\tilde{\beta}^j\|_2^2 + \lambda_1 \sum_{j=1}^{N} \|\tilde{\beta}^j\|_2 \\ = \ &\min \|Y - \tilde{X}\tilde{\beta}\|^2 + \lambda_2 \|\tilde{\beta}\|^2 + \lambda_1 \sum_{j=1}^{N} \|\tilde{\beta}^j\|_2 \end{aligned} \tag{7}$$

Set $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$, and $\hat{\beta} = \sqrt{(1+\lambda_2)}\tilde{\beta}$.. Then solving Eq. 7 is equivalent to solving

$$\min \|Y - \frac{1}{\sqrt{(1+\lambda_2)}} \tilde{X}\hat{\beta}\|^2 + \frac{\lambda_2}{1+\lambda_2} \|\hat{\beta}\|^2 + \gamma \sum_{j=1}^{N} \|\hat{\beta}^j\|_2 \tag{8}$$

Let $p = NL|\Omega|$, i.e, p is the number of columns of $\hat{X}$. Let $q = (T-L+1)|S|$, i.e, q is the length of Y and also the number of rows of $\tilde{X}$. Let $\hat{Y}_{(q+p)} = \begin{pmatrix} Y \\ 0_{(p)} \end{pmatrix}$, and $\hat{X}_{(q+p)\times p} = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \tilde{X} \\ \sqrt{\lambda_2} I \end{pmatrix}$. Then the problem is equivalent to solving

$$\min \|\hat{Y} - \hat{X}\hat{\beta}\|^2 + \gamma \textstyle\sum_{j=1}^{J} \|\hat{\beta}^j\|^2$$

which is the Group Lasso formulation. □

Similar to [17], the penalty parameters are tuned as follows. We consider a set of candidate parameters Λ_2 for λ_2, (for instance $\Lambda_2 = (0, 0.01, 0.1, 1, 10, 100)$). For each $\lambda_2 \in \Lambda_2$ we run the equivalent Group Lasso algorithm for $\gamma \in \Gamma$, where Γ is a set of candidate parameters for γ (e.g. $\Gamma = (0, 0.01\gamma_{\max}, 0.1\gamma_{\max}, \gamma_{\max})$, where $\gamma_{\max}$ is a value which is so high that no group gets selected.) Then we pick the pair $(\lambda_2^*, \gamma^*) = \arg\min \mathrm{BIC}(\lambda_2, \gamma)$, where

$$\mathrm{BIC}(\lambda, \gamma) = \frac{\|\hat{\mathrm{Y}} - \hat{\mathrm{X}}\beta_{\mathrm{GL}}(\lambda_2, \gamma)\|^2}{\mathrm{n}\sigma^2} + (\log(\mathrm{n})/\mathrm{n})\mathrm{df}_{\mathrm{GL}}(\lambda_2, \gamma), \tag{9}$$

where $\mathrm{df}_{\mathrm{GL}}$ is the degrees of freedom estimate for Group Lasso as proposed by [16], i.e.,

$$df_{\mathrm{GL}}(\lambda, \gamma) \approx \sum_{j=1}^{N} I(\|\hat{\beta}^j\|_2 > 0) + \sum_{j=1}^{N} \frac{\|\hat{\beta}^j\|}{\|\hat{\beta}^j_{\mathrm{OLS}}\|},$$

where $\hat{\beta} = \hat{\beta}_{\mathrm{GL}}(\lambda, \gamma)$ and $\hat{\beta}^{\mathrm{OLS}}$ is the ordinary least squares solution when using all the variables.

2.2 Extreme Value Modeling

2.2.1 Preliminaries: Extreme Value Modeling

We now give a brief review of extreme value theory [7]. We will show that a natural statistical model for the occurrence of extreme events is a Poisson point process that yields a generalized extreme value (GEV) distribution for the magnitude of the largest event in a fixed time period and a generalized Pareto distribution (GPD) for the amounts by which the magnitudes of extreme events exceed a specified threshold.

Let $X_1, \cdots, X_n$ be a sequence of independent and identically distributed random variables, and let $M_n = \max\{X_1, \cdots, X_n\}$. If there exist sequences of constants $a_n > 0$ and b_n such that

$$\Pr\left(\frac{M_n - b_n}{a_n}\right) \to G(z) \qquad \text{as } n \to \infty\,, \tag{10}$$

for some nondegenerate distribution function G, then G is a generalized extreme value distribution, with distribution function

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, \tag{11}$$

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, with $-\infty < \mu < \infty, \sigma > 0$, and $-\infty < \xi < \infty$.

If the limiting distribution (11) exists, then, for a large threshold u, the exceedance $Y = X - u$, conditional on $X > u$, is well approximated by a generalized Pareto distribution

$$H(y) = \Pr(X > u + y | X > u) = \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}, \tag{12}$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\}$, where

$$\tilde{\sigma} = \sigma + \xi(u - \mu)\,.$$

The parameters of the generalized Pareto distribution of threshold excesses are uniquely determined by those of the associated GEV distribution of block maxima.

These results provide two approaches for statistical modeling of extreme values. The block maxima (e.g. annual maxima of meteorological variables) can be modeled as independent observations from a GEV distribution, or the excesses over a high threshold can be modeled by a GPD. Both approaches have weaknesses. The GEV approach uses only one observation per block, which may be wasteful if more data than just the block maxima are available. In the GPD approach, the probability of exceeding the threshold is not available. These weaknesses can be overcome by formulating the behaviour of extreme events using a Poisson point process. This encompasses the GEV and GPD models and the process is completely defined by the same parameters that describe GEV distribution of block maxima. The model leads directly to a likelihood that enables a natural formulation of nonstationarity in threshold excesses, for example by including spatio-temporal correlation.

A point process on a set $\mathcal{A}$ is a stochastic rule that describes the occurrence and position of point events. For a set $A \subset \mathcal{A}$, we define the non-negative integer-valued random variable $N(A)$ to be the number of points in the set A. In a Poisson process the occurrence of events at different points $a \in \mathcal{A}$ is statistically independent and $N(A)$ has a Poisson distribution,

$$N(A) \sim \text{Poi}(\Lambda(A)), \tag{13}$$

with

$$\Lambda(A) = \int_A \lambda(a)\, da, \tag{14}$$

where the intensity function $\lambda(a)$, $a \in \mathcal{A}$, indicates the relative frequency of occurrence of events at different locations in $\mathcal{A}$. In extreme-value modeling the set $\mathcal{A}$ has the form $(-\infty, +\infty) \times [u, \infty)$, the two components respectively indicating time and event magnitude. The intensity function is

$$\lambda(t, x) = \sigma^{-1} \left[1 + \xi \frac{x - \mu}{\sigma}\right]^{-1/\xi - 1}, \tag{15}$$

which yields distributions of block maxima and of excesses over threshold u that have the forms (11) and (12) respectively.

2.2.2 *Spatio-Temporal Point Process*

Since the statistical characteristics of extreme climate data vary over space and time, the model specified by (13)–(15) cannot be used directly. We have therefore developed a more general version of the model, a spatio-temporal point process in which the location parameter μ and scale parameter σ are permitted to vary over space and time, and the threshold u varies over space.

To incorporate spatial and temporal correlation among the data, we build a hierarchical Bayesian spatio-temporal dynamic model [4]. This modeling strategy involves three stages. The first stage is the data model which models only observation process given a latent process. Stage 2 specifies the latent process; in our case, this is a Poisson point process and incorporates spatio-temporal dependence structures that are much more complicated than could be specified directly. In stage 3 we specify prior distributions for the parameters occurring in stage 2; here we can include external knowledge and expert opinion.

Let $X^i_{s,t}$ be the ith exceedance over threshold u_s at location s in year t, where $i = 1, \ldots, n_{s,t}, s = 1, \ldots, S$ and $t = 1, \ldots, T$. In the observation process, the likelihood function of the Poisson point process can be written as

$$\begin{aligned} &L(\mu_{s,t}, \sigma_{s,t}, \xi; X^1_{s,t}, \ldots, X^{n_{s,t}}_{s,t}, s = 1, \ldots, S, t = 1, \ldots, T) \\ &\propto \prod_{t=1}^{T} \prod_{s=1}^{S} \exp\left\{ - \left[1 + \xi \left(\frac{u_s - \mu_{s,t}}{\sigma_{s,t}} \right) \right]^{-1/\xi} \right\} \\ &\times \prod_{i=1}^{N} \sigma^{-1} \left[1 + \xi \left(\frac{x^i_{s,t} - \mu_{s,t}}{\sigma_{s,t}} \right) \right]^{-1/\xi - 1}, \end{aligned} \tag{16}$$

where $\mu_{s,t}$ and $\sigma_{s,t}$ are varied over space and time.

In the process model, we model the location parameter $\mu_{s,t}$ through a dynamic linear model and $\sigma_{s,t}$ is modeled in the same procedure:

$$\mu_t = B^\mu_s \theta^\mu_t + \epsilon^\mu_t, \tag{17}$$

where $\mu_t = (\mu_{1,t}, \ldots, \mu_{S,t})'$ at time t. θ^μ_t, a $K \times 1$ vector, is called state vector. B^μ_s in (17), is a $S \times K$ matrix which can reduce the spatial dimension from S to K ($K < S$). We choose B_s as a Matern kernel with fixed smoothness parameter [12]. ϵ_t is a random Gaussian process to include systematic error.

$$\theta^\mu_t = \Gamma^\mu \theta^\mu_{t-1} + \omega_t, \tag{18}$$

where Γ^μ in the transition equation is specified through an AR(1) process.

In stage 3, we assign noninformative priors to all the parameters. Given the data model (16), the process model (17) and (18), and the prior process, we can derive the posterior distribution. Markov Chain Monte Carlo (MCMC) algorithm can be used to draw sample from the full conditional distributions. The full conditional distributions of the variance parameters which characterize the random process ϵ^μ_t and ω^μ_t are inverse gamma distributions and can be drawn through Gibbs sampler. Some full conditional distributions of the parameters, such as μ_t and the temporal correlation parameters in Γ^μ, are hard to sample directly, and hence Metropolis-Hasting algorithm is used. θ_t are jointly sampled by forward filtering backward sampling (FFBS) algorithm [6]. After obtaining the MCMC samples, we can make inferences for the parameters in the model. We drew 15,000 samples and discarded the first 5,000. The chains were thinned by choosing every 10th samples to reduce the correlation. So 1,000 samples for each chain were left for analysis. Convergence was checked on trace plots of posterior samples.

It is usually more convenient to interpret extreme value models in terms of return levels, rather than individual parameter values. Let z_m be the return level associated with the return period m years; z_m is the level exceeded by the annual maximum in any particular year with probability $1/m$. Statistically, the return level is the $1/m$ upper quantile of generalized extreme distribution. Let n be the number of observations in a year, and z_m satisfies the equation

$$n \log p = \log(1 - 1/m),$$

where

$$p = 1 - n^{-1} \left[1 + \xi(z_m - \mu_{s,t})/\sigma_{s,t}\right]^{-1/\xi_i},$$

if $[1 + \xi(z_m - \mu_{s,t})/\sigma_{s,t}] > 0$, otherwise $p = 1$. Here $\mu_{s,t}, \sigma_{s,t}$, and ξ are the parameters of the point process for year t and location s. This equation can be solved for z_m using standard methods.

3. DATA

The mere amount of publicly available climate data is outright staggering. There are a large number of governmental and scientific institutions who publish measurements for a given geographical range on a multitude of relevant variables on the Web. This being said, it is nevertheless a major challenge to obtain consistent

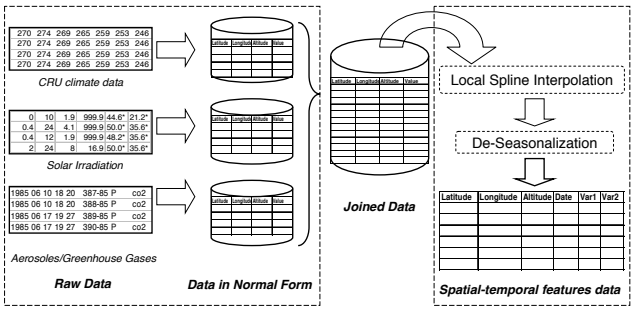

Figure 2: Data collection and pre-processing

longitudinal records that cover with comparable temporal and spacial resolution all relevant variables. Another problem is the large variety of formats in which data are available.

3.1 Data Sources and Collection

We compiled a comprehensive set of relevant variables for climate modeling in North America. Aside from the primary climate variables that we eventually wish to explain, the literature distinguishes human and natural agents or forcings that are known to affect the climate. These include solar irradiance and volcanic activities, greenhouse gases and aerosols (small particles dispersed in air). Figure 2 shows a schematic view of the data collection and preparation process. Table 1 lists the variables that we used in our analysis. We note that the "temperature extreme" variable is to be distinguished from all the others, in that they are *estimated*, using the extreme value modeling technique described in the previous section. We used for this study data from the following 5 sources:

1) **CRU:** Climate Research Unit provides monthly climatology data at http://www.cru.uea.ac.uk/cru/data for 11 surface variables including precipitation, wet-day frequency, mean, max, min temperature, vapor pressure, relative humidity, sunshine percent, cloud cover, frost frequency, wind speed from 1901 to present on a 0.5 degree latitude and longitude resolution. This grid data was interpolated from station data as a function of latitude, longitude, and elevation using thin-plate splines by New et al.[13]

2) **NOAA:** The data center http://www.cdc.noaa.gov/data/gridded/ of the National Oceanic and Atmospheric Administration is considered the "World's largest archive of climate data". We downloaded the greenhouse data from 170 worldwide stations from http://www.esrl.noaa.gov/gmd/dv/ftpdata.html.

3) **NASA:** NASA uses satellite images to estimates of the ambient aerosol optical thickness based on the resulting ultra-violet irradiation. We collected this data from http://iridl.ldeo.columbia.edu/SOURCES/.NASA/ .GSFC/.TOMS/.NIMBUS7/.

4) **NCDC:** The National Climate Data Center was our source for the different solar radiation measurements in 997 different locations at http://rredc.nrel.gov/solar/old_data/nsrdb/.

5) **CDIAC:** Daily temperature data are obtained from U.S. historical climatology network (http://cdiac.ornl.gov/epubs/ndp/ushcn/usa.html). The data of daily maximum temperature were collected from year 1948 to 2005 at 351 stations in U.S. We cleaned the data by removing invalid temperature observations.

3.2 Data Pre-Processing

The preparation of the data for the modeling involved a number of steps:

1) **Normalization:** Initially we transformed each dataset into monthly observations in a standard format including longitude, latitude, altitude, date, variable, value, unit, and source.

Variables (Variable group)	Type	Source
Methane (CH_4) Carbon-Dioxide (CO_2) Hydrogen (H_2) Carbon-Monoxide (CO)	Greenhouse Gases	NOAA
UV (AER)	Aerosol Index	NASA
Temperature (TMP) Temp Range (TMP) Temp Min (TMP) Temp Max (TMP) Precipitation (PRE) Vapor (VAP) Cloud Cover (CLD) Wet Days (WET) Frost Days (FRS)	Climate	CRU
Global Horizontal (SOL) Direct Normal (SOL) Global Extraterrestrial (SOL) Direct Extraterrestrial (SOL)	Solar Radiation	NCDC
1-year return level for temperature extreme (TMP.EXT)	Climate	Estimated using temp from CDIAC

Table 1: Variables and data sources.

2) **Interpolation and Smoothing:** We interpolated the data from NOAA and NCDC into a common 2.5x2.5 degree grid for North America to allow us to join multiple data sources. For this process we used thin plate splines on the monthly data to be consistent with the interpolation method used for the CRU data. Since the data from NASA and CRU were provided for a finer resolution grid, we performed spatial averaging to get data on the common 2.5x2.5 degree grid.

3) **De-seasonalization:** We performed de-seasonalization by removing seasonal averages.

4. EXPERIMENTS

As we noted in Introduction, we conduct two sets of experiments, one involving generic spatio-temporal data that are simulated from an artificial model, and the other involving the actual climate data we described in the previous section. The experiments involving real climate data consist of the following steps: 1) Using spatio-temporal extreme value modeling technique to estimate the 1-year return levels (1-year event magnitudes) of temperature; 2) Incorporating the estimated 1-year return levels as a proxy for extreme temperature in the spatio-temporal causal modeling using Group Elastic Net.

In the subsequent subsections, we describe the details of these experimental procedures and their results.

4.1 Simulation Experiments

We performed two sets of experiments on synthetic data to evaluate the performance of "Group Elastic Net" (which takes into account spatial interactions through spatial lagging and appropriate penalization in the regression), against that of a method that neglects such interactions and considers instead that a measurement at location s is only affected by variables at the same location. Specifically, the comparison method solves the following group Lasso

problem for each measurement x^i.

$$\hat{\beta} = \arg\min_{\beta} \sum_{s\in S}\sum_{t=L+1}^{T}(x^i_{t,s} - \sum_{j=1}^{N}\sum_{l=1}^{L}\beta^j_l x^j_{t-l,s})^2 + \lambda\sum_{j=1}^{N}\left(\sum_{l=1}^{L}(\beta^j_l)^2\right)^{1/2}. \quad (19)$$

We generated synthetic spatial-temporal data using a spatial-temporal vector autoregressive (VAR) model as generative model. More specifically, we considered $N = 10$ measurements $x^1, \ldots x^N$, taken on a 15×15 spatial grid. For each (interior) point $s = (s_1, s_2)$ we consider the neighborhood structure $\Omega = \{(\omega_1, \omega_2) \in \{-2, -1, 0, 1, 2\} \times \{-2, -1, 0, 1, 2\}$. We set the maximum lag $L = 3$. Let $\mathbf{x_{t,s}}$ denote the vector formed by all the measurements $x^i_{t,s}, i = 1, \ldots, N$. We considered the following generative model.

$$\mathbf{x}_{t,s} = \sum_{l=1}^{L}\sum_{\omega\in\Omega}\mathbf{A}_{l,\omega}\mathbf{x}_{t-l,s+\omega} + \eta.$$

The matrices $\mathbf{A}_{l,\omega}$ where generated as follows. We first generated an $N \times N$ adjacency matrix $\mathbf{A}$, where the entry $\mathbf{A}[i, j] = 1$ indicates that x^i causes x^j, and $\mathbf{A}[i, j] = 0$ otherwise. The value of each entry was chosen by sampling from a binomial distribution, where the probability that an entry equals to one was set to 0.2.

For the first set of experiments, we use a setup we call "random coefficient weighting." That is, for each pair (l, ω) and each i, j, we set $\mathbf{A}_{l,\omega}[i, j] = c_{l,\omega}(i, j) \cdot \mathbf{A}[i, j]$, where $c_{l,\omega}(i, j) \sim$ Unif$(-0.1, 0.1)$. For the second set of experiments, we use a different setup we refer to as "decaying coefficient weighting", which is meant to represent situations where the influence decays with the distance. Formally, we first focus on the central location $\omega_0 = (0, 0)$, and for each l and each i, j, we set $\mathbf{A}_{l,\omega_0}[i, j] = c_{l,\omega_0}(i, j) \cdot \mathbf{A}[i, j]$, where $c_{l,\omega_0}(i, j) \sim$ Unif$(-0.1, 0.1)$. Then for each pair $(l, \omega \neq \omega_0)$, for each i, j we set $\mathbf{A}_{l,\omega}[i, j] = \left(\frac{c_{l,\omega_0}(i,j)}{1+\|\omega\|} + \tilde{\eta}\right)\mathbf{A}[i, j]$, where $\tilde{\eta}$ is some random noise $\sim$ Uniform$(-0.01, 0.01)$.

For both sets of experiments the noise η was sampled according to a normal distribution $\mathcal{N}(0, 0.01)$. For each setup, we generated 10 models, and for each model simulated data for 100 time points. For Group Elastic Net, we used $\tilde{\Delta}_j = \mathrm{I}$ for the first set of experiments, and $\tilde{\Delta}_j = \mathrm{diag}(\exp(\|\omega\|/2), \omega \in \Omega)$ for the second (since in practice one may not know the exact type of distance based decay, e.g polynomial, exponential).

We measure the accuracy of each method with respect to their ability to correctly identify the underlying adjacency matrix $\mathbf{A}$. We report the average F_1 score along with standard deviation. The F_1 score is defined as $F_1 = 2\frac{PR}{P+R}$, where P is the precision and R the recall. The results are reported in Table 2. Under both settings, Group Elastic Net exhibits higher accuracy than the comparison method, and the difference in accuracy is greater for the "decaying coefficient weighting". This illustrates the importance of taking spatial interactions into account in the modeling.

Method	Grp Lasso	Grp Elastic Net
Random coef	0.53 ± 0.01	$\mathbf{0.60 \pm 0.01}$
Decaying coef	0.49 ± 0.02	$\mathbf{0.67 \pm 0.01}$

Table 2: The accuracy (F1) of two comparison methods: Group Lasso (no spatial interaction) and Group Elastic Net for spatial temporal VAR models with random and decaying coefficient weighting.

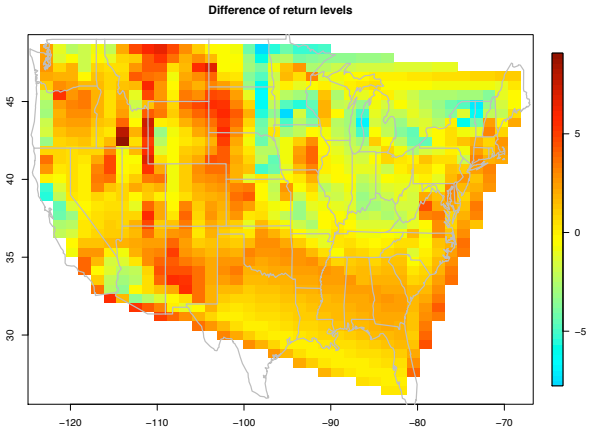

Figure 3: A comparison of average return levels from 1948 to 1980 and from 1981 to 2005.

4.2 Modeling extreme temperature

We used the daily temperature data from CDIAC for modeling extreme temperature. To obtain the exceedances over a threshold required for the modeling, we calculated the 95th quantile of the temperature distribution over the 58 years at each location and chose the observations which exceed the location-specific threshold. We removed the stations which have very few years with at least 2 exceedances. Thus the data used for our model are the daily maximum temperature exceedances for 58 years at 254 stations.

As discussed in Section 2.2.2, we interpret the extreme value analysis through return levels. We obtained the return level for years from 1948 to 2005 and each of the 254 stations. To investigate the evidence of global warming, for each station, we calculated the average return levels from 1948 to 1980 and from 1981 to 2005 and compared if there is any increase in terms of return level for these two periods. Figure 3 gives the return level difference which indicates the return levels increase over the past 58 years in midwest, western and eastern coastal areas of the United States. We observe a clear trend that the difference is mostly positive, with some of the regions exhibiting as much as 5 degrees Fahrenheit increase during this period.

4.3 Spatio-temporal modeling and attribution

We applied our spatio-temporal causal modeling method on two datasets: one monthly, the other yearly. Both contain data for 1990-2002 on a 2.5x2.5 degree grid for latitudes in $(30.475, 50.475)$, and longitudes in $(-119.75, -79.75)$. The monthly dataset contains the first 19 variables listed in Table 1. The annual dataset contains in addition the estimated return levels for the extreme temperature. Having two different time resolutions allows us to investigate short term and longer term influences. Note that since the return levels were estimated yearly, we did not incorporate them into the monthly dataset (Estimating monthly return levels will be the object of future work.)

For the spatial-temporal causal modeling, we used a 3x3 spatial neighborhood structure and a maximum time lag of 3 months for the monthly data, and 3 years for the yearly data. In our modeling, we considered the temperature variables as a group (TMP), as well as the solar variables (SOL), in addition to the natural grouping structure by spatial temporal series.

Figure 4 shows the results of attributing the changes in return level for extreme temperatures using the yearly dataset, while Figures 5 and 6 show the results on attributing the changes in temperature, using respectively the yearly and the monthly dataset. In assessing the *strength* of causal links identified in our outputs, we use two separate metrics. One is the l_2- norm of the regression coefficients corresponding to the variable group in question, which coincides with its contribution to the spatio-temporal penalty term in the Group Elastic Net modeling. The other is the *point* at which the causal link in question appears in the output graph, as the parameter dictating how much emphasis is placed on the model complexity penalty in BIC is varied. This is done by multiplying the estimated noise variance in the penalty term (σ^2 in Equation 9) by a varying constant, which determines the trade-off between the model fit and model complexity. (The noise variance estimate is known to add a certain degree of arbitrariness to BIC with finite samples.) Each of the figures exhibits several causal graphs corresponding to different values of this parameter, with models becoming denser going from left to right and top to bottom. Also in the figures the edge thickness represents the l_2- norm of the regression coefficient. It is apparent that the two measures coincide for the most part (order of appearance and edge thickness), and in particular, CO2 and other greenhouse gases are judged to have greater causal strength than solar radiance, according to both of these measures.

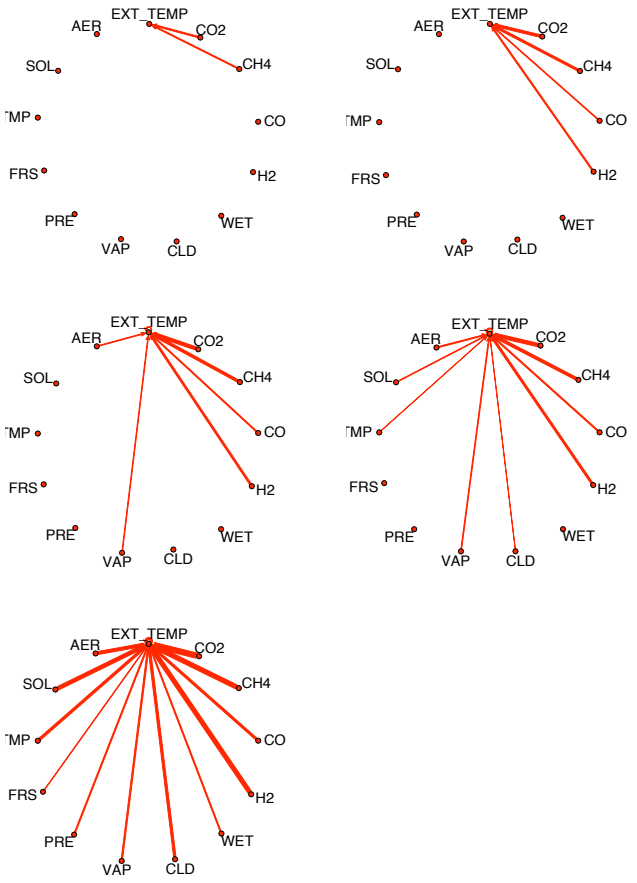

Figure 4: Attributing the change in 1-year return level for temperature extremes using annual data. Output causal structures for decreasing degrees of sparsity. Edge thickness represents the causality strength.

5. CONCLUDING REMARKS

In the present paper we initiated a data-centric approach to climate change attribution. The results to date are preliminary but encouraging, and in the future we plan to refine them, validate them with the domain experts, and explore ways in which they can provide assistance to the dominant, simulation-based, approach to climate modeling.

Acknowledgments

We would like to thank the following people for their contributions to this work in a variety of ways: Huijing Jiang, Elena Novakovskaia, Cezar Pendus, Saharon Rosset and Lloyd Treinish.

6. REFERENCES

[1] Climate change 2007 - the physical science basis *IPCC Fourth Assessment Report on scientific aspects of climate change for researchers, students, and policymakers*.

[2] Barnett, T.P., Pierce,D.W. and Schnur, R. (2001). Detection of anthropogenic climate change in the world's oceans. *Science*, 292.

[3] Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. New York: Wiley.

[4] Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Ration, Florida: Chapman & Hall.

[5] Christidis, N., Peter, S.A., Brown, S., Office, M. and Hegerl, J-C.G.C. (2005). Detection of changes in temperature extremes during the second half of the 20th century. *Geophys. Res. Lett.*, 32(L20716), 2005.

[6] Carter, C. K. and Kohn, R. (2001). On Gibbs sampling for state space models. *Biometrica*, 81, 541-553.

[7] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Berlin: Springer.

[8] Gillett, N.P., Zwiers,F.W., Weaver,A.J. and Stott, P.A. (2003). Detection of human influence on sea level pressure. *Nature*, 422(b).

[9] Granger, C. (1980). Testing for causlity: A personal viewpoint. Journal of Economic Dynamics and Control 2, 329-352.

[10] Karoly, D. J., Braganza, K., Stott, P. A., Arblaster, J.M. Meehl, Anthony, G.A., Broccoli, J. and Dixon, K.W. (2003) Detection of a human influence on north american climate. *Science*, 302.

[11] Luo, L. Wahba, G. and Johnson, D.R. (1998) Spatial-temporal analysis of temperature using smoothing spline anova. *J. Climate*, 11.

[12] Matern, B. (1960). *Spatial Variation*. New York: Springer.

[13] New, M., Hulme, M. and Jones, P.D. (1999) Representing twentieth century space-time climate variability. Part 1: development of a 1961-90 mean monthly terrestrial climatology. Journal of Climate 12, 829-856

[14] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58 (1), 267-288.

[15] P.A. Stott, D.A. Stone, and M.R. Allen. (2004) Human contribution to the european heatwave of 2003. *Nature*, 432.

[16] Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. J. R. Stat. B 68, 49-67.

[17] Zou, H., Hastie T. (2005) Regularization and variable selection via the Elastic Net. J. R. Statist. Soc. B 67(2) 301-320.

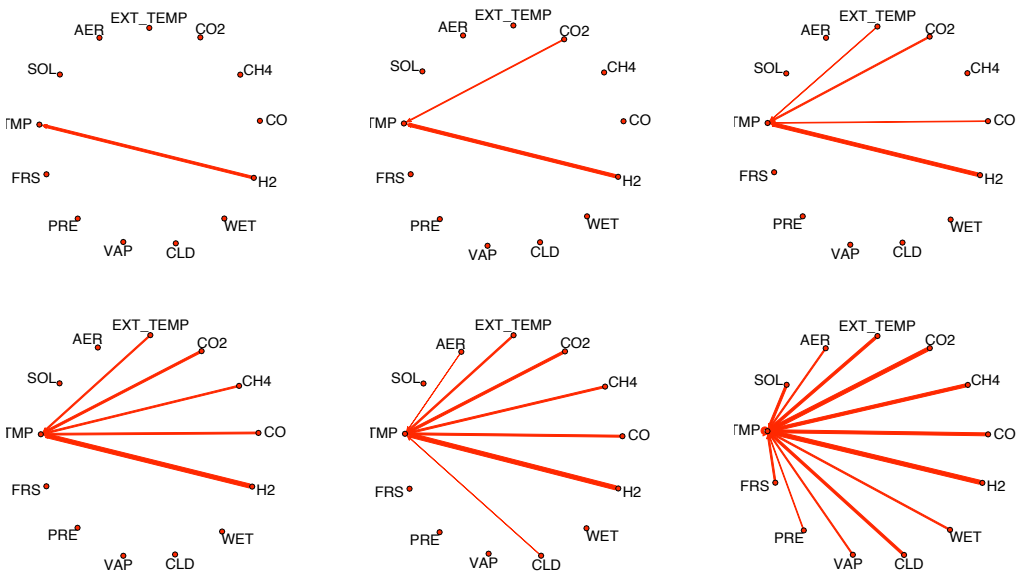

Figure 5: Attributing change in temperature using annual data. Output causal structures for decreasing degrees of sparsity. Edge thickness represents the causality strength.

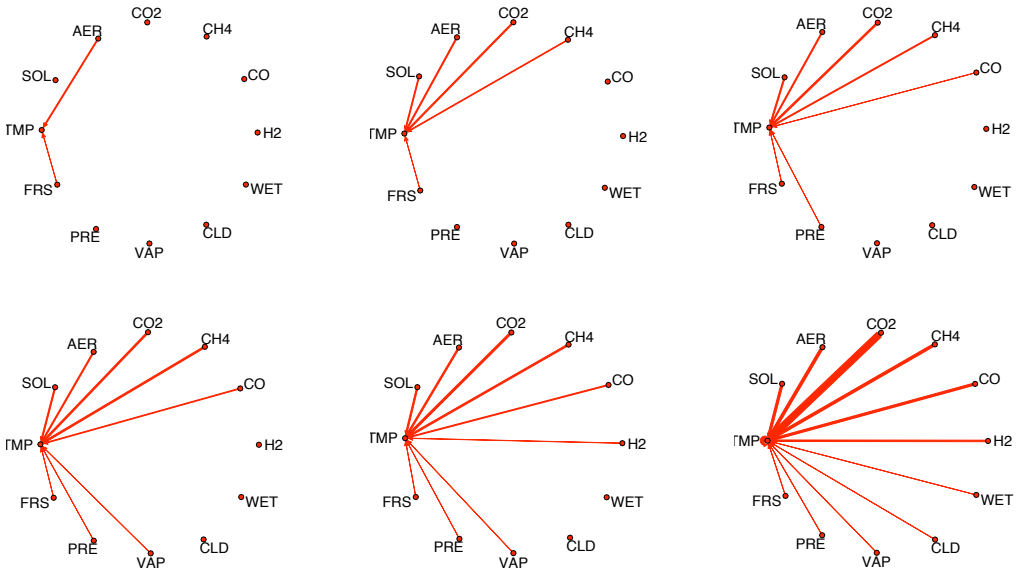

Figure 6: Attributing change in temperature using monthly data. Output causal structures for decreasing degrees of sparsity. Edge thickness represents the causality strength.