SPATIALLY PENALIZED REGRESSION FOR DEPENDENCE ANALYSIS OF RARE EVENTS: A STUDY IN PRECIPITATION EXTREMES

Debasish Das[1], Auroop Ganguly[2], Snigdhansu Chatterjee[3], Vipin Kumar[4] and Zoran Obradovic[1]

[1]Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA

[2]Civil and Environmental Engineering, Northeastern University, Boston, MA, USA

[3]School of Statistics, University of Minnesota, Minneapolis, MN, USA

[4]Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA

ABSTRACT

Discovery of dependence structure between precipitation extremes and other climate variables (covariates) within a smaller spatial and temporal neighborhood is an important step in better understanding the drivers of this complex phenomenon as well as short-term prediction of extremes occurrence. Apart from the inherent spatio-temporal variability of the dependence, it is further complicated by the availability of the covariates at different vertical levels. The above problem can be split into three different sub-problems. Firstly, a spatio-temporal neighborhood of influence has to be discovered, which can be different for different locations. Secondly, the dependence structure between the precipitation extremes and the covariates has to be discovered within this neighborhood and thirdly, it has to be investigated whether this dependence structure can be exploited for any predictive power. Climate scientists have already discovered some physics-based relations between some of the covariates (e.g. temperature, relative humidity, precipitable water etc.) and precipitation extremes. We are exploring data-dependent alternatives for these problems and any possibility of incorporating the physics-based relations into the resulting data model. In particular, we used elastic net-based sparse optimization technique which solves all three problems of neighborhood discovery, covariate dependence discovery and predictive modeling and at the same time maintains the interpretability of the resulting model. Preliminary results look promising and show potential for some interesting knowledge discovery. We are currently exploring non-linear correlations and the alternatives to combine the physics-based relationships into the data model.

***Index Terms*—** One, two, three, four, five

1. INTRODUCTION

Large-scale climate models solve a system of partial differential equations (PDEs) based on first principles, but also contain parameterizations for processes that are not so well understood. Unfortunately, processes pertaining to precipitation are among the least well understood and precipitation is not a state variable in the PDEs. In addition, precipitation is known to be extremely variable in space and time and the underlying processes are subject to thresholds and intermittences. However, as pointed out in the literature [1,2,3,10,11,12], precipitation extremes tend to have a dependence on atmospheric variables ranging from temperature, humidity and precipitable water, to updraft velocity and horizontal wind components. These atmospheric variables, which can be thought of as potential covariates for precipitation extremes, are often better predicted than precipitation itself. Thus, there have been (somewhat counter-intuitive) suggestions that precipitation extremes may be more predictable than precipitation mean processes, simply because the extremes may relate more directly to covariates that are better predicted from models. The prediction problem in the context of precipitation extremes therefore translates to extraction of information context from these covariates and translating them to predictive insights.

Dynamical downscaling based on regional climate models (RCM), while higher-resolution and physics-based, suffers from complex parameterizations and difficult boundary conditions. Statistical downscaling, which have used models ranging from simple linear regression to artificial neural networks, suffers from lack of interpretability.

One promising recent approach has been the development of physically-based approaches which attempt to relate the atmospheric covariates with precipitation extremes through what could be viewed as hypothesis-guided approaches. The physics operates at different scales or accounts for different processes than are handled within the large-scale computational models of climate, hence their added value [1,3,10]. While these approaches have demonstrated significant promise, they may not be able to leverage the full information content in atmospheric covariates and translate these to predictive insights, primarily because they have to rely on known physics-based hypotheses. The best approach would need to leverage the information content in the covariates through both the physics-based hypotheses and the data, while keeping the functional mappings between the covariates and the precipitation extremes interpretable and without losing the ability to generalize to non-stationary conditions.

2. METHODOLOGY

We propose sparse and spatially-penalized extremes regression as a way to fill this gap. Model parsimony is embedded into our formulations through sparse regularization and spatial penalties to reduce spurious or overly specific relations that may not generalize. As mentioned earlier, the entire problem can be divided into few sub-problems, First, the neighborhood of influence may depend on the selected location and prevailing climate and wind conditions. Here we select the neighborhood based on data rather than enforcing a specific shape or size *apriori*. Second, within that neighborhood, we need to find out a set of variables that contain information about target variable (precipitation extremes here) out of a larger pool of variables. As a first step, we consider linear dependence structures and leave nonlinear dependence analyses to future research. Dependence in climate data may often be reasonably well captured through linear or quasi-linear structures. However, in our proposed method we have combined both these problems into a single sparse regression learning problem with a spatial penalty. Third, in the context of precipitation extremes, percentile-based [1-2] and extreme value theoretic definitions [13,14] have been used. A fundamental issue is that extremes cannot be expected to follow the distribution of the original precipitation time-series since they represent distributional tails. Thus, transformations need to be constructed based on the statistical properties of the extreme values to make these values amenable to predictive modeling.

Here is a description of the problem from data-mining perspective. Let us denote the vector of precipitation extremes at a grid indexed by (i,j) on a certain geographical region of interest by $\boldsymbol{y}^{ij}$. Also, let us denote all candidate variables at (i,j)-th grid by $X^{ij} = \{X_1^{ij}, X_2^{ij} \dots X_M^{ij}\}$ and variables at all grids by $\mathbf{X} = \{X^{mn}\ \forall (m,n)\ \in S\}$ where S is the set of all grid-points within the region under consideration. We can combine variable and neighborhood selection into a single problem described as: For each variable $\boldsymbol{y}^{ij}$, we are required to find a set of variables/nodes $NE^{ij} = \{X_k \colon X_k \in \mathbf{X}\}$ so that $\boldsymbol{y}^{ij}$ is linearly dependent only on NE^{ij} and nothing else. We will use $\boldsymbol{y}^{ij}$ and $\boldsymbol{y}$ to refer the same variable. One way of dropping uninformative regressors is to use a penalized regression. Let $\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{\mathrm{t}=1}^{\mathrm{D}} \left(\mathbf{y}_\mathrm{t} - \mathbf{X}_\mathrm{t}^\mathrm{T}\boldsymbol{\beta}\right)^2$ be sum of squared residuals from a regression of $\mathbf{y}_\mathrm{t}$ on all available features, $\mathbf{X}_\mathrm{t}$ (D is number of data points). The solution to

$$\min_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_2^2 \qquad (1)$$

For $0 < \lambda < 1$ is the well-known ridge estimator that shrinks the least squares estimates of β_{ij} towards zero. Note that, $||\boldsymbol{\beta}||_2^2 = \sum_{p=1}^{L} \beta_p^2$, where L is length of vector $\mathbf{X}$, is the squared L_2 norm of $\boldsymbol{\beta}$. By the nature of the L_2 penalty, the ridge estimates will almost never be zero exactly. In consequence, uninformative predictors can still inflate prediction error variance.

Now, by replacing the L_2 penalty by an L_1 penalty $||\boldsymbol{\beta}||_1^1 = \sum_{l=1}^{L} |\beta_l\ |$ we get a LASSO estimator [7]. An important feature of the L_1 penalty is that some coefficient estimates can be exactly zero. As shown in [8], LASSO enjoys a `sparsity property'. However, if we simply replace the L_2 penalty by L_1 penalty, there are two problems. First, if $L > D$, LASSO can select at most D variables. Second, if there is a group of variables with high pair-wise correlation coefficients, LASSO tends to select only one variable from the group and does not care which one. But, these problems can be overcome if we use a convex combination of both L_1 and L_2 penalties. The resulting regression is called Elastic Net [6] and has the following objective function.

$$\min_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) + \lambda_1||\boldsymbol{\beta}||_1^1 + \lambda_2||\boldsymbol{\beta}||_2^2 \qquad (2)$$

In equation (2) we have a formulation of the problem of finding a spatio-temporal dependence structure between precipitation extremes and regional covariates at a particular grid-point in terms of coefficients $\boldsymbol{\beta}$. But in its current form it is missing important domain knowledge. Here, equal importance is given to covariates belonging to all neighbors as a potential feature irrespective of its distance from the grid-point for which the dependence structure is being estimated. We address this problem by letting the multipliers λ_1 and λ_2 depend directly on the normalized geodesic distance of the associated grid-point from which the corresponding feature belongs. The formulation is given by

$$\widehat{\boldsymbol{\beta}}^{\mathbf{ij}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{d=1}^{D(i,j)} \left(y_d^{ij} - \mathbf{X}_\mathbf{d}^{\mathbf{ij}\,\mathbf{T}} \boldsymbol{\beta}^{\mathbf{ij}}\right)^2 + \sum_{l=1}^{L} \lambda_{1l} \left|\beta_l^{ij}\right| + \sum_{l=1}^{L} \lambda_{2l} |\beta_l^{ij}|^2 \qquad (3)$$

where $\lambda_{1l} = \lambda_{10}.(g_l/g_l^{max})$

and $\lambda_{2l} = \lambda_{20}.(g_l/g_l^{max})$

Now, a variable belonging from a far away grid-point will have a larger penalty parameter than a variable belonging from a nearby grid-point and therefore will have a smaller probability of being assigned a non-zero β coefficient. Note that, this type of differential penalty can be used to incorporate a range of other domain knowledges into our data model.

3. RESULTS

We focus on NCEP-NCAR reanalysis data [9], which are climate reconstructions developed by assimilating multiple remote and in-situ sensor data into meteorological models. We used daily forecasts starting from 1948 until 2010

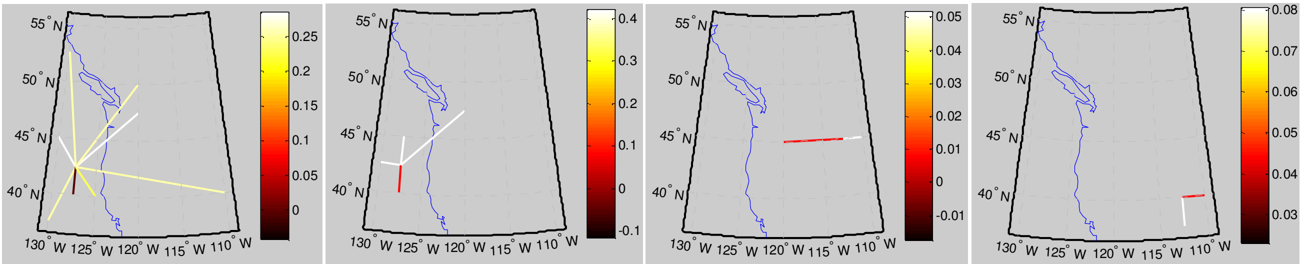

Figure 1: Graphical representation of spatial dependence of precipitation extremes on V-wind (a) without spatial penalty and (b) with spatial penalty for one grid-point and (c *and* d) temperature with spatial penalty for *two* different grid-points. The color signifies the strength of an edge.

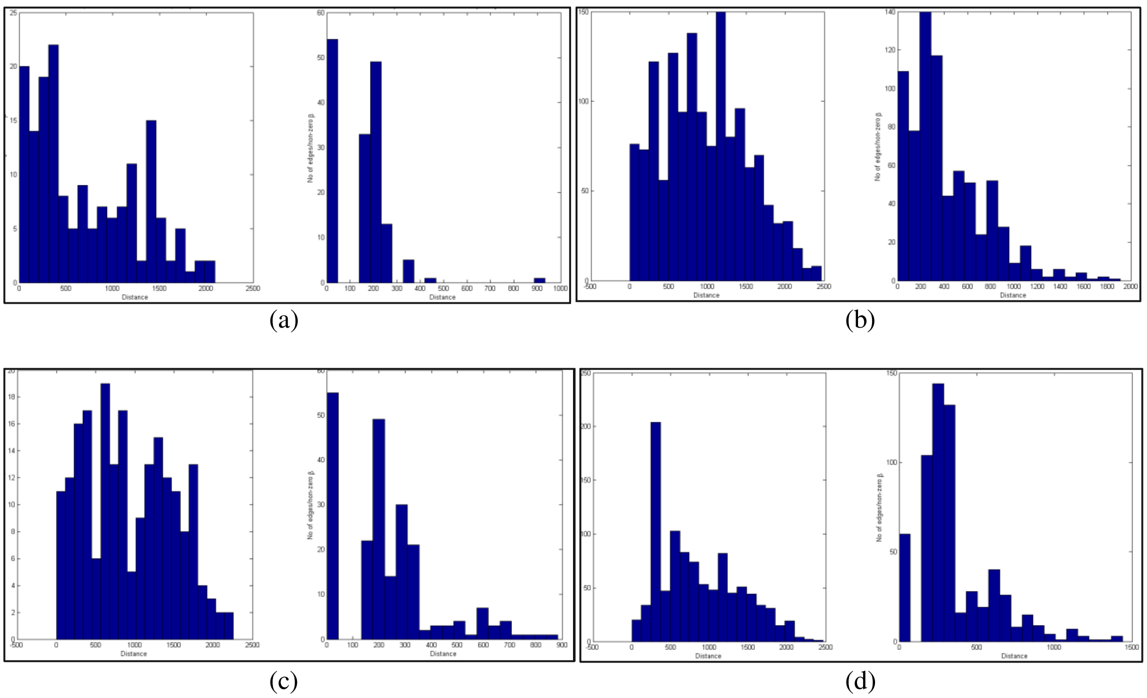

Figure 2: The distribution of edges for (a) Pressure; (b) Relative humidity; (c) Temperature and (d) V-wind.

for the following variables as potential covariates: i. Temperature (surface level); ii. Sea-level Pressure (surface level); iii. Relative Humidity (surface level); iv. Pressure (surface level); v. Precipitable Water (Entire Atmosphere); vi. Horizontal Wind Speed (North-south or V-wind). (surface level); vii. Horizontal Wind Speed (East-west or U-wind). (surface level); viii. Updraft Velocity (Omega) (surface level) and used covariates from the day the extreme occurred and 2 days before (total 3 days). We applied our method in North-west US for a proof-of-concept study. Our method showed promising results in terms of validating the insight presented in [1,2] which is obtained from hypothesis-guided physics models. However, our method shows more potential as it is able to discover the spatio-temporal variability of the existing relations, and at the same time discover the unknown relations. In figure 1, we have presented a few example graphs obtained using our method that shows the influence of V-wind (fig 1.b) and temperature (Fig 1.c and 1.d) at different neighboring grid-points on the precipitation extremes at a particular grid-point. From figures 1.c and 1.d, we can conclude that the dependence of precipitation extreme on temperature has a spatial variability. However, influence of V-wind on precipitation extremes can be considered as a novel insight. Moreover, in figure 1.a and 1.b, we show the benefit of using the spatial penalty in terms of achieving model parsimony.

We can represent the non-zero $\boldsymbol{\beta}$-values as edges connecting two nodes where one node represents the precipitation extremes in the grid-point on which the elastic net model is currently being trained and the other node is one of the potential features belonging from one of all the available grid-points (this includes the grid-point on which model is being trained). So, if there are $|S|$ total grid-points in the target region, we will have a total of

3x8x|*S*| possible **β**-values (however, most of them will be zero for a sparse model) for each grid. Again, we have one such model for each of the |*S*| grid-points. So, altogether there can be total of possible $24|S|^2$ **β**-values or edges. Figure 2 shows these edge distributions, as a function of the distance between the grid-points they connect, before and after using spatial penalties. We only present this for the NW US owing to lack of space, but this kind of analysis can be done for any target region. The distance will be zero for a non-zero **β** that connects with a variable in the same grid-point where the model is being trained. The plots are separated according to the covariates they correspond to. We can see that adding the spatial penalty results in more parsimonious models which are more easily interpreted by the domain scientists, while accuracies of the models remain intact. Some of the interesting information available from these plots about NW US are as follows: (a) winds, both vertical and horizontal, influence the precipitation extremes from a large number of neighboring grid-points, (b) pressure from neighboring grid-points has very small influence on precipitation extremes, and (c) both temperature and precipitable water have more localized influence on precipitation extremes.

4. FUTURE WORKS

Future research needs to consider non-linear dependencies inherent in the climate system, include atmospheric covariates in the vertical layer and incorporate the physical relations that have been developed in climate science, perhaps as pre-processors to the data algorithms. Combining the grid-based regression models and letting them share information is another direction. Statistical properties (including uncertainty quantification) of the sparse regression models that focus exclusively on extremes need to be examined. Combining the spatiotemporal neighborhood-based predictions with teleconnections, specifically the influence of ocean-based oscillators, for an overall assessment of precipitation extremes could be a way forward.

11. REFERENCES

[1] M. Sugiyama, H. Shiogama, and S. Emori. Precipitation extreme changes exceeding moisture content increases in MIROC and IPCC climate models. *Proc. Nat. Acad. Sci.* USA, 107(2):571–575, 2010.

[2] P. A. O'Gorman and T. Schneider. The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *Proc. Nat. Acad. Sci.* USA, 106(35):14773–14777, 2009.

[3] P. A. O'Gorman and T. Schneider. Scaling of precipitation extremes over a wide range of climates simulated with an idealized GCM. *Journal of Climate* 22, 5676-5685

[4] Q. Schiermeier. The real holes in climate science. *Nature* 463, 284-287 (2010).

[5] S. Ghosh, D. Das, S.C. Kao and A. R. Ganguly, Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes, *Nature Climate Change*, doi:10.1038/nclimate1327, 2011.

[6] H. Zhou and T. Hastie, Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser.* B 67 301–320, 2005.

[7] R Tibshirani, Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser.* B 58 267–288. 1996.

[8] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist Assoc.* 96, 1348-1360, 2001

[9] Kalnay et al., The NCEP/NCAR 40-year reanalysis project, *Bull. Amer. Meteor. Soc.*, 77, 437-470, 1996

[10] C. J. Muller and P. A. O'Gorman. An energetic perspective on the regional response of precipitation to climate change. *Nature Climate Change,* 2011.

[11] S. C. Liu et al. Temperature dependence of global precipitation extremes. *Geophysical Research Letters*, vol. 36, L17702, 2009.

[12] R. P. Allan and B. J. Soden. Atmospheric Warming and the Amplification of Precipitation Extremes. *Science*, 2008.

[13] V. V. Kharin et. al. Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *Journal of Climate*, 20, 1419—1444, 2007.

[14] S. C. Kao, and A. R. Ganguly, Intensity, duration, and frequency of precipitation extremes under 21st-century warming scenarios, *Journal of Geophysical Research - Atmospheres*, vol. 116, D16119, 14 pp., 2011.

[15] B. Efron, T. Hastie, I. Johnstone, and R.Tibshirani, (2004), 'Least angle regression', The Annals of Statistics 32, 407–499

[16] J. Friedman, T. Hastie and R. Tibshirani, Regularized Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 2010.