A Bayesian Sparse Generalized Linear Model With an Application to Multiscale Covariate Discovery for Observed Rainfall Extremes Over the United States

Debasish Das, Auroop R. Ganguly, and Zoran Obradovic

***Abstract*—Predictive insights on extreme and rare events are important across multiple disciplines ranging from hydrology, climate, and remote sensing to finance and security. Characterizing the dependence of extremes on covariates can help in identification of plausible causal drivers and may even inform predictive modeling. However, despite progress in the incorporation of covariates in the statistical theory of extremes and in sparse covariate discovery algorithms, progress has been limited for high-dimensional data where the number of covariates is large. In this paper, we propose a general-purpose sparse Bayesian framework for covariate discovery based on a Poisson description of extremes frequency and a hierarchical Bayesian description of a sparse regression model. We obtain posteriors over regression coefficients, which indicate dependence of extremes on the corresponding covariates, using a variational Bayes approximation. Experiments with synthetic data demonstrate the ability of the approach to accurately characterize dependence structures. The method is applied to discover the covariates affecting the frequency of precipitation extremes obtained from station-level observations over nine climatologically homogeneous regions within the continental U.S. The candidate covariates at multiple spatial scales represent station-level as well as regional and seasonal atmospheric condition, indices that attempt to capture large-scale ocean-based climate oscillators and hence natural climate variability, as well as global warming. Our results confirm the dependence structures that may be expected from known precipitation physics and generate novel insights, which can inform physical understanding and perhaps even predictive modeling.**

***Index Terms*—Bayesian model, climate change, covariate discovery, sparse regression, statistical downscaling.**

I. Introduction

DESCRIPTIVE analysis and predictive modeling of extreme values are growing in importance and urgency across societal priorities, ranging from remote sensing, natural hazards, geosciences including hydrology, weather, and climate to security of physical and cyber systems, as well as financial markets [1], telecommunication signals, and even relatively newer fields like social media [2]. Extremes are particularly relevant for climate due to the connection of consistent change in extreme weather patterns with the warming climate, as identified by the Intergovernmental Panel on Climate Change in their Special Report on Extremes [3]. Characterizing the dependence of extreme events on multiple covariates is often a first step in developing causal insights involving physical and empirical relations.

Both the frequency and intensity of regional rainfall extremes are undeniably showing change in the recent times, leading to increased number of flood events and related catastrophes around the world. Statistically significant increase in heavy to extreme rainfall has been shown in scientific studies over regions spanning multiple continents including Europe [4]–[6], Asia [7]–[9], Australia [10], [11] and the U.S. [12]–[14]. However, an accurate understanding of the sources of variability and change patterns of extreme rainfalls is necessary for management and adaptation strategies to succeed. Identifying the covariates with known pattern of variability on which extreme rainfall shows statistical dependence may be immensely helpful in predicting future change in extreme rainfall patterns and designing adaptation strategies. To that end, a large number of studies exist, which relate variability of rainfall extremes with large-scale climate indices ("teleconnections") with predictable pattern of fluctuation. Examples include linking change in European winter extremes to the change in North Atlantic Oscillation (NAO) [15], [16], relation of Atlantic Multi-decadal Oscillation with rainfall (not extremes) and river flows in the U.S. [17], [18] and Mexico [18], influence of Pacific Decadal Oscillation (PDO) on Arizona winter precipitation [19], ENSO influence on intraseasonal extreme rainfall frequencies in the U.S. [20], and influence of three different climate indices on Eastern Mediterranean rainfall extremes (using climate model simulations) [21] among many others. Although these studies vary in terms of analytic methods, these can be roughly categorized as statistical methods driven by specific hypothesis about dependence between two variables to prove or disprove the hypothesis. At their core, these are efforts to link the temporal trends in rainfall extremes with the regular temporal pattern of variations shown by the large-scale indices. However, while teleconnections are important, local and regional atmospheric conditions typically tend to dominate in the context

Manuscript received October 1, 2013; revised June 5, 2014 and December 12, 2014; accepted April 23, 2015. This work was supported by the National Science Foundation Expeditions in Computing Grant "Understanding Climate Change: A Data Driven Approach," award number 1029166.

D. Das is with the Big Data Analytics Group, Verizon Communications, New York, NY 10013 USA (e-mail: dave.das@verizon.com).

A. R. Ganguly is with the Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115 USA.

Z. Obradovic is with the Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA 19122 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2015.2445911

of climate-related extremes. Spatial variation in hydrologicz extremes at regional level has been studied using methods like regional frequency analysis [22] where the distribution of extremes over space has been assumed to follow the same distribution except for a site-specific scaling factor, and attempts were made to link the variation of extremes with a spatial explanatory variable [14], [23]. Only a small number of studies consider both spatial and temporal variations of extremes and attribute them to the appropriate explanatory variables. Kriging was used in [10] to account for the spatial correlation among extremes while modeling the temporal change. However, no spatial explanatory variable was used. On the other hand, Gregersen *et al.* [24] used regional spatial explanatory variables to handle spatial correlation in addition to using large-scale indices to explain temporal variability of the frequencies of rainfall extremes over Denmark and estimated the contribution of each explanatory variables (covariates) toward explaining spatial or temporal variation using a unregularized multivariate regression technique. While we borrow their approach of using regional-scale spatial variables to explain the spatial variability of the rainfall frequencies, our goal is to use as many variables as possible—both atmospheric variables at multiple spatial scales and large-scale climate indices—to use as potential covariates to explain any possible variability in extreme frequencies. The intent behind the choice of covariates at different spatial scales is to capture the influence of atmospheric processes at different spatial scales on precipitation extremes.

However, it is well known that solutions to unregularized multiple regression are not identifiable when the number of potential covariates is larger than the number of samples. Linear regression methods are generally trained to only minimize the prediction error on the target variable and therefore often infer nonzero coefficients for irrelevant covariates due to overfitting, even when the number of samples is larger than the number of potential covariates. Using a regularizer that constrains the L2-norm (ridge regression) of the coefficients will shrink the coefficient values, but the resulting regression model still suffers from overfitting as it also infers nonzero coefficients for the irrelevant covariates. The solution is to use a regularizer that constrains the L1-norm of the coefficient vector and thereby induces a sparse solution where the coefficients tend to shrink toward zero if the corresponding covariate is irrelevant. The theoretical and experimental proof of this advantage of the L1 regularizer over the L2 regularizer can be found in [25], which shows that sample complexity (minimum number of samples to learn a correct model) grows at least linearly with the number of irrelevant features for the L2 regularizer, while sample complexity grows only logarithmically for the L1 regularizer. Since we are interested in developing a generic statistical framework for modeling extremes, we expect to have only few training data in many applications while dealing with a large number of covariates. Therefore, the L1 regularizer is a much better choice in terms of avoiding overfitting.

Sparse models are more interpretable in the presence of a large number of potential covariates since it selects relevant covariates only. From the view-point of bias-variance tradeoff, the L1 regularizer is more restrictive than L2 and therefore reduces variance of the model by increasing the bias. As a result, L1-regularized models generalize better than the L2-regularized models. Sparse regression techniques based on L1 regularizers such as LASSO [26], basis pursuit [27], and elastic net [28] have gained prominence as tools for finding dependence between variables in many application areas due to the development of efficient algorithms [26], [29] that can handle millions of potential feature variables. These methods have been used extensively for covariate selection tasks in many fields starting from healthcare [30], [31] and biology [32], [33] to signal and image processing [34], [35] (in compressed sensing) and geosciences [36]. Although the sparse regression methods are overwhelmingly used to model target variables that are Gaussian, these methods have been extended to generalized linear model (GLM) framework to fit variables that are not Gaussian [37]. However, the ability of the sparse GLM algorithms to recover the true underlying sparse dependence structure depends heavily on the choice of tunable hyperparameter used in the methods. Here, we used a Bayesian hierarchical description of the sparse GLM that has three major advantages over the widely used non-Bayesian techniques. First, unlike non-Bayesian approaches, no hyperparameters are required to be tuned here. Second, non-Gaussian target distributions are accommodated naturally into the hierarchical Bayesian framework, and third, unlike non-Bayesian methods, which provide a deterministic estimate of the regression coefficients, Bayesian methods provide probability distributions over the regression coefficients. As evidenced later by the experimental results on synthetic data, this additional information about the coefficients facilitates a statistically sound inference of dependence structure and thereby drastically increases the accuracy of the resulting sparse solution. Although Bayesian versions of sparse regressions [38], [39] and its extension to accommodate GLM [40] were proposed in literature before, the later used expectation propagation method which is not guaranteed to converge and often produces nonsparse posterior even with a sparse prior. We developed a variational Bayes (VB) inference algorithm to estimate the posteriors, which is both fast and guaranteed to converge. We obtained a sparse solution empirically by using t-test on the posterior distribution of the regression coefficients for their significance.

The goal of this paper is to correctly identify the covariates that explain the variability of the precipitation extremes frequency, while predicting the extremes frequencies is out of the scope of this paper. We applied our method for covariate discovery for rainfall extremes as neither intensity nor frequency of the extremes follows Gaussian distribution. The statistical approach developed to model extremes is the extreme value theory (EVT) [41] that provides a natural family of probability distributions for modeling extremes. We will discuss this briefly in the next section. However, we need to be careful not to infer causality directly from the statistical dependence relations discovered by our method. The proposed method can be regarded as a selector of novel hypotheses out of a pool of a large number of potential ones which can then be validated using the physical understanding of causality. Our hope is that this will lead to better understanding of the processes that affect the changes in the pattern of precipitation extremes. We would also need to be careful in interpreting the significance of covariates that are highly correlated. This is an issue with any regression problem, which is inherited by the Bayesian approaches as well. Unregularized ordinary least square solutions assume no linear dependence (and therefore no significant linear correlation) among the predictors, and it fails if the predictors are highly correlated since the design matrix becomes singular in that case. Regularized least squares do not suffer from that problem. L2 regularizers tend to group the correlated predictors by assigning them similar coefficients but fail to produce a sparse solution as mentioned before. On the other hand, L1 regularizers tend to randomly assign a nonzero coefficient to one of the predictors out of a group of correlated ones whereby assigning zero coefficients to others in that group. The predictor selected may change with even a slight change in the regularization parameter, which is generally tuned in a heuristic manner. However, the Bayesian version of the L1 regularizer does not suffer from this inconsistency since there is no hyperparameter to be tuned. Since the presence of correlated climate variables is an important issue for our application, we will discuss more on the implication of having correlated covariates later when we will describe the data set.

Despite their importance, problems related to extremes were rarely addressed by the knowledge discovery community. In [42], a group elastic net method was described, which discovers Granger causality between a set of time-series variables which was used for climate change attribution. However, it is still a method for analyzing dependence among Gaussian-distributed variables. In [2], a latent variable model was used to obtain dependence between a set of extremes time-series of a single climate variable, but no covariates were considered.

1) Our Contribution: We propose a general Bayesian framework to adapt the L1-regularized sparse regression (LASSO) approach for dependence discovery for frequency of extremes. We regarded the average rates of the extremes frequency as latent variable and estimated their linear dependence on the covariates using the Bayesian version of the L1-regularized regression. We derived a fully Bayesian description of the joint distribution of observed frequencies and all latent variables and developed an efficient variational approximation algorithm to obtain the posterior distribution of the regression coefficients. This enabled us to discover the relevant features along with an estimate of uncertainty. The proposed framework can be applied as a statistical tool for dependence analysis of extremes in many other fields where extremes are important.

From the climate science perspective, to our knowledge, this is the first attempt to use Bayesian sparse regression technique for improved understanding of frequencies of rainfall extremes by linking their variability with a large number of atmospheric variables at different spatial scales as well as with large-scale climate indices.

II. Methods and Data

A. Basics of Extreme Value Modeling

The probabilities of extreme events are represented by the tail of distributions. As mentioned earlier, the statistical properties of extremes are described by EVT [41]. Extreme events are generally characterized by their intensity, duration, and frequency. The intensity of extremes may be described by either generalized extreme value distribution or generalized Pareto distribution depending on whether samples are obtained using block maxima (e.g., annual maxima) or peaks-over-threshold approach. On the other hand, the frequency and the interarrival time of extremes are modeled by Poisson point process approach where the number of extremes N within a fixed period is described by Poisson distribution [41], which is given by the following pdf:

$$\text{Pois}(N|\lambda) = \frac{\lambda^N e^{-\lambda}}{N!} \tag{1}$$

where λ is the average rate of occurrence of extremes within the unit period of time. In this paper, we only considered the frequency or number of occurrences of extremes, and modeling intensity is left as a future work.

B. Problem Setup and the Bayesian Framework

Let us denote the observed numbers of occurrences (frequencies or counts) of extremes in n distinct years by $\{N_i : i = 1, \ldots, n\}$ for a given location. We also have n observations of p covariates that may or may not influence the frequency of extremes. They are given by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i$ is a p-dimensional vector. Now, we assume that each of these frequencies N_i is drawn from distinct Poisson distributions with fixed but unknown average rates $\lambda_1, \lambda_2, \ldots, \lambda_n$. Given the aforementioned observations, our goal is to find a sparse subset of covariates (we may use "feature variables" or "predictors" alternately to mean covariates) that influence the frequencies of extremes. In order to obtain a robust, interpretable, and therefore more generalizable model, a natural choice is a sparse linear model. However, directly modeling the relation between observed frequencies and the feature variables using a linear model may not be appropriate since the relation may well be nonlinear. Instead, we use a GLM [43], where the average rate parameters of Poisson distributions are regarded as latent variables that are more stable than the actual observed frequencies and can be described reasonably well by a linear model. Therefore, instead of modeling the actual observations, we model the relationship between the latent variables and the feature variables using a sparse linear regression model. However, we model $\log \lambda_i$ instead of λ_i in order to enforce positivity on λ_i (also, $\log \lambda_i$ is a natural parameter for Poisson distribution). The final linear model is given by

$$\log(\lambda_i) = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i + \varepsilon, \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_1^1 \leq t \tag{2}$$

where $\varepsilon \sim \mathcal{N}(0, \tau^{-1})$. Here, $\boldsymbol{\beta} = [\beta_1 \beta_2, \ldots, \beta_p]$ is the vector of regression coefficients corresponding to p different features. The constraint $\|\boldsymbol{\beta}\|_1^1 \leq t$ refers to the L1 regularization. In non-Bayesian approach, that constraint term becomes the part of the loss function $\mathcal{L}$, which is given by $\mathcal{L} = \sum_i (\log \lambda_i - \boldsymbol{\beta}\mathbf{x}_i)^2 + \gamma \|\boldsymbol{\beta}\|_1^1$. Now, we denote $\psi_i = \log \lambda_i$, and the joint distribution of the frequencies N_i and parameters ψ_i is given by

$$p(N_i, \psi_i|\boldsymbol{\beta}, \tau, \mathbf{X}) = \prod_{i=1}^{n} p\left(N_i|e^{\psi_i}\right) p(\psi_i|\boldsymbol{\beta}, \tau, \mathbf{x}_i). \quad (3)$$

The distribution $p(N_i|e^{\psi_i})$ can be modeled by a Poisson distribution, and $p(\psi_i|\boldsymbol{\beta}, \tau, \mathbf{x}_i)$ can be modeled by the Gaussian given by $\mathcal{N}(\boldsymbol{\beta}^T\mathbf{x}_i, \tau^{-1})$. In order to implement sparsity, we can use a Laplace prior over $\boldsymbol{\beta}$ [38], [39]. The linear dependence structure between the logarithm of average rates and the feature variables is captured by the values of $\boldsymbol{\beta}$. A zero value of any of the β_j means that the extremes frequencies have no dependence on the corresponding feature, while a nonzero β_j that is statistically significant means otherwise. We can find the parameters $\boldsymbol{\beta}$ by maximizing the log-marginal $p(\{N_i\}|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{X})$ with respect to the parameters. However, directly optimizing the aforementioned log-marginal will not be feasible as it involves marginalization of the joint distribution in (3) over the latent variables. Instead, we propose to use a VB approximation algorithm [44] to find an approximate posterior distribution over each of the parameters which is described in the following section.

C. Variational Approximation for Bayesian Sparse Regression

We denote the sets $\{\psi_i\}$ and $\{N_i\}$ by vectors $\boldsymbol{\Psi}$ and $\mathbf{N}$, respectively. Let us also denote all of the unknown parameters by $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}\}$ and regard them as latent variables as well. A set of latent variables is therefore given by $\mathbf{Z} = \{\boldsymbol{\Psi}, \boldsymbol{\Theta}\}$. Moreover, from now on, we will ignore observed feature variables $\mathbf{X}$ from the list of conditioning variables. From a Bayesian view-point, the joint distribution $p(\mathbf{N}, \mathbf{Z})$ can be factorized in conditionals by assuming priors over each of the parameters in a hierarchical fashion. As mentioned earlier, sparsity over $\boldsymbol{\beta}$ can be implemented by enforcing a Laplace prior on the components of $\boldsymbol{\beta}$. We use the following scale-mixture form of Laplace prior [39] to make the inference tractable

$$\begin{aligned} p(\boldsymbol{\beta}; \tau, \boldsymbol{\gamma}) &= \prod_{j=1}^{p} \frac{\sqrt{\gamma_j \tau}}{2} \exp\left(-\sqrt{\gamma_j \tau}|\beta_j|\right) \\ &= \prod_{j=1}^{p} \int \mathcal{N}\left(\beta_j; 0, \tau^{-1}\alpha_j^{-1}\right) \text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) d\alpha_j \end{aligned} \quad (4)$$

where τ is the precision parameter of the additive noise in (2), α_j is the latent precision parameter for the coefficient β_j, and InvGa(.) is the Inverse gamma distribution. Moreover, gamma priors are imposed on τ and feature-specific individual penalty parameters γ_j [38], [39] with parameters (a_0, b_0) and (c_0, d_0), respectively. The priors are chosen for their conjugacy property that ensures tractability of inference. We typically assign $a_0 = b_0 = c_0 = d_0 = 10^{-6}$, although other values of these parameters will lead to the same estimate of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}$ does not depend on these parameters. The final joint distribution $p(\mathbf{N}, \mathbf{Z})$ is given by

$$\begin{aligned} p(\mathbf{N}, \mathbf{Z}) = \prod_{i=1}^{n} &\text{Pois}\left(N_i; e^{\psi_i}\right) \mathcal{N}\left(\psi_i; \boldsymbol{\beta}^T\mathbf{x}_i, \tau^{-1}\right) \\ &\times \text{Ga}(\tau; c_0, d_0) \prod_{j=1}^{p} \mathcal{N}\left(\beta_j; 0, \tau^{-1}\alpha_j^{-1}\right) \\ &\times \text{InvGa}\left(\alpha_j; 1, \frac{\gamma_j}{2}\right) \text{Ga}(\gamma_j; a_0, b_0). \end{aligned} \quad (5)$$

Computing the posterior $p(\boldsymbol{\Psi}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}|\mathbf{N})$ on all of the latent variables is analytically intractable due to the variables τ and α being coupled in the likelihood. Instead, we assume a factorized form variational approximation for the posterior distribution given by

$$\begin{aligned} p(\boldsymbol{\Psi}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}|\mathbf{N}) &= q(\boldsymbol{\Psi}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \\ &= q_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}) q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) q_{\tau}(\tau) q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}). \end{aligned} \quad (6)$$

We then solve for each component using a mean-field variational method [43]–[45]. Using Jensen's inequality, it can be shown that, given any arbitrary distribution $q(\boldsymbol{\Psi}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})$, the log-marginal $\log p(\mathbf{N})$ is lower bounded by the term

$$J = \int q(\boldsymbol{\Psi}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \log \frac{p(N, \boldsymbol{\Psi}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{q(\boldsymbol{\Psi}, \boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, \boldsymbol{\gamma})} d\boldsymbol{\Psi} d\boldsymbol{\beta} d\boldsymbol{\alpha} d\boldsymbol{\gamma} d\tau. \quad (7)$$

Now, by maximizing J with respect to each component of the variational distribution given in (6), we find the following update equations for each of the components.

1) Distribution of $\boldsymbol{\Psi}$:

$$q_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}) = \prod_{i=1}^{n} q_{\boldsymbol{\Psi}}(\psi_i) = \prod_{i=1}^{n} \frac{1}{Z_i} e^{\left\{\tilde{N}_i\psi_i - \frac{\langle\tau\rangle}{2}\psi_i^2 - e^{\psi_i}\right\}} \quad (8)$$

with $\tilde{N}_i = N_i + \langle\tau\rangle\mathbf{x}_i^{\mathrm{T}}\langle\boldsymbol{\beta}\rangle$.

Here, $\langle f(u)\rangle$ denotes expectation taken with respect to the variational distribution of random variable u [i.e., $\langle\tau\rangle$ denotes expectation of τ taken with respect to $q_\tau(\tau)$]. In order to make the inference tractable, we used Laplace method to approximate $q_{\boldsymbol{\Psi}}(\psi_i)$ by a Gaussian whose mean is given by the mode $\hat{\psi}_i$ of $q_{\boldsymbol{\Psi}}(\psi_i)$, which is obtained by maximizing $f(\psi_i) = e^{\{\tilde{N}_i\psi_i - (\langle\tau\rangle/2)\psi_i^2 - e^{\psi_i}\}}$ with respect to ψ_i. Solving $(df(\psi_i)/d\psi_i)|_{\psi_i=\hat{\psi}_i} = 0$, we get $\hat{\psi}_i = \tilde{N}_i/\langle\tau\rangle - \mathcal{W}(\exp(\tilde{N}_i/\langle\tau\rangle))/\langle\tau\rangle$, where $\mathcal{W}(.)$ is Lambert's W-function. Now, making a second-order Taylor expansion of $\ln f(\psi_i)$ around $\hat{\psi}_i$ and then exponentiating both sides, we get $f(\psi_i) = f(\hat{\psi}_i)\exp\{-(A/2)(\psi_i - \hat{\psi}_i)^2\}$, where $A = -(d^2/d\psi_i^2)\ln f(\psi_i)|_{\psi_i=\hat{\psi}_i} = (|\langle\tau\rangle + e_i^{\psi}|_{\psi_i=\hat{\psi}_i})^{-1}$. Note that the first-order term vanishes since $\hat{\psi}_i$ is a local maximum of $f(\psi_i)$ (and therefore of $\ln f(\psi_i)$).

Therefore, the Gaussian approximation of $q_{\boldsymbol{\Psi}}(\psi_i)$ is given by

$$q_{\boldsymbol{\Psi}}(\psi_i) \approx \mathcal{N}\left(\psi_i; m_i, \sigma_i^2\right) \quad (9)$$

where

$$\begin{aligned}\langle\psi_i\rangle &= m_i = \tilde{N}_i/\langle\tau\rangle - \mathcal{W}\left(\exp\left(\frac{\tilde{N}_i}{\langle\tau\rangle}\right)\right)/\langle\tau\rangle\\ \sigma_i &= \left(|\langle\tau\rangle + e^z|_{z=m_i}\right)^{-1}\\ \langle\psi_i^2\rangle &= m_i^2 + \sigma_i^2.\end{aligned}$$

2) Distribution of $\boldsymbol{\beta}$

$$\begin{aligned} & q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta};\boldsymbol{\mu},\boldsymbol{\Sigma})\\ \text{with} \quad & \boldsymbol{\Sigma} = \left(\langle\tau\rangle\mathbf{X}^{\mathrm{T}}\mathbf{X} + \langle\tau\rangle\mathrm{diag}\left(\langle\boldsymbol{\alpha}\rangle\right)\right)^{-1}\\ \text{and} \quad & \boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\mathbf{X}^{\mathrm{T}}\langle\boldsymbol{\Psi}\rangle\right)\langle\tau\rangle. \end{aligned} \tag{10}$$

The moments are given by $\langle\boldsymbol{\beta}\rangle = \boldsymbol{\mu}$; $\langle\boldsymbol{\beta}_j^2\rangle = \Sigma_{jj} + \mu_j^2$, and $\langle\boldsymbol{\beta}\boldsymbol{\beta}^T\rangle = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$.

3) Distribution of $\boldsymbol{\alpha}$

$$q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \prod_{j=1}^{p} q_{\boldsymbol{\alpha}}(\alpha_j) = \prod_{j=1}^{p}\mathrm{InvGaussian}(\alpha_j; g_j, h_j) \tag{11}$$

with $g_j = \sqrt{\langle\gamma_j\rangle/\langle\tau\rangle\langle\beta_j^2\rangle}$ and $h_j = \langle\gamma_j\rangle$.

The relevant moments are $\langle\alpha_j\rangle = g_j$; $\langle\alpha_j^{-1}\rangle = g_j^{-1} + h_j^{-1}$.

4) Distribution of $\boldsymbol{\gamma}$

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \prod_{j=1}^{p} q_{\boldsymbol{\gamma}}(\gamma_j) = \prod_{j=1}^{p}\mathrm{Ga}(\gamma_j; a_j, b_j) \tag{12}$$

with $a_j = a_0 + 1$ and $b_j = b_0 + (1/2)\langle\alpha_j^{-1}\rangle$.

Relevant moment is $\langle\gamma_j\rangle = a_j/b_j$.

5) Distribution of τ

$$q_\tau(\tau) = \mathrm{Ga}(\tau; c, d) \tag{13}$$

with $c = c_0 + ((n+p)/2)$ and $d = d_0 + (I/2) + (J/2)$

$$\begin{aligned}\text{where} \quad I &= \sum_{i=1}^{n}\left(\langle\psi_i^2\rangle - 2\langle\psi_i\rangle\mathbf{x}_i^{\mathrm{T}}\langle\boldsymbol{\beta}\rangle + \mathbf{x}_i^{\mathrm{T}}\left\langle\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\rangle\mathbf{x}_i\right)\\ \text{and} \quad J &= \sum_{j=1}^{p}\langle\alpha_j\rangle\left\langle\beta_j^2\right\rangle.\end{aligned}$$

Relevant moment is $\langle\tau\rangle = c/d$.

The parameters of each of the distributions have a dependence on the moments of one or more of the other variables. We can only find an optimum solution via iterative updates starting with the random initial values of the relevant moments since no closed-form solution exists. This is equivalent to applying a gradient ascent to maximize the lower bound J. We can compute the variational lower bound J by replacing $q_\Lambda(\Lambda)$, $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$, $q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$, $q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})$, and $q_\tau(\tau)$ in (7). In each update iteration, we compute the lower bound J and continue the process as long as the bound keeps increasing. Note that, once the approximate solution is reached, we can compute the marginal distributions over coefficients β_i which are Gaussian with mean μ_i and variance Σ_{ii}. We can thereby perform a t-test to determine whether the corresponding covariate influences the extremes frequencies.

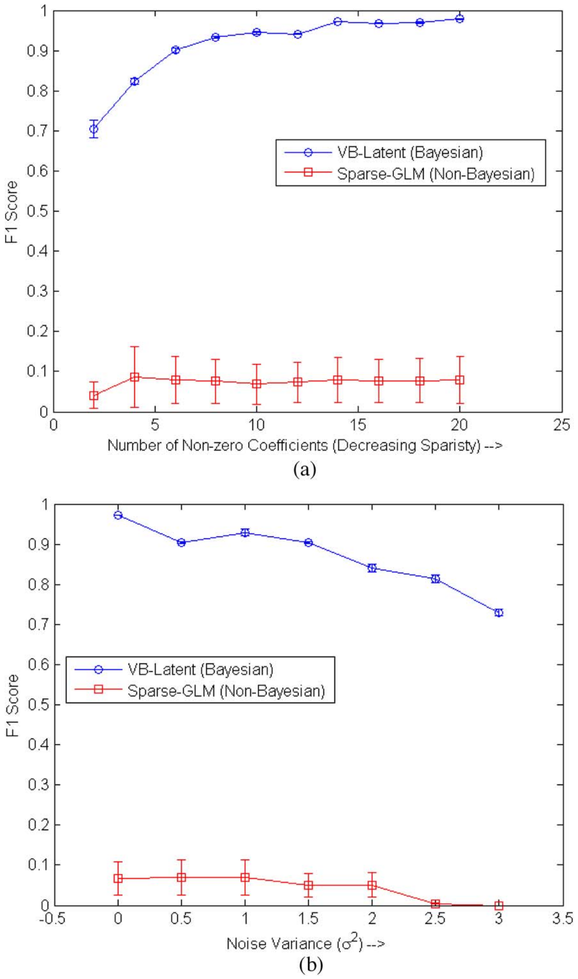
D. Validation on Synthetic Data

In order to validate the performance of our algorithm in terms of how accurately it recovers the true dependence structure, we tested our algorithm on a number of different synthetic data sets with varying complexity. In our first set of experiments, we used ten different $\boldsymbol{\beta}$ values having decreasing sparsity varying from 2 nonzero elements to 20 nonzero elements incremented by 2 at a time, and noise variance was fixed at 0.1. The values of nonzero β_i in these experiments are kept fixed at 0.5. We generated 200 samples of the features $\mathbf{x}_i$ (dimensionality p fixed at 40; $i = 1, \ldots, 200$) from normal distributions $N(0,1)$ and generated λ_i using (2), i.e., by exponentiating the linear combination of features $\mathbf{X}_i$ in the right-hand side. We sampled the frequencies N_i from Poisson distributions having average parameters λ_i. We sampled more than 200 data-points, and in the final data set, we kept only 200 data-points for which N_i is less than 30 to simulate natural scenarios where extremes are rare (the number of rainfall extremes in a year hardly exceeds 30). In the second set of experiments, we fixed the sparsity to ten nonzero coefficients and varied noise variance $\sigma^2 = \tau^{-1}$ [see (2)] from 0 to 3 incremented by 0.5. Experimental data sets are then generated following the same procedure described for the first set of experiments. We repeated both experiments ten times and reported the average F_1-score along with the standard deviation for each set of experiments. F_1-score measures how accurately our method recovers the zero/nonzero structure of the coefficients. It is given by $F_1 = (2PR/(P+R))$, where P is the precision[1] and R is the recall.[2] Since the primary goal of this paper is identifying the covariates correctly, we have evaluated our method based on its ability to correctly identify the relevant covariates and not their ability to predict correctly. As mentioned before, we decided whether each β_i is zero or not based on a t-test that rejects the null hypothesis ($\beta_i \neq 0$) at 1% confidence level.

As a baseline, we reported the F_1-score of non-Bayesian sparse GLM which was implemented using standard software package in MATLAB. The hyperparameter for the non-Bayesian algorithm was tuned using a tenfold cross-validation. The results of both sets of experiments are shown in Fig. 1(a) and (b), respectively. We can see that the Bayesian method is far superior in terms of *accurately* and *precisely* finding the dependence structure. The reason for this large difference in performance between two methods becomes clear when we plot the coefficients estimated by both methods for a single experiment along with the actual coefficients in Fig. 2(a) and (b). Although in both cases initial estimates are highly inaccurate in identifying irrelevant features (0 coefficients), non-Bayesian sparse GLM does not provide any additional information to correct the inaccuracies, while the Bayesian approach provides a distribution over the coefficients. One may argue that a thresholding scheme may be used to convert some of the small coefficients obtained using the non-Bayesian approach to zero. However, the problem with that approach is that there is no

[1] Defined by the ratio of correctly identified zero coefficients (true positives) and total number of estimated zero coefficients (true positives + false positives).

[2] Defined by the ratio of correctly identified zero coefficients (true positives) and total number of actual zero coefficients (true positives + false negatives).

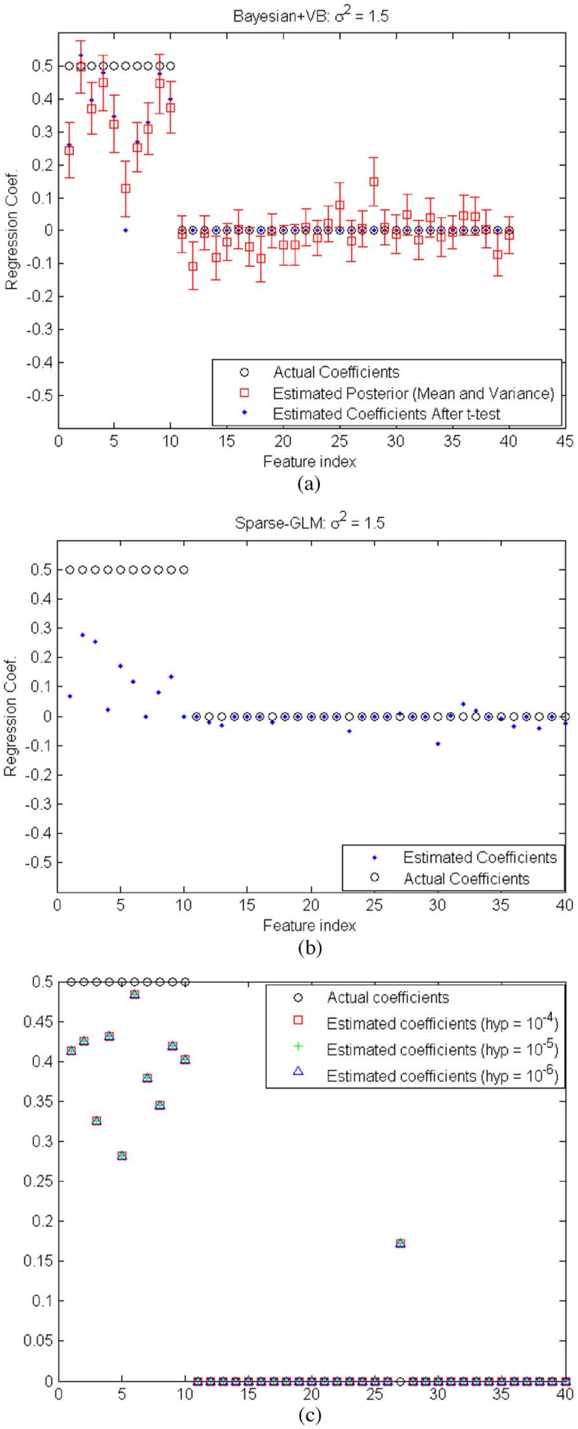


(a)

(b)

Fig. 1. Comparison of accuracy (F_1-score) of Bayesian sparse GLM (our method) with that of non-Bayesian sparse GLM. The mean and standard deviation of F_1-scores from ten repetitions are plotted as functions of (a) number of nonzero coefficients (sparsity) and (b) noise variance σ^2.

statistically sound way of threshold selection that is not problem specific, unlike t-test, which can only be used for a Bayesian approach. A simple t-test at 1% significance level on the results of the Bayesian approach produces almost a perfect result. In Fig. 2(c), we show a sample result (10 nonzero coefficients with a value of 0.5 and $\sigma^2 = 1.5$) with different values of a_0, b_0, c_0, d_0. We can see that the values of these parameters do not affect the estimated values of the coefficients.



(a)

(b)

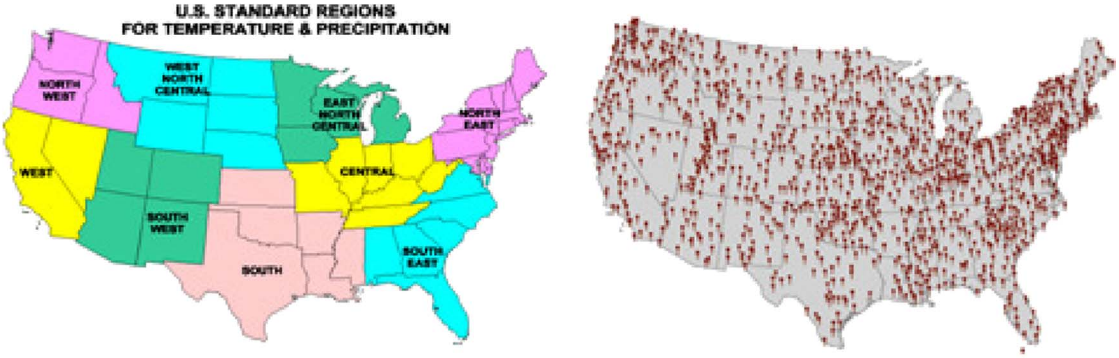
(c)

Fig. 2. Sample results showing the actual coefficients and the estimated coefficients by (a) Bayesian sparse GLM (our method) and (b) non-Bayesian sparse GLM. (c) Sample results showing that changing the values of a_0, b_0, c_0, d_0 have no effect on the estimated values of the coefficients.

E. Data Sources and Preprocessing

We applied our algorithm to estimate the dependence of the number of yearly occurrences of precipitation extremes on a number of potential local, regional, and large-scale climate variables for nine climatologically homogeneous regions over continental U.S. shown in Fig. 3(a). We obtained the daily precipitation and minimum and maximum temperature station data from 1218 stations across continental U.S. stored in the repository of the U.S. Historical Climatology Network (USHCN) [46]. The distribution of stations is shown in Fig. 3(b).

As mentioned earlier, our target variable is the yearly frequency of precipitation extremes over stations within a region. Extremes were defined as events exceeding a threshold set at 99 percentile of all precipitation events at a particular station, and we counted the number of such events in each year. We used a number of covariates for all regions and let our algorithm find the covariates that influence the precipitation extremes frequency. We aggregated all of the data-points for each year and over all of the stations within a given climatologically homogeneous region and obtained one sparse model for each region



(a) (b)

Fig. 3. (a) Homogeneous regions for temperature and precipitation over continental U.S. (b) Distribution of stations for the USHCN.

using our method. For example, if a region has 100 stations and each station has 40 years of data on average, then after aggregation, we have 4000 data-points for that region to build our model. The covariates that we used fall in one of the three broad categories—local, regional, or climate indices (global). Local covariates are included to explain the station-level spatial and temporal variability of the extremes as our model does not handle them explicitly. Regional variables are spatially averaged local variables over the entire region under consideration that are used to represent the regional-level climate dynamics that influences extremes over all of the stations equally. Regional variables do not change spatially over a region but are different for different regions. Both station-level and regional-level climate variables can be further categorized in mean precipitation variables (annual and summer mean) and other atmospheric variables [annual and seasonal mean temperatures, sea-level pressure (SLP), and convective available potential energy (CAPE)]. Additionally, the elevation of the station was used as a local-level variable for which no regional average is used. Finally, the climate indices are global variables that represent large-scale variability in climate variables. Usually, they are principal components (although they are called empirical orthogonal function in climate literature) of the space-time anomaly matrix of global climate variables like sea surface temperature, SLP, etc. Anomalies in climate variables are computed by subtracting the seasonal mean from the daily values. Additionally, we have also used the global mean temperature anomaly (GMT) as a global feature variable. A list of covariates used from each category is given in Table I.

As mentioned earlier, stationwise daily time-series of temperature and precipitation data were obtained from the USHCN repository [46]. SLP and CAPE for each station were collected from the closest grid-point of the gridded North American Regional Reanalysis (NARR) data set [47]. However, the NARR data set is available starting only from 1979. The station elevation was obtained from the high-resolution elevation data set made available by U.S. Geological Survey [48]. Finally, the climate indices were obtained from NOAA's Climate Prediction Center data repository [49], most of which are available from 1950.

We used the covariates from 1979 to 2011 as this is the common period for which all of the covariates are available. If more than 50% of the daily observations out of a year are missing for any of the variables at a specific station, we discarded all variables for that year and for that specific location. For valid years, we took the annual average (or seasonal, wherever applicable) of the daily atmospheric variables. On the other hand, global climate indices are available monthly. We averaged monthly indices over a year if at least two monthly values are available. Otherwise, we discarded the corresponding year. Only two valid monthly values (August and September) were available for Pacific Transition (PT) pattern. For the remaining ten months, the pattern is not a leading mode of variability. On the other hand, for Tropical/Northern Hemisphere pattern, only three monthly values are available every year for the same reason. For all other indices, the maximum number of missing months is 1. For all of the atmospheric variables, both local and regional, we obtained the anomaly time-series by subtracting the temporal mean of the time-series. Finally, we normalized all of the potential covariates before applying our algorithm.

III. Results and Discussion

A. Correlation Among Feature Variables

As mentioned earlier, unregularized regression algorithms do not work in the presence of highly correlated features. Although L_1-regularized sparse regressions do not fail in such cases, they tend to behave inconsistently in terms of selecting correlated feature variables. If there is a group of variables that are highly correlated within the candidate set of variables, non-Bayesian L_1-regularized regressions tend to randomly select one of the variables from this correlated set [26], [28], which may change unpredictably owing to small changes in the choice of the hyperparameter. Bayesian sparse models also tend to pick one variable from a set of highly correlated ones, but it selects the same variable consistently since there is no hyperparameter involved. One alternative approach to handle correlated variables is to first project the data set into a low-dimensional subspace that minimizes the correlation among the projected variables using principal component analysis and then apply the sparse regression method. However, the covariates chosen in this method will not be interpretable as they do not correspond to the actual covariates. A number of other alternative methods like elastic net [28] or fused LASSO [50] use a convex combination of two regularizers to achieve both sparsity

TABLE I
LIST OF CANDIDATE COVARIATES FROM EACH CATEGORY OF VARIABLES USED IN THE STUDY

Category	Variables
Atmospheric variables: station level	Mean Annual Minimum Temperature (Ann-Min-Temp(Stn)), Mean Annual Maximum Temperature (Ann-Max-Temp(Stn)), Mean Winter Minimum Temperature (Win-Min-Temp(Stn)), Mean Winter Maximum Temperature (Win-Max-Temp(Stn)), Mean Spring Minimum Temperature (MAM-Min-Temp(Stn)), Mean Spring Maximum Temperature (MAM-Max-Temp(Stn)), Mean Summer Minimum Temperature (Summ-Min-Temp(Stn)), Mean Summer Maximum Temperature (Summ-Max-Temp(Stn)), Mean Autumn Minimum Temperature (SON-Min-Temp(Stn)), Mean Autumn Maximum Temperature (SON-Max-Temp(Stn)), Sea Level Pressure (SLP(Stn)), Convective Available Potential Energy (CAPE(Stn))
Other variables: station level	Elevation (Elev(Stn))
Mean rainfall: regional average	Mean Annual Precipitation (Ann-Mean-Prcp(Rgn)), Mean Summer Precipitation (Summ-Mean-Prcp(Stn))
Atmospheric variables: regional average	Mean Annual Minimum Temperature (Ann-Min-Temp(Rgn)), Mean Annual Maximum Temperature (Ann-Max-Temp(Rgn)), Mean Winter Minimum Temperature (Win-Min-Temp(Rgn)), Mean Winter Maximum Temperature (Win-Max-Temp(Rgn)), Mean Spring Minimum Temperature (MAM-Min-Temp(Rgn)), Mean Spring Maximum Temperature (MAM-Max-Temp(Rgn)), Mean Summer Minimum Temperature (Summ-Min-Temp(Rgn)), Mean Summer Maximum Temperature (Summ-Max-Temp(Rgn)), Mean Autumn Minimum Temperature (SON-Min-Temp(Rgn)), Mean Autumn Maximum Temperature (SON-Max-Temp(Rgn)), Sea Level Pressure (SLP(Rgn)), Convective Available Potential Energy (CAPE(Rgn))
Climate Indices	North Atlantic Oscillation (NAO), East Atlantic Pattern (EA), West Pacific Pattern (WP), East Pacific/North Pacific Pattern (EPNP), Pacific/North American Pattern (PNA), East Atlantic/West Russia Pattern (EAWR), Scandinavia Pattern (SCA), Tropical/Northern Hemisphere Pattern (TNH), Polar/Eurasia Pattern (POL), Pacific Transition Pattern (PT), Nino 1+2, Nino 3, Nino 3.4, Nino 4, Southern Oscillation Index (SOI), Pacific Decadal Oscillation (PDO), Northern Pacific Oscillation (NP), Tropical/Northern Atlantic Index (TNA), Tropical/Southern Atlantic Index (TSA), Western Hemisphere Warm Pool (WHWP)
Global	Global Mean Temperature Anomaly (GMT)

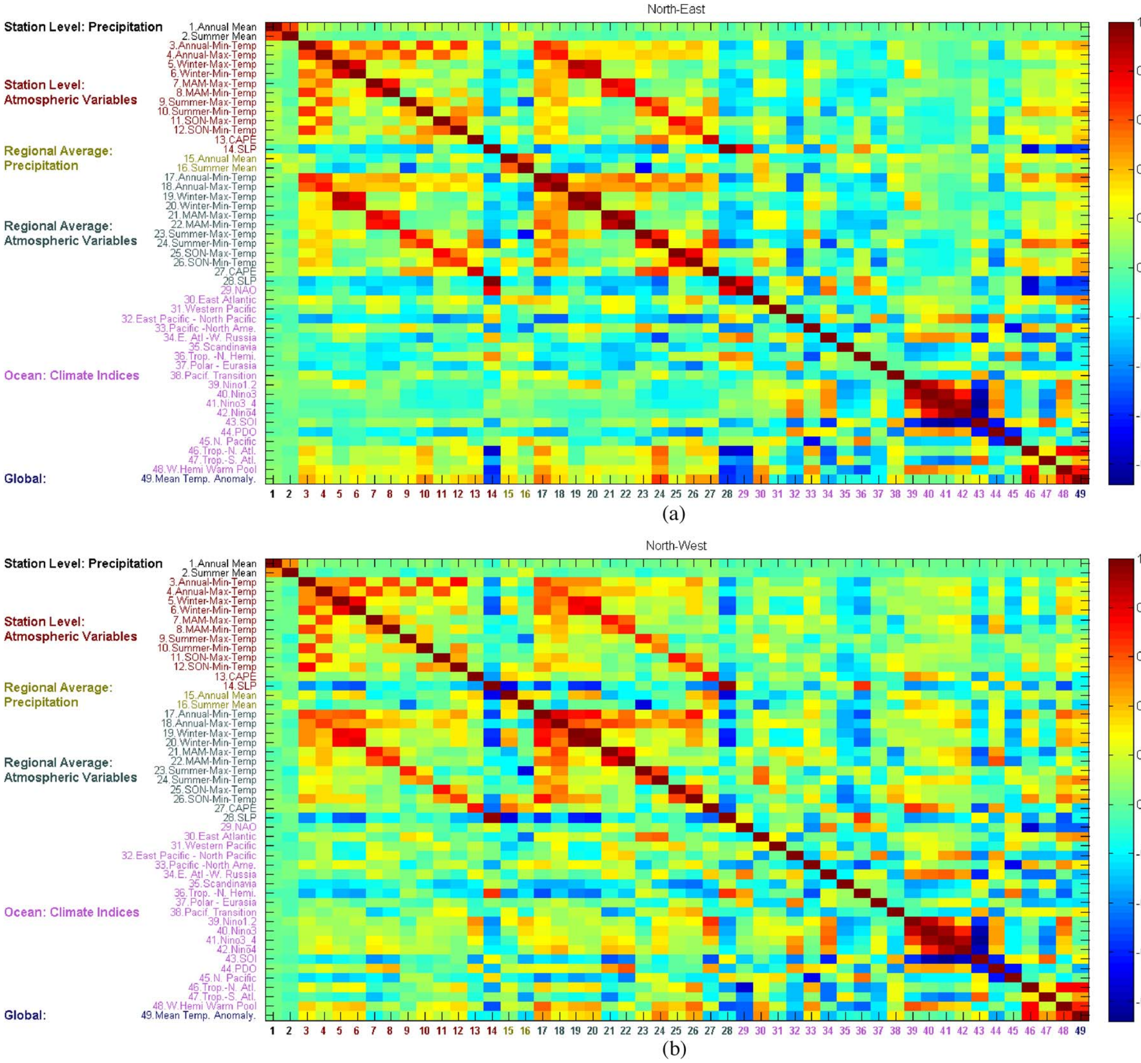
and grouping of correlated variables, thereby adding additional hyperparameters to be tuned. Therefore, before we start to interpret the relevant covariates discovered by our method, it is important to know the correlation structure among the covariates. In Fig. 4(a) and (b), we show the correlation between each pair of candidate covariates for northeast and northwest regions, respectively. The correlation patterns in the other regions look very similar. We can see that highly correlated variables are mostly the station-level and region-level averages of the same variable. This implies that these variables do not exhibit much spatial variability over the particular region. Therefore, if one of these variables is significant, it is hardly important which one is actually selected. The same argument can be used for other pairs (or groups) of variables that are highly correlated. However, at the time of interpretation of the dependence results, it will be important to keep these correlation structures in perspective.

B. Significant Covariates

Now, we show the significant covariates discovered using our method for each of the regions considered in Figs. 5 and 6. In Fig. 5, we display the mean and variance of posterior Gaussian distribution over each of the regression coefficients for the northeast region. However, Fig. 6 shows the set of covariates only which are statistically significant, where the statistical significance is determined using a t-test based on the estimated posterior distribution. Although only a few posterior means are actually zero, the coefficients show different levels of variance which indicates varying levels of uncertainty over the coefficients of different covariates. We used the uncertainty information to perform a t-test that rejects the null hypothesis (H_0) that the coefficient is zero at 1% significance level, which forces many coefficients to be reduced to 0.

C. Spatial Correlation of Extreme Frequencies

In order to show that the spatial variability in the latent average parameters, if any, is explained by the choice of covariates, we show the plot of correlation as a function of distance among time-series of extreme frequencies (N), logarithm of latent average parameters (ψ_i), and residuals ($r_i = \psi_i - \boldsymbol{\beta}^T \mathbf{x}_i$) after fitting the sparse model in Fig. 7(a)–(c), respectively, for the northeast region. These plots can be regarded as the



(a)

(b)

Fig. 4. Correlation matrices between candidate covariates for the (a) northeast region and (b) northwest region.

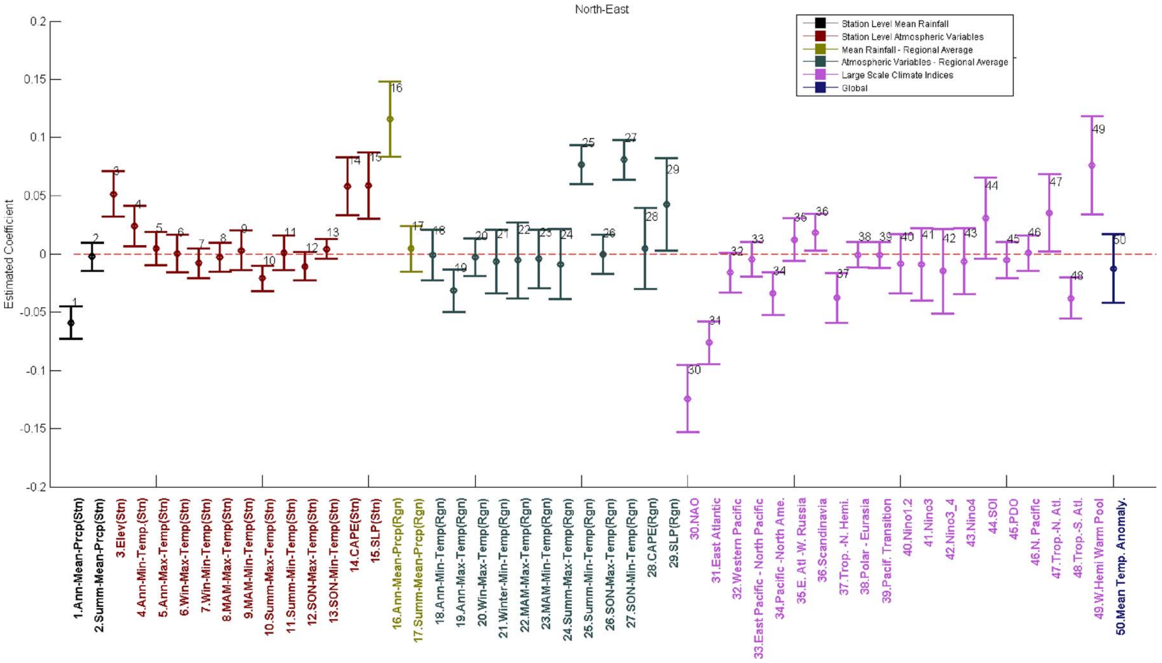


Fig. 5. Mean and variance of the marginal posterior Gaussian distribution over regression coefficients corresponding to each of the covariates for the northeast region.

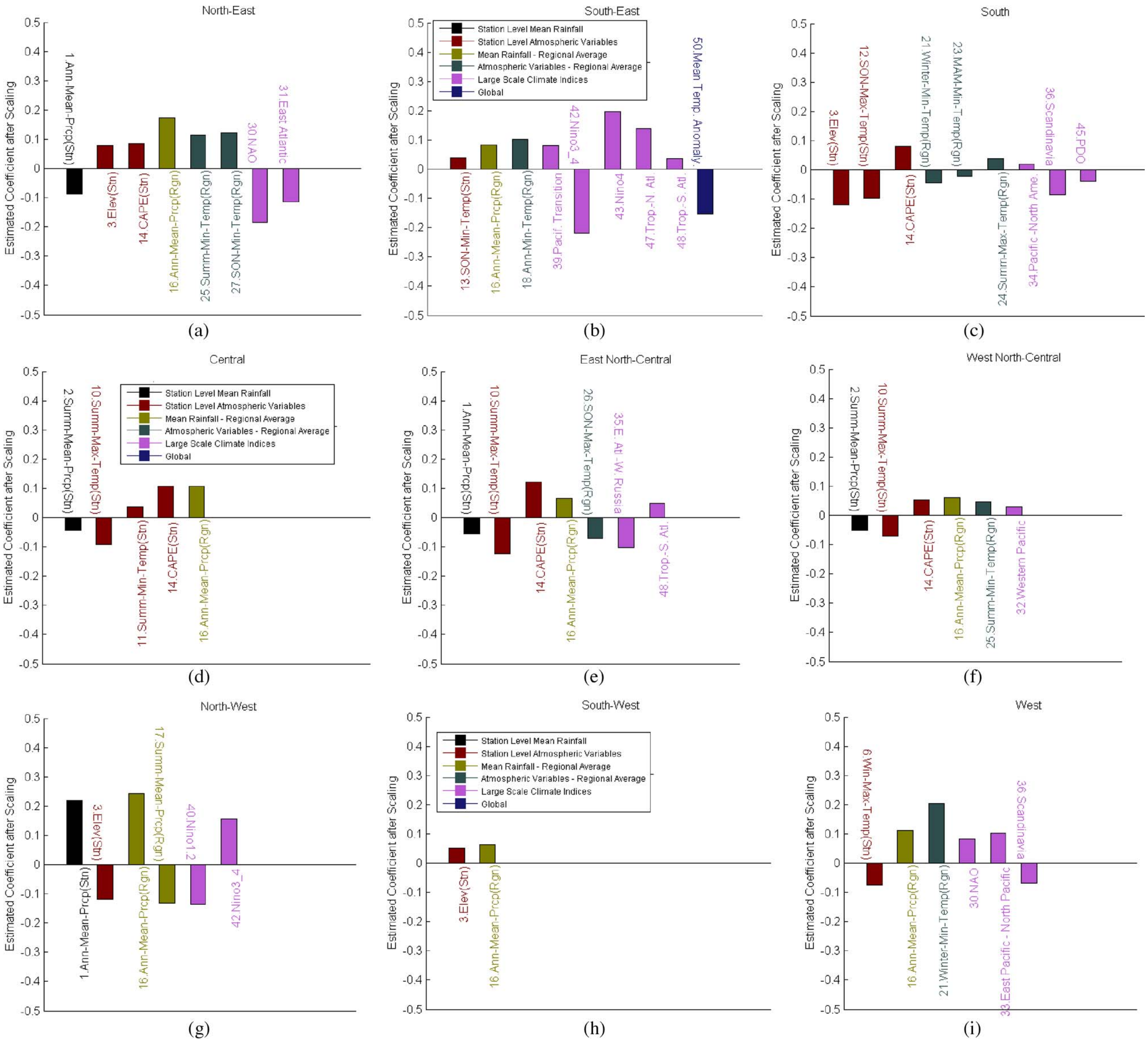


Fig. 6. Nonzero coefficients surviving the t-test for the (a) northeast region, (b) southeast region, (c) southern region, (d) central region, (e) east north-central region, (f) west north-central region, (g) northwest region, (h) southwest region, and (i) western region.

spatial version of *correlograms*, since correlations are plotted as functions of spatial distance instead of the temporal delay. For each of the variables (N, ψ, r), we computed the correlation among all pairs of stations within the target region as well as the distance between the pairs of stations. Then, the distances were sorted, and corresponding correlations were arranged in equidistance bins. We only considered correlations that are statistically significant. The correlations in each bin are then averaged and plotted as a function of equidistance bins.

Although the correlograms shown in Fig. 7 are plotted for the northeast region only, same plots for other regions look very similar. For all regions, the correlations among extreme counts show a clear spatial variability but smaller in value due to the inherent non-Gaussian nature of the extreme frequencies. The correlation increases drastically, although the spatial pattern is preserved, when we use logarithm of average rate parameter (ψ), instead of the actual extremes count. This is expected since ψ is a more stable variable than the actual frequencies and therefore amenable to a linear model. However, spatial correlation almost vanishes, and the spatial pattern all but disappears when residuals (r) are used to generate the same correlation plot. This suggests that the spatial variability observed in the previous two cases is well explained by the chosen features. This also shows that the linear model is sufficient to model the latent average parameters of the extreme frequencies.

D. Insights

A number of interesting observations can be made from the plots of regression coefficients for various regions that suggest dependence of extremes frequencies on corresponding covariates. Most of the regions show existence of predictive information in annual and/or summer mean rainfall values (either station level or regional average) about rainfall extremes,

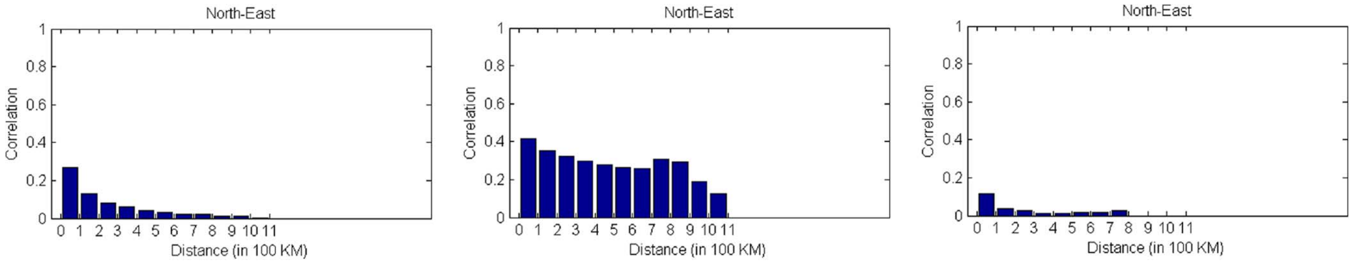


Fig. 7. Spatial correlograms showing the average correlation between stations as a function of spatial distance for (a) extreme counts (N_i), (b) logarithm of latent average parameters (ψ_i) estimated by our algorithm, and (c) residuals ($r_i = \psi_i - \boldsymbol{\beta}\mathbf{x}_i$) after linear fit.

and in the north-west (NW) region, it is particularly large. Note that we excluded the extreme rainfall amounts (rainfall events exceeding the threshold for the station) when computing the mean rainfall at a station. A similar relationship was found using various methods for Washington state [51], Canada [52], Australia [53], and Denmark [23]. The mean annual precipitation (MAP) was found to be a major predictor of extreme percentiles of rainfall intensity by [54] over a large region encompassing U.S., Europe and parts of Asia, northern Africa, and south America. However, in our study, no such predictive value of MAP for extreme frequencies was found in the southern U.S. Moreover, in the southeastern, southwestern, and western U.S., the regional average of MAP is found to be predictive of extreme frequencies rather than the station-level MAP. These results suggest that the regional-scale analysis of rainfall may lead to insight that is not in agreement with the global insight, and additional studies and methods may be required to find regions around the globe where mean rainfall has predictive information about extremes.

In the NW region, the regression coefficients of annual and summer mean rainfalls have opposite signs. Although it might seem counterintuitive at a first glance, a closer look at the rainfall pattern of the region reveals that summer is drier than winter in NW and summer average rainfall does not show any significant correlation with annual average in that region [see Fig. 4(b)]. Therefore, based on this result, one can claim that, if this region gets a lower average rainfall in the summer, there is a high probability that it might get more extremes in the winter on average, which, to our knowledge, is a new insight.

Influence of temperature on both convective and large-scale precipitation extremes has been documented in literature. Theoretically, Clausius–Clapeyron relationship dictates that the maximum amount of water vapor contained in an air-column is determined by the temperature of the column. Increasing global temperature is found to have influence on increasing extreme rainfalls at a global level over a multidecadal time-scale [55], [56], although the relation is not straightforward at a regional level and over a smaller time-scale [57]. An indicator for regional-level convective process, which leads to convective rainfall, is CAPE that quantifies convective pull in an air-column due to the temperature difference in the vertical layers of the column. Dependence of rainfall extreme frequencies on the CAPE therefore implies prevalence of convective rainfall in that region. Our results show that the northeast, south, central, east north-central, and west north-central regions have nonzero coefficients for CAPE, which implies that a sizeable number of extremes in these regions possibly occur due to the regional convective processes. Moreover, in the central region, there is no influence of climate indices, which may mean that majority of extreme rainfalls in this region are caused by convection. On the other hand, southeast or any of the regions in western U.S. show dependence, which may imply that regional convection contributes only a small number of extremes in these regions.

The effect of elevation on the occurrence of rainfall extremes has been well documented for the northwest region. However, we found a dependence of rainfall extremes on elevation in the northeast and southern regions as well. We did not find dependence on SLP in any of the regions.

It can also be seen that most of the regions show dependence of rainfall extremes on one or more large-scale climate indices except land-locked regions like the central and southwest regions. Since most of the climate indices explain large-scale variability of the ocean variables, dependence on these variables also represents ocean influence on rainfall extremes. As mentioned earlier, similar relations between extreme rainfall and large-scale indices have been studied for contiguous U.S. [17], [20] and at more regional level [19], [58] using both observed and model simulations. We can see that the southeast region shows particularly a strong dependence on multiple ocean variables. A plausible explanation can be that the majority of the rainfall extremes in this region are caused by tropical cyclones of various intensities [59], [60], which, in turn, are generated over oceans due to the variability in ocean variables. Past studies have established a link between precipitation patterns in Florida with El Nino, PDO, and NAO patterns [17], [61], [62]; our method have found links between precipitation extremes frequency in the southeast region with Tropical North and South Atlantic (TNA and TSA) patterns, Nino indices, and PT patterns.

Southeast is the only region that shows dependence on global mean temperature anomaly, albeit the dependence being negative. This suggests reduction of the extreme rainfall events in this region with global warming, which contradicts the existing knowledge about the relation between the two. Although the trend in rainfall extreme frequencies in the southeastern U.S. has not been previously linked to global temperature anomaly before, weakly declining trend of heavy rainfall has been reported along the coastal section of the region, which has been linked to the fluctuation of the high-pressure region in the subtropical Atlantic (Bermuda high) [63]. This is an example of an interesting insight that can be regarded as a new hypothesis and needs to be validated using a hypothesis-guided approach that involves understanding the physics behind it.

E. Discussions

Since we used a Bayesian approach to estimate the posterior distribution over the regression coefficients, we get a natural estimate of uncertainty bound around the estimated mean coefficient which enables us to test the significance of the coefficient in a statistically sound manner. The estimated uncertainty bounds for the northeast region are shown in Fig. 5, and only the significant coefficients determined from these bounds are shown in Fig. 6. It can be seen that the southwest region has only two significant covariates, which is significantly smaller than the other regions. One possible reason is that this region does not get too much rainfall, and therefore, many years will experience no extremes at all. Therefore, many data-points were not used to build the regression model. This resulted in large uncertainty in posterior distribution of the regression coefficients which were then subsequently proved insignificant.

As mentioned earlier, the correlation among the potential covariates needs to be considered before inferring any dependence relationship. As an example, the influence on Nino indices has to be interpreted carefully since some of them highly correlated among themselves and therefore contains redundant information. In the north-west region, two Nino indices show opposite effects on precipitation extremes. However, these two covariates are highly correlated, and therefore, it is difficult to infer from this result which variable has actual influence on the rainfall extremes. In this case, their significance needs to be tested individually using further statistical tests. We can ensure selection of all correlated variables if at least one of them is actually important by using a Gaussian prior on the coefficients to encourage grouping in addition to the sparsity-inducing Laplace prior currently being used. However, that will introduce a tunable hyperparameter (to balance between weights of two priors) and may cause inconsistent results due to the small variation of the hyperparameter. Moreover, we will still have to decide the covariates that actually influence the rainfall extremes using further statistical tests. Improving our Bayesian model to automatically handle the correlated features remains to be a future work.

IV. Conclusion

In this paper, we have presented a statistically rigorous framework for finding covariates that influence the extreme frequencies. We have extended the Bayesian hierarchical description of sparse regression to accommodate GLM with Poisson-distributed target variable and developed an efficient VB inference algorithm to estimate the probability distribution over the regression parameters. Our method shows a high level of accuracy on synthetic data set and outperforms the non-Bayesian alternative by a large margin. We have also shown the value of using our method for finding the covariates that influence precipitation extremes frequency out of a large pool of candidate covariates of different spatial scale.

The insights that we have obtained using our methods can be summarized in three categories from the perspective of climate science. First, we have rediscovered some known relations like the following: 1) predictive value of station-level and/or regional-level mean annual and summer rainfalls about the rainfall extreme frequencies and 2) dependence of extreme rainfall frequencies in the northwest region on elevation. Second, we have obtained new insights that are fairly intuitive in nature like the following: 1) dependence of rainfall extremes in the central region on CAPE but not on climate indices that suggest that majority of rainfall extremes are convective in that region, whereas in the southeastern and western U.S., our method suggests otherwise, and 2) rainfall extremes frequency depends on elevation in the northeast. Finally, we have obtained a few insights that are new and counterintuitive and therefore need more focused statistical and physical validation. The inverse relations between rainfall extremes frequency with summer mean rainfall in the northwest U.S. and with global mean temperature anomaly in the southeast U.S. are examples of such insights. We have shown the importance of considering covariates at different spatial scales in order to fully capture the multiscale processes that lead to rainfall extremes. Moreover, we have demonstrated that our method can be used as an alternative to hypothesis-guided approaches for discovering new insights related to precipitation extremes. However, one important caveat is that these new insights have to be ultimately corroborated by the climate physics.

The dependence analysis performed in this paper is mainly exploratory in nature. We plan to extend the model to perform predictive analysis in the future. We also plan to develop a similar sparse model for dependence analysis of extremes intensity which is significantly more challenging since it involves distributions outside of the exponential family. Finally, we plan to incorporate a spatiotemporal autoregressive model within our probabilistic framework to leverage the spatiotemporal correlation inherent in climate and other spatiotemporal data sets.

References

[1] R.-D. Reiss and M. Thomas, *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*, 3rd ed. New York, NY, USA: Springer-Verlag, 2007.

[2] M. Bahadori and Y. Liu, "Granger causality analysis in irregular time series," in *Proc. SIAM Int. Conf. Data Mining*, 2012, pp. 660–671.

[3] C. B. Field, V. Barros, T. F. Stocker, and Q. Dahe, Eds., *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[4] M. Brunetti, M. Maugeri, and T. Nanni, "Changes in total precipitation, rainy days and extreme events in northeastern Italy," *Int. J. Climatol.*, vol. 21, no. 7, pp. 861–871, Jun. 2001.

[5] V. Ntegeka and P. Willems, "Trends and multidecadal oscillations in rainfall extremes, based on a more than 100-year time series of 10 min rainfall intensities at Uccle, Belgium," *Water Resources Res.*, vol. 44, no. 7, pp. 1–15, 2008.

[6] H. Madsen, K. Arnbjerg-Nielsen, and P. Steen, "Update of regional intensity–duration–frequency curves in Denmark: Tendency towards increased storm intensities," *Atmos. Res.*, vol. 92, no. 3, pp. 343–349, 2009.

[7] B. Su, B. Xiao, D. Zhu, and T. Jiang, "Trends in frequency of precipitation extremes in the Yangtze River basin, China: 1960–2003," *Hydrol. Sci.*, vol. 50, no. 3, pp. 479–492, 2005.

[8] S. Shahid, "Trends in extreme rainfall events of Bangladesh," *Theor. Appl. Climatol.*, vol. 104, no. 3/4, pp. 489–499, Jun. 2011.

[9] J.-Y. Tu and C. Chou, "Changes in precipitation frequency and intensity in the vicinity of Taiwan: Typhoon versus non-typhoon events," *Environ. Res. Lett.*, vol. 8, no. 1, Mar. 2013, Art. ID. 014023.

[10] S. K. Aryal *et al.*, "Characterizing and modeling temporal and spatial trends in rainfall extremes," *J. Hydrometeorol.*, vol. 10, no. 1 pp. 241–253, Feb. 2009.

[11] R. Suppiah and K. J. Hennessy, "Trends in total rainfall, heavy rain events and number of dry days in Australia, 1910–1990," *Int. J. Climatol.*, vol. 18, no. 10, pp. 1141–1164, Aug. 1998.

[12] A. T. DeGaetano, "Time-dependent changes in extreme-precipitation return-period amounts in the continental United States," *J. Appl. Meteorol. Climatol.*, vol. 48, no. 10, pp. 2086–2099, Oct. 2009.

[13] G. Villarini *et al.*, "On the frequency of heavy rainfall for the Midwest of the United States," *J. Hydrol.*, vol. 400, no. 1/2, pp. 103–120, Mar. 2011.

[14] A. A. Bradley, "Regional frequency analysis methods for evaluating changes in hydrologic extremes," *Water Resources Res.*, vol. 34, no. 4, pp. 741–750, Apr. 1998.

[15] A. A. Scaife, C. K. Folland, L. V. Alexander, A. Moberg, and J. R. Knight, "European climate extremes and the North Atlantic Oscillation," *J. Clim.*, vol. 21, no. 1, pp. 72–83, Jan. 2008.

[16] M. R. Haylock and C. M. Goodess, "Interannual variability of European extreme winter rainfall and links with mean large-scale circulation," *Int. J. Climatol.*, vol. 24, no. 6, pp. 759–776, May 2004.

[17] D. B. Enfield, A. M. Mestas-Nunez, and P. J. Trimble, "The Atlantic Multidecadal Oscillation and its relation to rainfall and river flows in the continental U.S.," *Geophys. Res. Lett.*, vol. 28, no. 10, pp. 2077–2080, May 2001.

[18] S. Curtis, "The Atlantic Multidecadal Oscillation and extreme daily precipitation over the US and Mexico during the hurricane season," *Clim. Dyn.*, vol. 30, no. 4, pp. 343–351, Mar. 2008.

[19] G. Goodrich, "Influence of the Pacific Decadal Oscillation on winter precipitation and drought during years of neutral ENSO in the Western United States," *Weather Forcast.*, vol. 22, no. 1, pp. 116–124, Feb. 2007.

[20] A. Gershunov and T. Barnett, "ENSO influence on intraseasonal extreme rainfall and temperature frequencies in the contiguous United States: Observations and model results," *J. Clim.*, vol. 11, no. 7, pp. 1575–1586, Jul. 1998.

[21] M. Hatzaki, H. A. Flocas, C. Oikonomou, and C. Giannakopoulos, "Future changes in the relationship of precipitation intensity in Eastern Mediterranean with large scale circulation," *Adv. Geosci.*, vol. 23, pp. 31–36, 2010.

[22] J. R. M. Hosking and J. R. Wallis, *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[23] H. Madsen, P. S. Mikkelsen, D. Rosbjerg, and P. Harremoe, "Regional estimation of rainfall intensity–duration–frequency curves using generalized least squares regression of partial duration series statistics," *Water Resources Res.*, vol. 38, no. 11, pp. 21-1–21-11, Nov. 2002.

[24] I. B. Gregersen, H. Madsen, D. Rosbjerg, and K. Arnbjerg-Nielsen, "A spatial and non-stationary model for the frequency of extreme rainfall events," *Water Resources Res.*, vol. 49, no. 1, pp. 127–136, Jan. 2013.

[25] A. Y. Ng, "Feature selection, L1 versus L2 regularization, and rotational invariance," in *Proc. Int. Conf. Mach. Learn.*, 2004, p. 78.

[26] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Stat. Soc., Ser. B (Statist. Methodology)*, vol. 58, no. 1, pp. 267–288, 1996.

[27] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[28] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc., Ser. B (Statist. Methodology)*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[29] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, Jan. 2010.

[30] S. Ryali, K. Supekar, D. A. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fMRI data" *NeuroImage*, vol. 51, no. 2, pp. 752–764, Jun. 2010.

[31] D. Lee, Y. Lee, Y. Pawitan, and W. Lee, "Sparse partial least-squares regression for high-throughput survival data analysis," *Stat. Med.*, vol. 32, no. 30, pp. 5340–5352, Dec. 2013.

[32] Y. Lu, Y. Zhou, W. Qu, M. Deng, and C. Zhang, "A LASSO regression model for the construction of microRNA-target regulatory networks," *Bioinformatics*, vol. 27, no. 17, pp. 2406–2413, Sep. 2011.

[33] S. Kaneko, A. Hirakawa, and C. Hamada, "Gene selection using a high-dimensional regression model with microarrays in cancer prognostic studies" *Cancer Inform.*, vol. 11, pp. 29–39, Jan. 2012.

[34] X. Lv, G. Bi, and C. Wan, "The group LASSO for stable recovery of block-sparse signal representation," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1371–1382, Apr. 2011.

[35] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[36] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Collaborative sparse regression for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 341–354, Jan. 2014.

[37] M. Seeger and H. Nickisch, "Large scale Bayesian inference and experimental design for sparse linear models," *SIAM J. Imag. Sci.*, vol. 4, no. 1, pp. 166–199, 2011.

[38] T. Park and G. Casella, "The Bayesian LASSO," *J. Amer. Stat. Assoc.*, vol. 103, no. 482, pp. 681–686, Jun. 2008.

[39] A. Kabán, "On Bayesian classification with Laplace priors," *Pattern Recog. Lett.*, vol. 28, no. 10, pp. 1271–1282, Jul. 2007.

[40] M. Seeger, S. Gerwinn, and M. Bethge, "Bayesian inference for sparse generalized linear models," *Mach. Learn., ECML*, vol. 4701, pp. 298–309, 2007.

[41] S. Cole, *An Introduction to Statistical Modeling of Extreme Values*. New York, NY, USA: Springer-Verlag, 2001.

[42] A. C. Lozano *et al.*, "Spatial–temporal causal modeling for climate change attribution," in *Proc. 15th ACM SIGKDD Int. Conf. KDD*, 2009, pp. 587–596.

[43] P. McCullach and J. A. Nelder, "Generalized linear models," in *Monographs on Statistics and Applied Probability*, 1st ed. London, U.K.: Chapman & Hall, 1983.

[44] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, Nov. 2008.

[45] G. Parisi, *Statistical Field Theory*. New York, NY, USA: Addison-Wesley, 1988.

[46] D. R. Easterling, T. R. Karl, J. H. Lawrimore, and S. A. D. Greco, "United States Historical Climatology Network Daily Temperature, Precipitation, and Snow Data for 1871–1997," Carbon Dioxide Inf. Anal. Center, Oak Ridge Natl. Lab., U.S. Dept. Energy, Oak Ridge, TN, USA, Tech. Rep., May 1999.

[47] F. Mesinger *et al.*, "North American regional reanalysis," *Bullet. Amer. Meteorol. Soc.*, vol. 87, no. 3, pp. 343–360, Mar. 2006.

[48] D. Gesch, G. Evans, J. Mauck, J. Hutchinson, and W. Carswell, Jr., "The National Map—Elevation: U.S. Geological Survey Fact Sheet," U.S. Geological Survey, Reston, VA, USA, Tech. Rep., Jul. 2009.

[49] "Climate indices: Monthly atmospheric and ocean time series." [Online]. Available: http://www.esrl.noaa.gov/psd/data/climateindices/list/

[50] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused LASSO," *J. R. Stat. Soc., Ser. B (Statist. Methodology)*, vol. 67, no. 1, pp. 91–108, Feb. 2005.

[51] J. R. Wallis, M. G. Schaefer, B. L. Barker, and G. H. Taylor, "Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Washington state," *Hydrol. Earth Syst. Sci.*, vol. 11, no. 1, pp. 415–442, 2007.

[52] Y. Alila, "A hierarchical approach for the regionalization of precipitation annual maxima in Canada," *J. Geophys. Res.*, vol. 104, no. D24, pp. 31 645–31 655, Dec. 1999.

[53] K. Haddad, A. Rahman, and J. Green, "Design rainfall estimation in Australia: A case study using L moments and generalized least squares regression," *Stochastic Environ. Res. Risk Assess.*, vol. 25, no. 6, pp. 815–825, 2011.

[54] R. E. Benestad, D. Nychka, and L. O. Mearns, "Spatially and temporally consistent prediction of heavy precipitation from mean values—Supporting material," *Nature Clim. Change*, vol. 2, pp. 544–547, 2012.

[55] 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. United States Amer.*, vol. 106, no. 35, pp. 14 773–14 777, 2009.

[56] P. A. O'Gorman, "Sensitivity of tropical precipitation extremes to climate change," *Nature Geosci.*, vol. 5, no. 10, pp. 697–700, Sep. 2012.

[57] V. Mishra, J. M. Wallace, and D. P. Lettenmaier, "Relationship between hourly extreme precipitation and local air temperature in the United States," *Geophys. Res. Lett.*, vol. 39, no. 16, Aug. 2012, Art ID. L16403.

[58] M. J. DeFlorio, D. W. Pierce, D. R. Cayan, and A. J. Miller, "Western U.S. extreme precipitation events and their relation to ENSO and PDO in CCSM4," *J. Clim.*, vol. 26, no. 12, pp. 4231–4243, Jun. 2013.

[59] J. M. Shepherd, A. Grundstein, and T. L. Mote, "Quantifying the contribution of tropical cyclones to extreme rainfall along the coastal southeastern United States," *Geophys. Res. Lett.*, vol. 34, no. 23, Dec. 2007, Art ID. L23810.

[60] D. B. Knight and R. E. Davis, "Contribution of tropical cyclones to extreme rainfall events in the southeastern United States," *J. Geophys. Res.*, vol. 114, no. D23, Dec. 2009, Art ID. D23102.

[61] R. S. Teegavarapu, *Floods in a Changing Climate: Extreme Precipitation*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[62] J. Obeysekera *et al.*, "Consideration of Long-Term Climatic Variability in Regional Modeling for SFWMD Planning & Operations," South Florida Water Manage. District, West Palm Beach, FL, USA, Tech. Rep., 2007.

[63] B. D. Keim, "Preliminary analysis of the temporal patterns of heavy rainfall across the southeastern United States," *Prof. Geographer*, vol. 49, no. 1, pp. 94–104, Feb. 1997.

Debasish Das received the Master's degree in electrical engineering and the Doctorate degree in computer science from Temple University, Philadelphia, PA, USA.

He spent the last few years of his doctoral studies working as a Research Associate in the Department of Civil and Environmental Engineering at Northeastern University, Boston, MA, USA, where he worked on data-driven methods to understand climate change. He is currently a Data Scientist with Verizon Big Data Analytics Group, Verizon Communications, New York, NY, USA. His research interests include Bayesian machine learning, nonparametric models, and their applications.

Auroop R. Ganguly received the Ph.D. degree in civil and environmental engineering from Massachusetts Institute of Technology, Cambridge, MA, USA, with a specialization in hydrology.

He was with the Oak Ridge National Laboratory for seven years and with Oracle Corporation, including a startup subsequently acquired by Oracle, for about six years, where his final positions were Senior Scientist and Senior Product Manager, respectively. He is currently an Associate Professor of civil and environmental engineering and the Director of the Sustainability and Data Sciences Laboratory with Northeastern University, Boston, MA. He has published in journals such as Nature, Nature Climate Change, Proceedings of the National Academy of Sciences, Scientific Reports, Physical Review E, and IEEE transactions and published edited book *Knowledge Discovery from Sensor Data*. His interests include ancient history, mythology, comparative religion, and science fiction.

Dr. Ganguly served as a Guest Editor for special issues of Nonlinear Processes in Geophysics on physics-guided data mining in climate extremes and Intelligent Data Analysis on data stream mining and currently serves as an Associate Editor of the journals Water Resources Research and Journal of Computing in Civil Engineering. He was a recipient of best paper awards at data mining conferences such as the SIAM International Conference on Data Mining.

Zoran Obradovic He is currently an L.H. Carnell Professor of data analytics with Temple University, Philadelphia, PA, USA, where he is also a Professor in the Department of Computer and Information Sciences with a secondary appointment in Statistics and is the Director of the Center for Data Analytics and Biomedical Informatics. His work is published in over 300 articles and is cited over 14 700 times (H-index 47). His research interests include data mining, complex networks applications, and decision support systems.

Prof. Obradovic was a General Cochair for the 2013 and 2014 SIAM International Conference on Data Mining and has been the program or track chair at many data mining conferences. In 2014-2015, he chairs the SIAM Activity Group on Data Mining and Analytics. He is the Executive Editor of the Statistical Analysis and Data Mining journal, which is the official publication of the American Statistical Association, and is an Editorial Board Member of eleven journals.