EVA for large spatial data sets

Emeric Thibaud and Brian Reich

Colorado State University and North Carolina State University

June 17, 2016

Joint work with Sam Morris (NCSU) and Dan Cooley (CSU)

I reached my 5-year return level!

Thanks to the organizers. The meeting has be great.

Sam Morris

Spatial extremes

- EVA can benefit greatly from spatial methods
- Spatial methods can map risk and borrow strength over space to estimate rare-event probabilities
- Accounting for spatial dependence is necessary for valid inference
- Methods and software in this area are developing rapidly to meet a growing demand

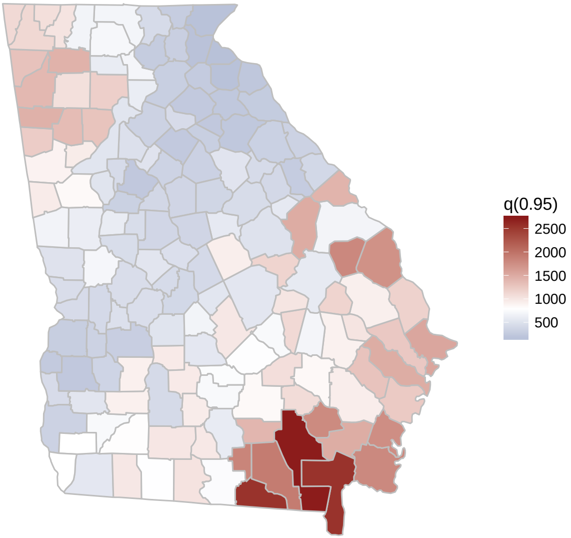
Current approaches and limitations

- Theory suggests that a max-stable process is a good option for spatial extremes
- The max-stable process gives a complicated likelihood function with no closed-form except in trivial cases
- Current Bayesian approaches can handle only a moderate number of spatial locations
- This is limiting because most modern applications have hundreds or thousands of stations
- Because of these challenges advanced methods for e.g., multivariate or nonstationary data are limited

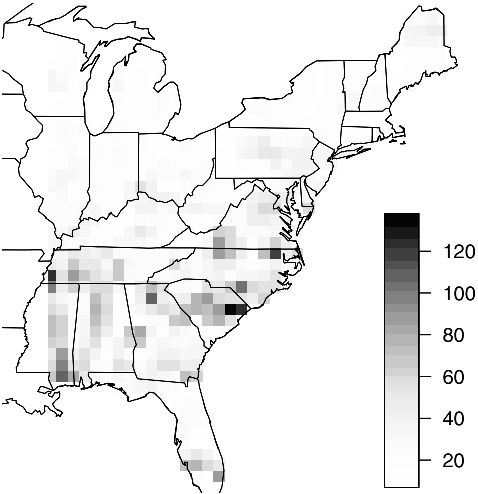
Outline of talk

- The objective of this work is to develop Bayesian methods that can be scaled up to high dimensions
- Approach 1: Low-rank empirical basis approximation
- Approach 2: Spatial skew-t process
- These two approaches are applied to both block maxima and peaks over a threshold

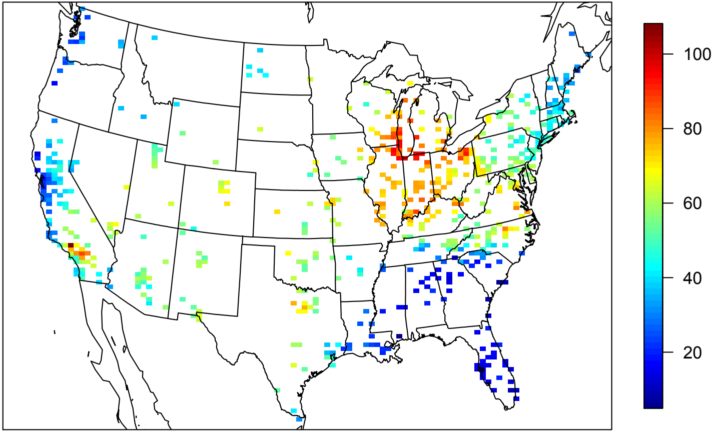
Application 1 – Forest fires in GA



Application 2 – Climate model precip output



Application 3 – Air pollution (ozone)



Low-rank methods for Gaussian data

- Principle components analysis (aka., empirically orthogonal functions) is a valuable tool for climate data
- PCA uses the sample covariance matrix between spatial locations to identify highly-correlated sites
- This is useful for understanding spatial patterns and identifying homogeneous sub-regions
- It is also a dimension-reduction tool; the times series of the PCA loadings are approximately independent

Low-rank methods for Gaussian data

- Assume there are n spatial locations and $\text{Y}_1, ..., \text{Y}_m$ are the data for m independent replications
- PCA decomposes the $n \times n$ sample covariance as $S = BVB^T$
- Eigenvector maps (B's columns) reveal the large-scale spatial patterns
- Dimension reduction: often $L << n$ is sufficient
- Denote the first L columns of B as B_L
- Replication t's loadings are $A_t = B_L^T \text{Y}_t$
- The $L << n$ elements of the A_t should be independent

Low-rank methods for Gaussian data

- This idea extends from a discrete process at n locations to a Gaussian process (GP) $Y_t(s)$
- Karhunen-Loeve: Any GP $Y_t(s)$ can be written

$$Y_t(\mathrm{s}) = \sum_{l=1}^{\infty} B_l(s) A_{lt}$$

- $B_l(\mathrm{s})$ are orthonormal spatial eigenfunctions
- A_l are independent normals with variance v_l
- Covariance: $\mathrm{Cov}[Y_t(\mathrm{s}), Y_t(\mathrm{s}')] = \sum_l B_l(\mathrm{s}) B_l(\mathrm{s}') v_l$

Challenges in EVA

- Covariance decompositions cannot be applied for extremes because the covariance may not exist
- Also, covariance focuses on deviations around the mean and not the extremes
- In this talk we propose a method to identify empirical basis functions (EBF) for extremes
- We use the EBFs for both exploratory analysis and model building

Max-stable processes

- Let $Y_1(s)$, ..., $Y_m(s)$ be iid spatial processes
- The pointwise maximum process is

$$\tilde{Y}(s) = \bigvee_{l=1}^{m} Y_l(s)$$

- If there exist constants a_L and b_L so that

$$Z(s) = a_m + b_m \tilde{Y}(s)$$

converges to a valid process as $m \to \infty$, then Z is max-stable

- The marginal distribution of Z at each s is GEV

Spectral representation theorem

- Any max-stable process can be written as a pointwise maximum of m processes (de Haan)
- Max-linear model (Wang and Stoev):

$$Z(s) = \bigvee_{l=1}^{L} B_l(s) A_l$$

where $B_l(s) > 0$, $\sum_l B_l(s) = 1$ for all s, and $A_l \stackrel{iid}{\sim}$ GEV

- If we view the $B_l(s)$ as basis functions constant over time, these can play the role of PCs/eigen-functions
- The A_l change over time and play the role of the loadings

Low-rank positive-stable representation

- It is unlikely that realizations will identically equal the point-wise maximum of L processes
- Following Reich and Shaby (2012), let $Z_t(\mathrm{s})$, the value at site s and time t, be

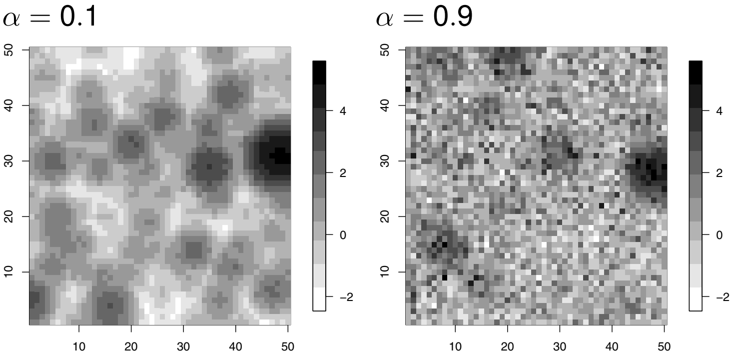
$$Z_t(\mathrm{s}) = \theta_t(\mathrm{s})\varepsilon_t(\mathrm{s})$$

  where θ_t is a spatial process and $\varepsilon_t(\mathrm{s}) \overset{iid}{\sim} \mathrm{GEV}(1,\alpha,\alpha)$
- The spatial process is

$$\theta_t(\mathrm{s}) = \left(\sum_{l=1}^{L} B_l(\mathrm{s})^{1/\alpha} A_{tl}\right)^{\alpha}$$

  where $A_{tl} \sim \mathrm{PS}(\alpha)$

The parameter $\alpha \in (0, 1)$ controls the "nugget"



Low-rank positive-stable representation

- Z_t is max-stable marginally over the random effects A_{tl}
- The joint is GEV - asymmetric Laplace
- Dependence is measured by the extremal coefficient ϑ, defined via

$$\text{Prob}[Z_t(\mathbf{s}_1) < c, Z_t(\mathbf{s}_2) < c] = \text{Prob}[Z_t(\mathbf{s}_1) < c]^{\vartheta(\mathbf{s}_1,\mathbf{s}_2)}$$

- For the low-rank PS model

$$\vartheta(\mathbf{s}_1, \mathbf{s}_2) = \sum_{l=1}^{L} \left[B_l(\mathbf{s}_1)^{1/\alpha} + B_l(\mathbf{s}_2)^{1/\alpha} \right]^{\alpha} \in [1, 2]$$

Estimating the EBFs, $B_l(\mathrm{s})$

1. Use a rank transformation to standardize data for each s
2. Estimate the extremal dependence between each pair of sites (using χ or madogram), $\hat{\vartheta}(\mathrm{s}_i, \mathrm{s}_j)$
3. Spatially (4D) smooth the sample dependence measures
4. Constrained least squares (next slide) to minimize the distance between sample ($\hat{\vartheta}$) and model (ϑ as a function of the B) spatial dependence
5. Order the terms by $v_l = \sum_{\mathbf{s}} B_l(\mathrm{s})$

Estimating the EBFs, $B_l(\mathrm{s})$

- The objective function to estimate the B_l is

$$\sum_{i<j}\left[\hat{\vartheta}(\mathrm{s}_i,\mathrm{s}_j)-\vartheta(\mathrm{s}_i,\mathrm{s}_j)\right]^2$$

  where $\vartheta(\mathrm{s}_i,\mathrm{s}_j)$ is a function of B_l

- The EBFs must satisfy $B_l(\mathrm{s})>0$ and $\sum_l B_l(\mathrm{s})=1$ for all s

- The solution is approximated by cycling through the sites and solving a series of constrained optimization problems

Contrasts with PCA

- Basis functions are not orthonormal
- Loadings are positive stable, not Gaussian
- Loadings A_{lt} may not be independent
- Computing A and B is not as simple as a few matrix operations

Analogy with PCA

- Reduces the dimension from n to L
- Maps of $B_l(s)$ tell us about the most important spatial patterns
- Captures a non-stationary spatial dependence structure
- The v_l tell us how many important features are present
- Loadings A_{lt} can be estimated and fed into future analyses

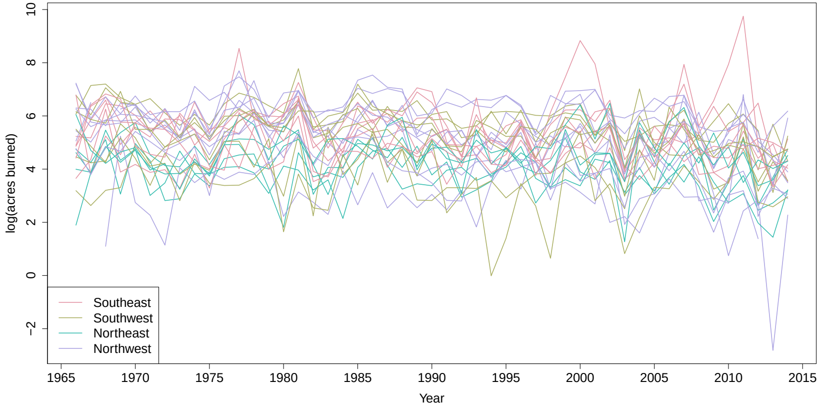
Bayesian implementation

- Given the basis function $B_l(\mathbf{s})$ we can proceed with MCMC to estimate the remaining parameters
- GEV location: $\mu_t(\mathbf{s}) = \beta_{\mu,int}(\mathbf{s}) + \beta_{\mu,time}(\mathbf{s})t$
- GEV scale: $\log[\sigma_t(\mathbf{s})] = \beta_{\sigma,int}(\mathbf{s}) + \beta_{\sigma,time}(\mathbf{s})t$
- GEV shape: ξ for all $\mathbf{s}$ and t
- The β have Gaussian process priors
- We use cross-validation (quantile and Bries scores) to select L
- Alternative: select L so that $\sum_{l=1}^{L} v_l = 0.8$

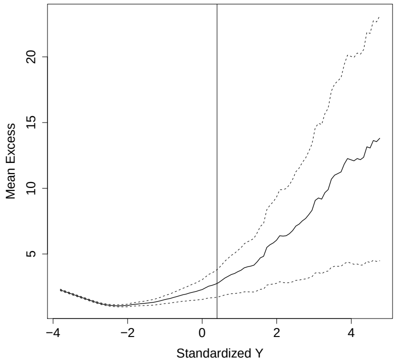
Application 1 - forest fires in GA

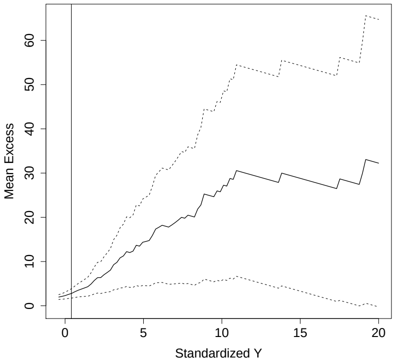
- The data are the number of acres burned by forest fires each year (1965-2014) in each county of Georgia
- We censor the data at the local 95^{th} percentile, $T(s)$
- The censored data are modeled as GEV with spatially-varying location and scale
- The objectives are to map fire risk and determine if it is changing with time

GA Fires – time series for each county

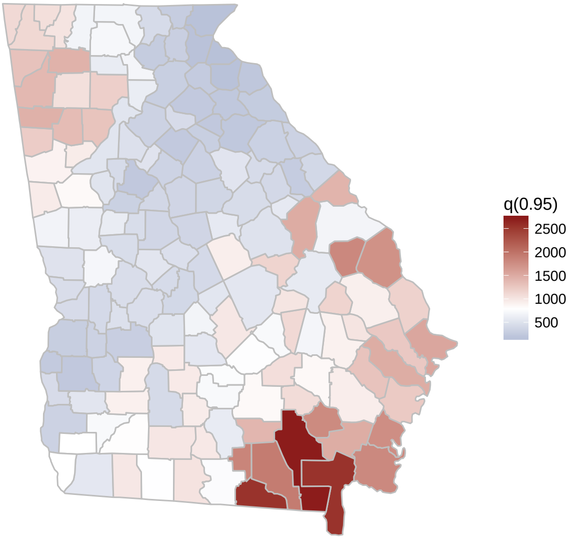


GA Fires – picking the threshold





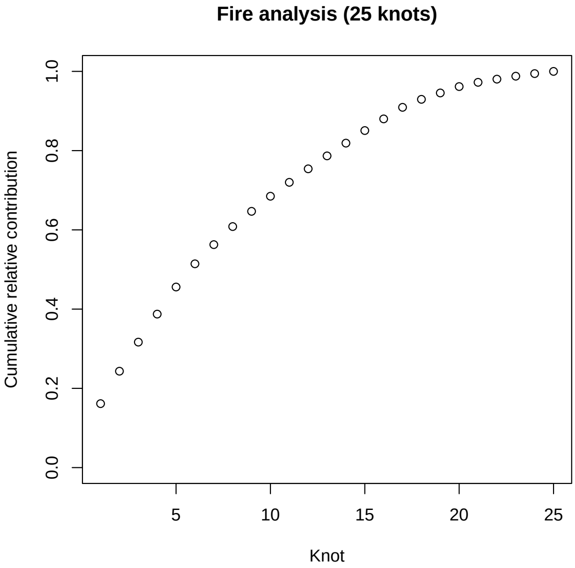
GA Fires – 95th percentile by county, $T(s)$



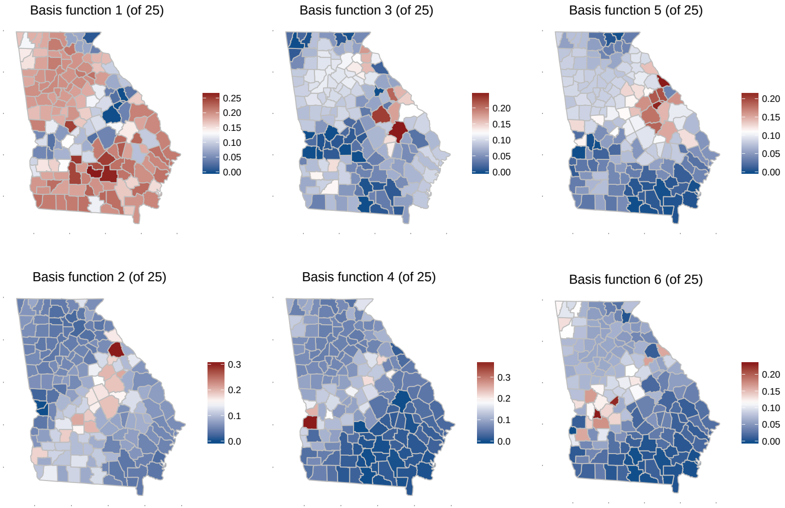
Model comparisons

L	Brier Score	Quantile score
5	5.64	135.7
10	5.33	127.3
15	5.00	128.3
20	4.93	122.4
25	**4.78**	**116.9**
40	4.72	115.7

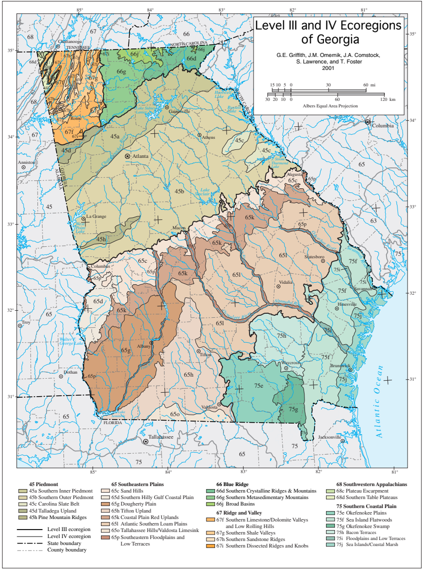
GA Fires – EBF weights, v_l



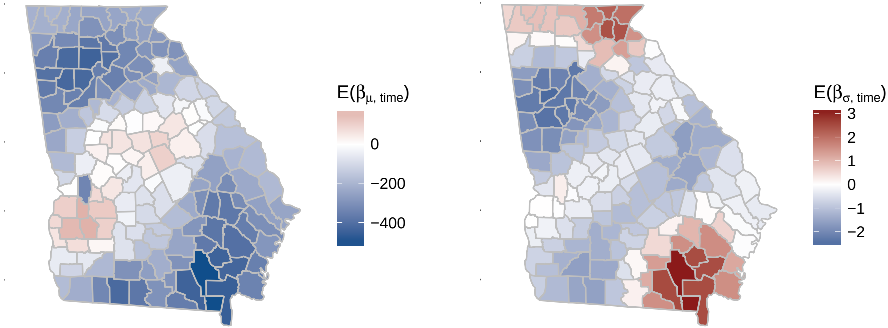
GA Fires – EBF's $B_l(s)$



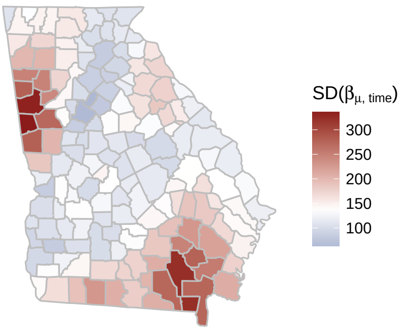
GA Fires – Ecoregions

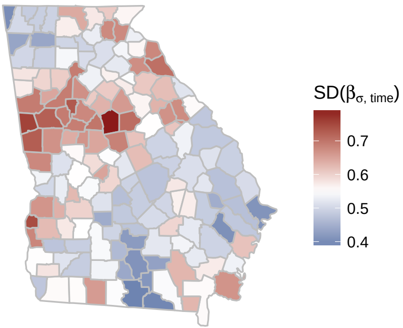


Time trend ($\beta_{\mu,time}$, $\beta_{\sigma,time}$) – posterior mean

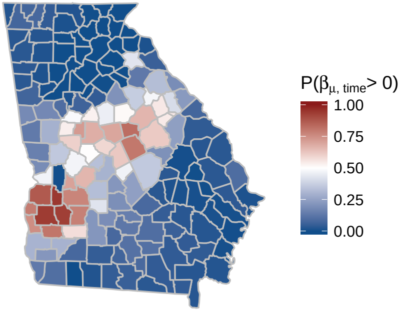


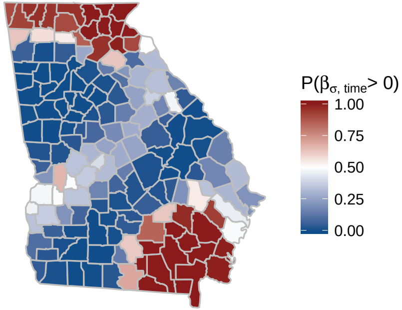
Time trend ($\beta_{\mu,time}$, $\beta_{\sigma,time}$) – posterior SD





Time trend ($\beta_{\mu,time}$, $\beta_{\sigma,time}$) – prob > 0

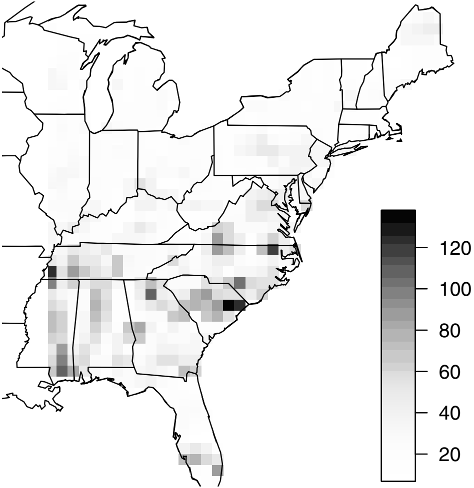




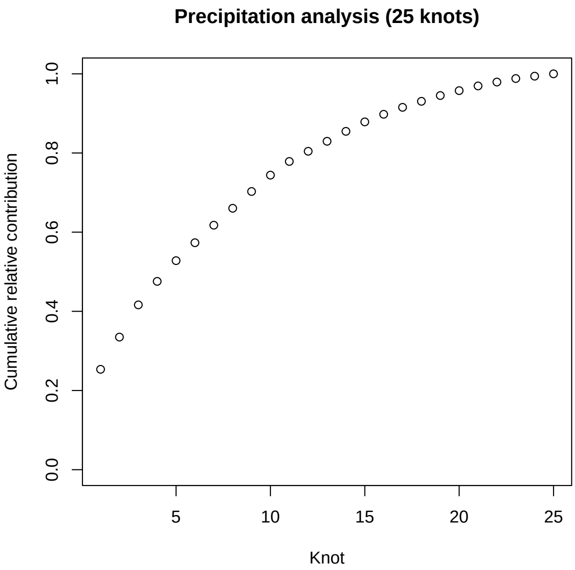
Application 2 – NARCCAP climate model output

- Data consist of annual maximum precipitation at 697 grid cells in the Eastern US
- The model is run separately for 1969-2000 and 2039-2077
- The objective is the compare the extremes in the two climate periods
- We fit the same model as for the fire data except without censoring
- We fit the model separately for the two periods

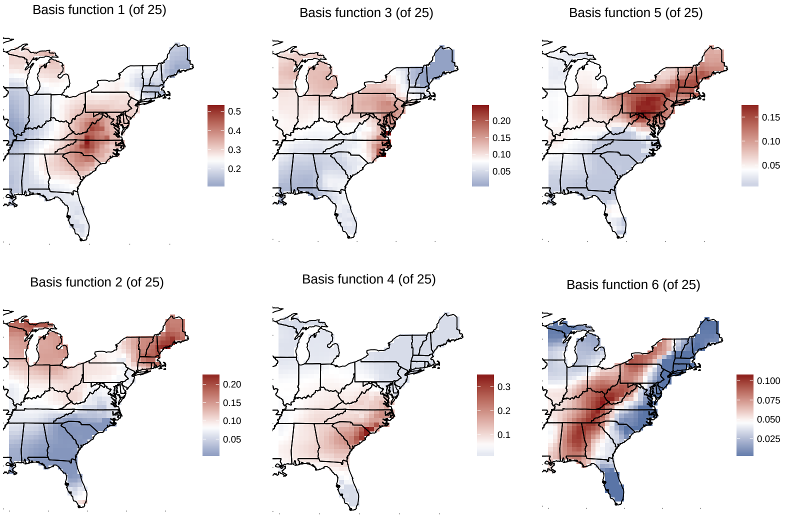
Climate model output for 1969



Precip – EBF weights, v_l



Precip – EBFs $B_l(s)$



Back to a Gaussian process model

- The max-stable process is an elegant approach, but does that mean it's the right model?
- In reality, it is only an approximation
- There are less complicated approximations
- For example, we could model daily data as a Gaussian process (GP)
- If the goal is spatial interpolation, perhaps this is competitive?

GP - asymptotic independence

- A GP leads to simple interpretation and computing, but asymptotic independence.
- The extremal dependence between $Y_t(\mathrm{s}_1)$ and $Y_t(\mathrm{s}_2)$ is

$$\chi(\mathrm{s}_1, \mathrm{s}_2) = \lim_{c \to \infty} \mathrm{Prob}[Y_t(\mathrm{s}_1) > c | Y_t(\mathrm{s}_2) > c]$$

- If $Y_t(\mathrm{s}_1)$ and $Y_t(\mathrm{s}_2)$ are bivariate normal then $\chi(\mathrm{s}_1, \mathrm{s}_2) = 0$, i.e., asymptotic independence
- This suggests Kriging will not capture extremes
- But so much is known for the Gaussian case: nonstationarity, multivariate, numerical approximations,...
- Rather than toss it out, can we patch it up?

Spatial skew-t process

A spatial skew-t process (Azzalinia and Capitanio, 2014) resembles a GP but exhibits asymptotic dependence

$$
\begin{aligned}
Y_t(\mathrm{s}) &= \mathrm{X}(\mathrm{s})^T\boldsymbol{\beta} + \lambda\sigma_t|z_t| + \sigma_t v_t(\mathrm{s}) \\
z_t &\sim \mathrm{Normal}(0,1) \\
\sigma_t^2 &\sim \mathrm{InvGamma}(a/2, b/2) \\
v_t &\sim \text{Spatial GP}
\end{aligned}
$$

- Location: $\mathrm{X}(\mathrm{s})^T\boldsymbol{\beta}$
- Scale: $b > 0$
- Skewness: $\lambda \in \mathcal{R}$
- Degrees of freedom: $a > 0$

Good properties

- Flexible t marginal distribution with four parameters including the degrees of freedom which allows for heavy tails ($a = 1$ gives a Cauchy)

- Computation on the order of a GP; the only extra steps are z_t and σ_t which have conjugate full conditionals

- Asymptotic dependence: $\chi(\mathbf{s}_1, \mathbf{s}_2) > 0$ for all $\mathbf{s}_1$ and $\mathbf{s}_2$

Bad properties and *remedies*

- Modeling all the data (bulk and extreme) can lead to poor tail probability estimates if the model is misspecified
- *We use a censored likelihood to focus on the tails*
- Long-range dependence: $\chi(\mathrm{s}_1, \mathrm{s}_2) > 0$ for all s_1 and s_2 even if s_1 and s_2 are far apart
- This occurs because all sites share z_t and σ_t
- *We propose a local skew-t process*

Censored likelihood

- Censored likelihood: We censor the data

$$\tilde{Y}_t(\mathrm{s}) = \begin{cases} T & \text{for } Y_t(\mathrm{s}) \leq T \\ Y_t(\mathrm{s}) & \text{for } Y_t(\mathrm{s}) > T \end{cases}$$

- Censoring is handled using standard Bayesian imputation methods

- The threshold T is chosen by cross-validation

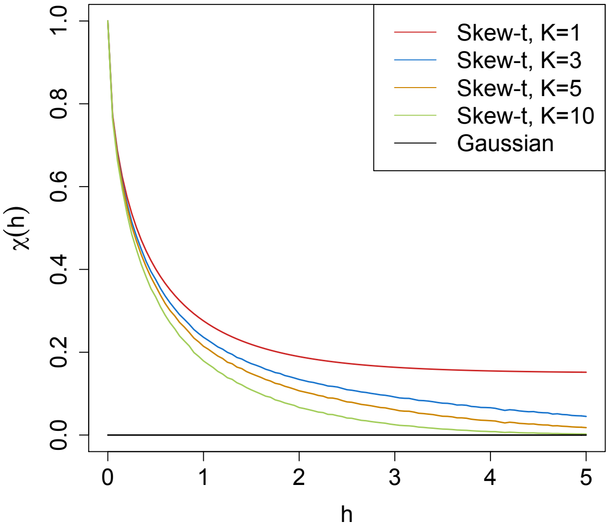
Local skew-t process

- Let the knots $\mathbf{v}_1, ..., \mathbf{v}_K$ follow a homogeneous Poisson process over the domain of interest (in practice we fix K)
- Associated with each is $z_{tk} \sim \text{Normal}(0, 1)$ and $\sigma^2_{tk} \sim \text{InvGamma}(a/2, b/2)$
- The knots partition the domain if we assign location $\mathbf{s}$ to subregion $k = \arg\min_l ||\mathbf{s} - \mathbf{v}_l||$.
- If $\mathbf{s}$ is in subregion k then

$$Y_t(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta} + \lambda \sigma_{tk} |z_{tk}| + \sigma_{tk} v_t(\mathbf{s})$$

- The marginal distribution remains a t, but partitioning breaks long-range spatial dependence

Extremal coefficient by $h = ||s_1 - s_2||$



Results of a simulation study

In terms of Brier and quantile scores for spatial prediction:

- Data generated as a GP:

  skew-t is close to GP
  max-stable is 15% worse than GP

- Data generated as a skew-t:

  skew-t is 15% better than GP
  max-stable is 30% worse than GP

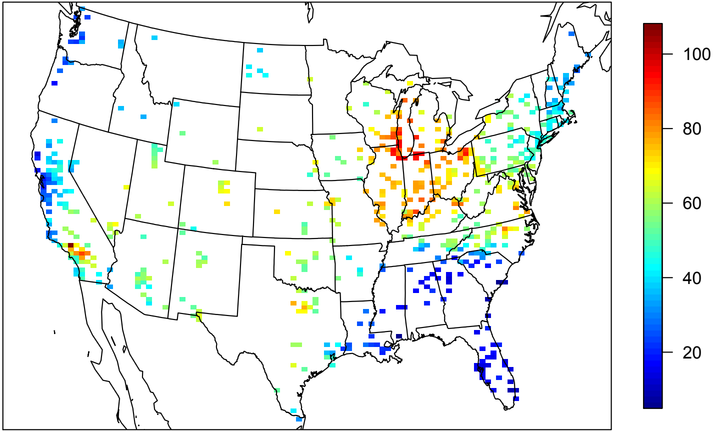
- Data generated as max-stable:

  skew-t is close to GP
  max-stable performs 10% better than GP

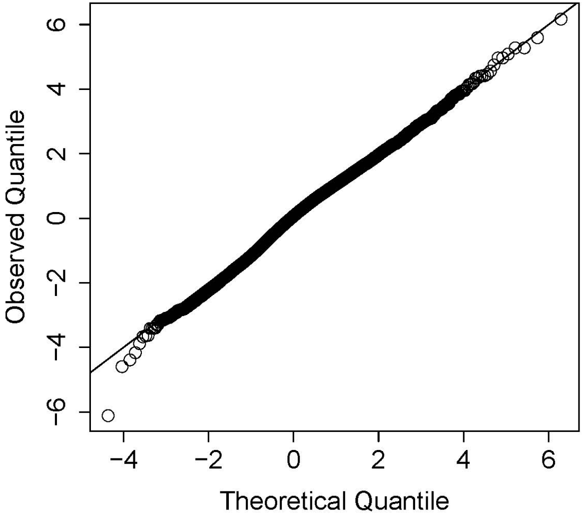
Application to ozone

- The USEPA has an extensive network of ozone monitors throughout the US
- We will analyze ozone for 31 days in July, 2005 at $n = 1,089$ stations
- Currently the EPA regulates the annual 99^{th} percentile
- Our objective is to map the probability of an extreme ozone event

Ozone on July 10



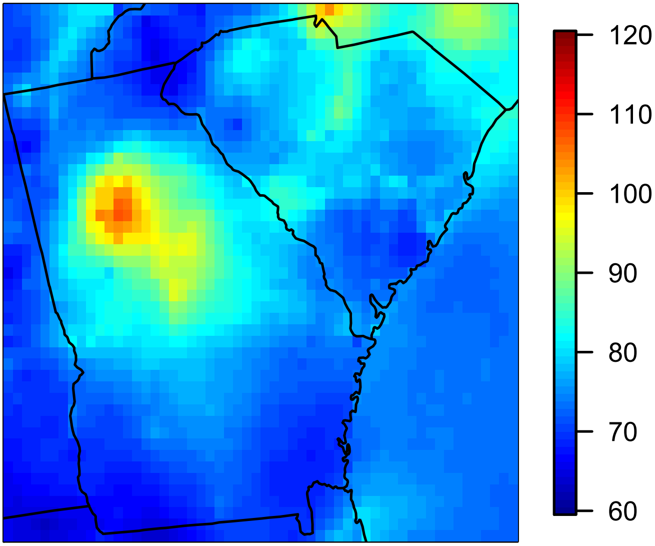
Skew-t qqplot



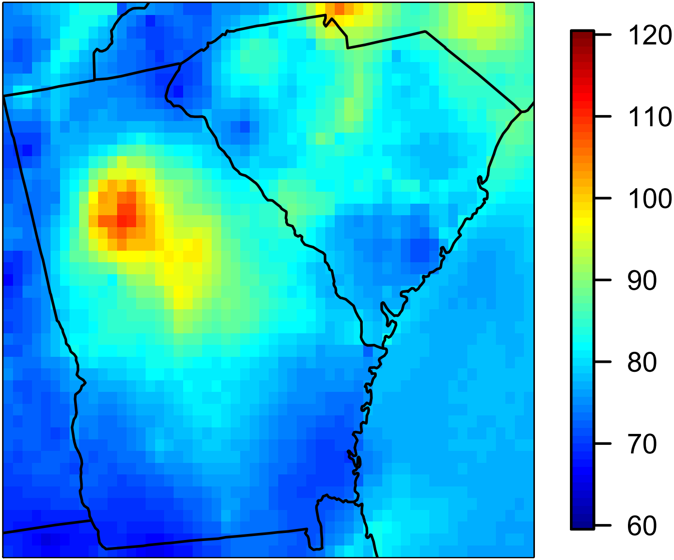
Cross-validation

- We split the sites into training and testing
- The model was fit assuming independence over days
- We found that $K = 15$ knots and censoring at T equal to the median gave the best results
- Results were not sensitive to these tuning parameters
- This model was 8% more accurate (Brier score) than GP
- The max-stable model fit was 15% less accurate than GP

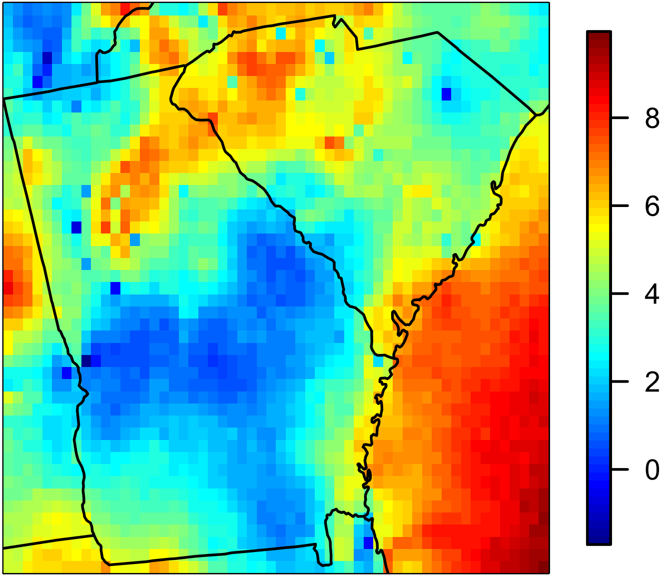
Fitted 99^{th} percentile - Gaussian



Fitted 99^{th} percentile - Skew-t



Difference (Skew-t - Gaussian)



Summary

- We proposed two methods to handle large spatial datasets: EBF and skew-t
- After this exploration, I personally feel:
  - EBF nice for exploratory analysis
  - Skew-t is a nice balance between theoretical properties and computational feasibility
- This should at least be used as a benchmark for more sophisticated approaches
- Work supported by NSF, NIH, DOI, and EPA
- Thanks!