

Spatiotemporal Modeling of Extreme Events

**Samuel Morris (NC State), Brian Reich (NC State),
and Dan Cooley (Colorado State)**

JSM 2014, Boston

Motivation

- ▶ Average behavior is important to understand, but it does not paint the whole picture.
 - ▶ e.g. When constructing river levees, engineers need to be able to estimate a 100-year or 1000-year flood levels.
- ▶ Spatial extreme borrows strength across space to estimate return levels and make predictions at unknown locations..
- ▶ In geostatistical analysis, kriging uses spatial correlation for prediction.
- ▶ Want to explore similar methods for extremes.

Standard analysis - Block maxima

- ▶ Uses yearly maxima
- ▶ Discards many observations
- ▶ Models are fit using the generalized extreme value distribution
- ▶ For a spatial analysis, max-stable processes give an appropriate limiting distribution

Standard analysis - Peaks over threshold

- ▶ Incorporates more data than block maxima
- ▶ Select a threshold, T , and fit data above the threshold using the generalized Pareto distribution
- ▶ Temporal dependence may be an issue between observations (e.g. flood levels don't dissipate overnight)
- ▶ Generalized Pareto distribution (GPD) is used for the exceedances.

Multivariate analysis

- ▶ Multivariate max-stable and GPD models have nice features, but they are
 - ▶ computationally challenging to work with
 - ▶ joint distribution only available in low dimension
- ▶ Pairwise likelihood approach (Huser and Davison, 2014)

Model objectives

- ▶ Our objective is to build a model that
 - ▶ has marginal distribution with a flexible tail
 - ▶ has asymptotic spatial dependence
 - ▶ has computation on the order of Gaussian models for large space-time datasets

Thresholding data

- ▶ We threshold the observed data at a high threshold T .
- ▶ Thresholded data:

$$Y_t^*(\mathbf{s}) = \begin{cases} Y_t(\mathbf{s}) & Y_t(\mathbf{s}) > T \\ T & Y_t(\mathbf{s}) \leq T \end{cases}$$

- ▶ Allows tails of the distribution to speak for themselves.

χ coefficient

- ▶ The χ coefficient is a measure of extremal dependence
- ▶ Specifically, we focus on $\chi(\mathbf{h})$ for the upper tail given by

$$\chi(\mathbf{h}) = \lim_{c \rightarrow \infty} \Pr(Y(\mathbf{s}) > c \mid Y(\mathbf{s} + \mathbf{h}) > c)$$

- ▶ If $\chi(\mathbf{h}) = 0$, then observations are asymptotically independent at distance \mathbf{h} .
- ▶ We expect $\lim_{\mathbf{h} \rightarrow \infty} \chi(h) = 0$.

Gaussian spatial model

- ▶ In geostatistics $Y(\mathbf{s})$ are often modeled using a Gaussian process with mean function $\mu(\mathbf{s})$ and covariance function $\rho(\mathbf{h})$.
- ▶ Model properties:
 - ▶ Nice computing properties (closed-form likelihood)
 - ▶ For a Gaussian spatial model $\lim_{c \rightarrow \infty} \chi(c) = 0$ regardless of the strength of the correlation in the bulk of the distribution.
 - ▶ Tail is not flexible (Gaussian is light tailed)

Spatial skew- t distribution

- ▶ Assume observed data $Y_t(\mathbf{s})$ come from a skew- t (Zhang and El-Shaarawi, 2012)

$$Y_t(\mathbf{s}) = X_t(\mathbf{s})\beta + \alpha z_t + v_t(\mathbf{s})$$

where

- ▶ $\alpha \in \mathcal{R}$ controls the skewness
- ▶ $z_t \stackrel{iid}{\sim} N_{(0,\infty)}(0, \sigma_t^2)$ is a random effect
- ▶ $v_t(\mathbf{s})$ is a Gaussian process with variance σ_t^2 and Matérn correlation
- ▶ $\sigma_t^2 \stackrel{iid}{\sim} \text{IG}(a, b)$

Spatial skew- t distribution

- ▶ **Conditioned** on z_t and σ_t^2 , $Y_t(\mathbf{s})$ is a Gaussian spatial model
- ▶ Can use standard geostatistical methods to fit this model.
- ▶ Predictions can be made through Kriging.
- ▶ **Marginalizing** over z_t and σ_t^2 (via MCMC),

$$Y_t(\mathbf{s}) \sim \text{skew-}t(\mu, \Sigma^*, \alpha, \text{df} = 2a)$$

where

- ▶ μ is the location
- ▶ a, b are the IG parameters for σ_t^2
- ▶ $\Sigma^* = \frac{b}{a}\Sigma$ is a scale matrix, and Σ is a Matérn covariance matrix
- ▶ $\alpha \in \mathcal{R}$ controls the skewness

Spatial skew- t distribution

- ▶ Model properties
 - ▶ Has flexible tail controlled by skewness α and degrees of freedom $2a$
 - ▶ For a skew- t distribution $\lim_{c \rightarrow \infty} \chi(c) > 0$ Padoan, 2011)
 - ▶ Computation that is on the order of Gaussian computation
- ▶ For this distribution, $\chi(\mathbf{h})$ shows asymptotic dependence that does not approach 0 as $\mathbf{h} \rightarrow \infty$
- ▶ This occurs because all observations (near and far) share the same z_t and σ_t^2 .
- ▶ We deal with this through a daily random partition (similar to Huser and Davison).

Daily random partition

- ▶ Daily random partition allows z_t and σ_t^2 to vary by site.

$$Y_t(\mathbf{s}) = X_t(\mathbf{s})\beta + \alpha z_t(\mathbf{s}) + \sigma(\mathbf{s})v_t(\mathbf{s})$$

- ▶ Consider a set of daily knots $w_{tk} \sim \text{Uniform}$ that define a random daily partition P_{t1}, \dots, P_{tK} such that

$$P_{tk} = \{s : k = \arg \min_{\ell} ||\mathbf{s} - w_{t\ell}||\}$$

- ▶ For $\mathbf{s} \in P_{tk}$

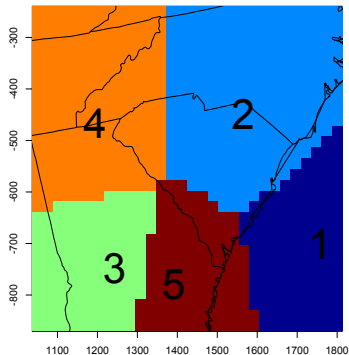
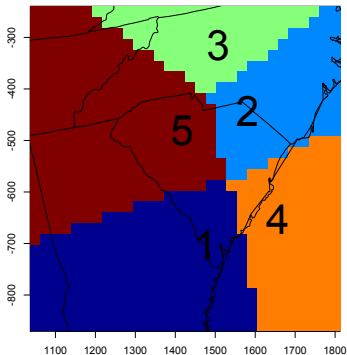
$$z_t(\mathbf{s}) = z_{tk}$$

$$\sigma_t^2(\mathbf{s}) = \sigma_{tk}^2$$

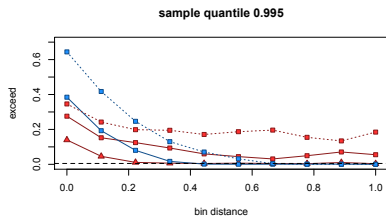
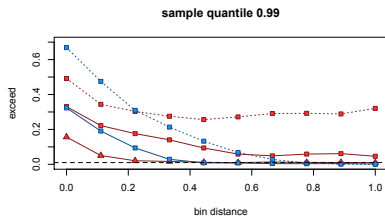
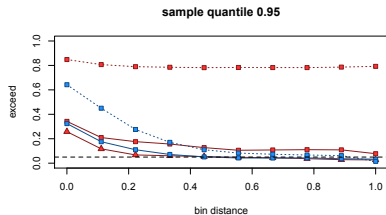
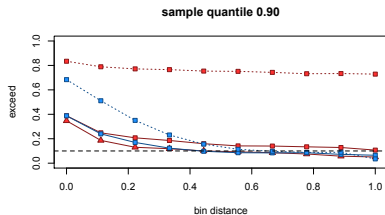
- ▶ Within each partition $Y_t(\mathbf{s})$ has the same MV skew-t distribution as before.

Example daily partition

Two sample partitions (number is at partition center)

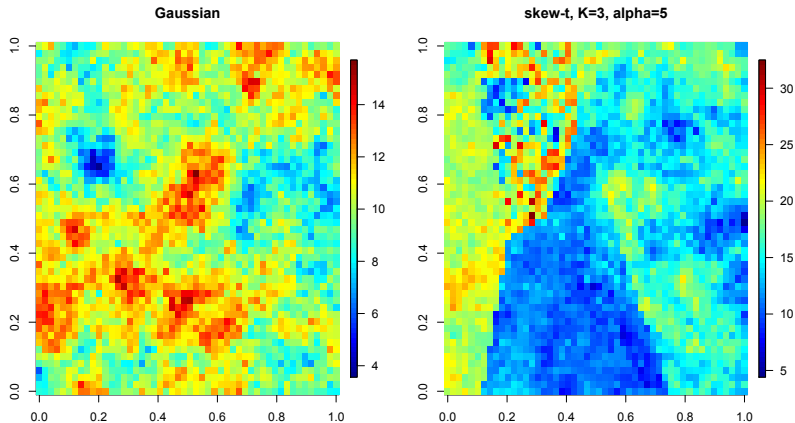


Simulated χ plots



- ▲ Gaussian
- t, K=1
- t, K=5
- skew-t, K=1
- skew-t, K=5

Sample simulated datasets



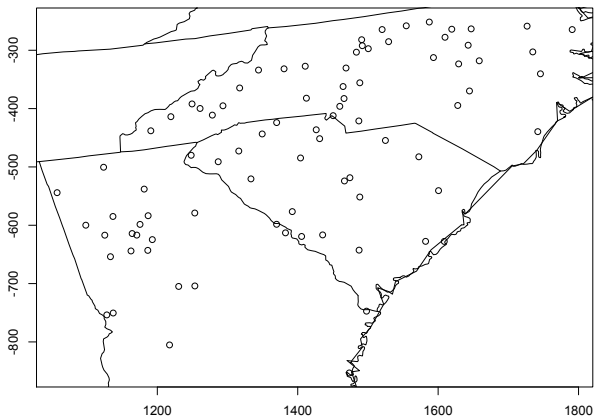
Spatiotemporal Model

- ▶ We can account for time in one of two ways
 - ▶ The mean: e.g. $AR(1)$
 - ▶ Three dimensional covariance model (e.g. Huser and Davison, 2014)

MCMC details

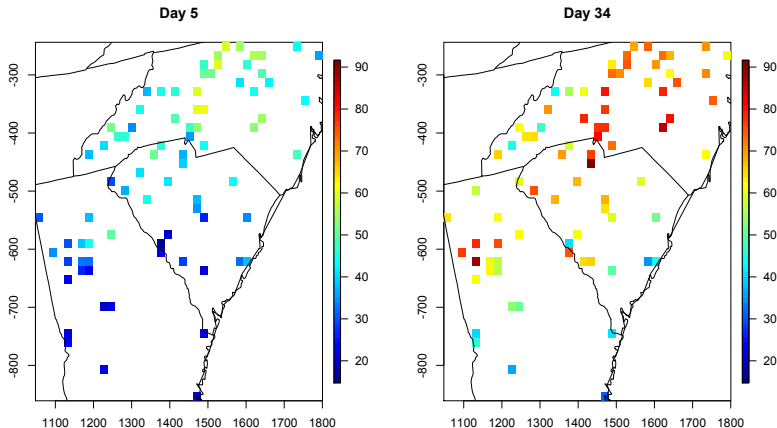
- ▶ Three main steps:
 1. Impute missing observations and censored data below T
 2. Update parameters with standard random walk Metropolis Hastings or Gibbs sampling
 3. Make spatial predictions
- ▶ Priors are selected to be conjugate when possible.

Ozone monitoring station locations



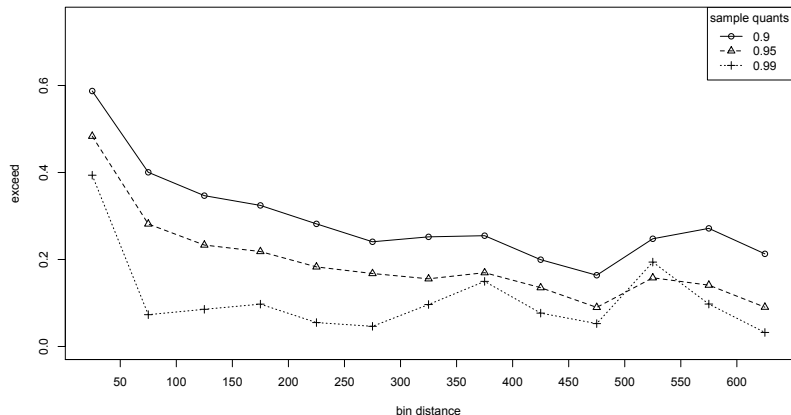
Data analysis

Max 8-hour ozone measurements at 85 sites in NC, SC, and GA
for days 5 and 34.



Exploratory data analysis

χ -plot for residuals selected ozone sample quantiles



Model comparisons

- ▶ 9 different analysis methods incorporating
 - ▶ Gaussian vs t vs skew- t marginal distribution
 - ▶ $K = 1$ partition vs $K = 5$ partitions
 - ▶ No thresholding vs thresholding at $T = 0.90$ sample quantile
- ▶ All methods use a Matérn or exponential covariance ($\nu = 0.5$)
- ▶ Compare quantile and Brier scores using 5-fold cross validation (Gneiting and Raftery, 2007)
- ▶ Mean function modeled using a first-order spatial trend

Quantile score for cross-validation

- ▶ The quantile score for the τ th quantile is

$$2\{I[y < \hat{q}(\tau)] - \tau\}(\hat{q} - y)$$

where:

- ▶ y is a test set value
- ▶ $\hat{q}(\tau)$ is the estimated τ th quantile

- ▶ The Brier score for predicting exceedance of threshold c is

$$[e(c) - P(c)]^2$$

where

- ▶ y is a test set value
- ▶ $e(c) = I[y > c]$
- ▶ $P(c)$ is the predicted probability of exceeding c

Five-fold cross-validation results

Marginal	K	T	Quantile				
			0.900	0.950	0.990	0.995	0.999
Gaussian	1	0	16.38	15.76	14.52	14.08	13.22
t	1	0	16.15	15.51	14.00	13.43	12.32
t	5	0	13.61	12.66	10.96	10.40	9.34
skew t	1	0	9.24	7.27	4.13	3.27	1.96
skew t	5	0	15.81	14.46	11.57	10.57	8.60
t	1	0.9	5.52	3.58	1.77	1.47	1.10
t	5	0.9	5.98	4.27	2.41	2.03	1.49
skew t	1	0.9	4.91	3.16	1.45	1.16	0.82
skew t	3	0.9	5.58	3.78	1.93	1.58	1.11

- Brier score results are similar.

Simulation study

- ▶ 6 different data settings:
 - ▶ Gaussian vs t vs skew- t marginal distribution
 - ▶ $K = 1$ partition vs $K = 5$ partitions
- ▶ Results are similar to the results from the data analysis
- ▶ Biggest gains come from thresholding.
- ▶ Using skew models give additional gain, but small relative to gain for thresholding.

Future work

- ▶ Comparison with extreme value analysis methods
- ▶ Reporting of compliance with EPA ozone standards
- ▶ Including time in the model via standard spatiotemporal Gaussian models.

Questions

- ▶ Questions?
- ▶ Thank you for your attention.
- ▶ Acknowledgment: This work was funded by EPA STAR award R835228.

References

- ▶ Demarta, S. and McNeil, A. J. (2007) The t copula and related copulas. *International Statistical Review*, **73**, 111–129.
- ▶ Huser, R. and Davison, A. C. (2014) Space-time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, 439–461.
- ▶ Padoan, S. A. (2011) Multivariate extreme models based on underlying skew- t and skew-normal distributions. *Journal of Multivariate Analysis*, **102**, 977–991.
- ▶ Zhang, H. and El-Shaarawi, A. (2010) On spatial skew-Gaussian processes and applications. *Environmetrics*, **21**, 33–47.