Spatiotemporal Modeling of Extreme Events

Samuel Morris and Brian Reich

North Carolina State University

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.
- In geostatistical analysis, kriging uses spatial correlation to help inform prediction at unknown locations.
- Want to explore computationally easy methods that are available in higher dimensions

Standard non-spatial analysis

- Block maxima:
  - Uses yearly maxima
  - Discards many observations
  - Models are fit using the generalized extreme value distribution
- Generalized extreme value distribution (GEV):

$$\Pr(Y_j < y) = G_j(y) = \exp\left\{ - \left[\left(1 + \xi_j \frac{y - \mu_j}{\sigma_j} \right)_+^{-1/\xi_j} \right] \right\}$$

Standard non-spatial 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
  - Autocorrelation may be an issue between observations (e.g. flood levels don't dissipate overnight)
- Generalized Pareto distribution (GPD):

$$\Pr(Y_j > y | Y_j > T) = F_j(y) = \left(1 + \xi_j \frac{y - T}{\sigma_j}\right)_+^{-1/\xi_j}$$

Multivariate analysis

- Multivariate max-stable and GPD models have nice features, but they are
  - computationally hard 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 a flexible tail
  - has asymptotic spatial dependence
  - 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.

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 skew
  - $z_t \overset{iid}{\sim} N_{(0,\infty)}(0, \sigma_t^2)$ is a random effect
  - $v_t(\mathbf{s})$ is a Gaussian process with variance σ_t^2 and Matérn correlation
  - $\sigma_t^2 \overset{iid}{\sim} \text{IG}(a, b)$

Skew-normal distribution

- Conditioned on z_t and σ_t^2, $Y_t(\mathbf{s})$ is Gaussian
  - After marginalizing over σ_t^2, this becomes a multivariate t-distribution with $2a$ degrees of freedom.
- 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),

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

Long-range dependence

- The χ coefficient is a measure of extremal spatial correlation

$$\chi(\mathbf{h}) = \Pr(Y_t(\mathbf{s}) > c \mid Y_t(\mathbf{s}+\mathbf{h}) > c)$$

- This value shows asymptotic dependence that does not approach 0 as $\mathbf{h} \to \infty$.
- Deal with this through a daily random partition.

Random daily partition

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

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

- Consider a set of daily knots $\{w_{t1}, \ldots, w_{tK}\}$ that define a daily partition $\{P_{t1}, \ldots, P_{tK}\}$ such that

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

- For $\mathbf{s} \in P_{tk}$

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

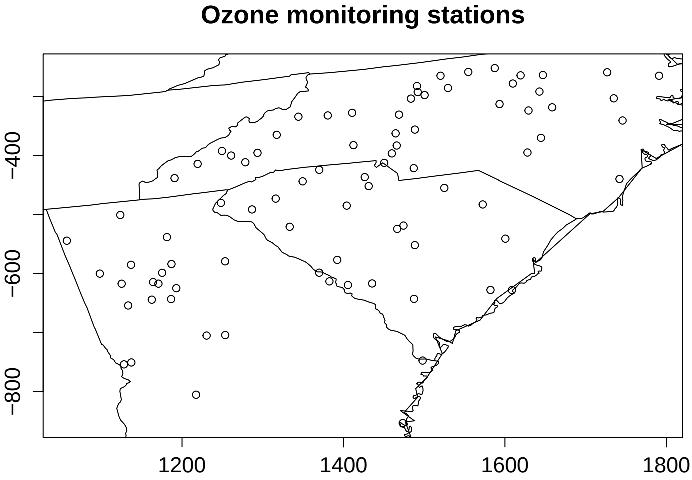
- Within each partition $Y_t(\mathbf{s})$ has the same distribution as before.

MCMC details

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

Data analysis

- Ozone analysis at 85 sites in NC, SC, and GA for 92 days



Model comparisons

- 9 different analysis methods incorporating
  - Gaussian vs t
  - 1 partition vs 5 partitions
  - No thresholding vs thresholded
    - Thresholded data at 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
- Mean function modeled using a first-order spatial trend

Cross-validation results

	Quantile					
	0.900	0.950	0.980	0.990	0.995	0.999
Gaussian	16.38	15.76	15.02	14.52	14.08	13.22
t-1 ($T = 0$)	16.15	15.51	14.62	14.00	13.43	12.32
t-5 ($T = 0$)	13.61	12.66	11.61	10.96	10.40	9.34
t-1 ($T = 0.9$)	5.52	3.58	2.28	1.77	1.47	1.10
t-5 ($T = 0.9$)	5.98	4.27	2.99	2.41	2.03	1.49
skew t-1 ($T = 0$)	9.24	7.27	5.26	4.13	3.27	1.96
skew t-1 ($T = 0.9$)	**4.91**	**3.16**	**1.94**	**1.45**	**1.16**	**0.82**
skew t-5 ($T = 0$)	15.81	14.46	12.74	11.57	10.57	8.60
skew t-5 ($T = 0.9$)						

- Quantile scores – Still waiting on last set

Simulation study settings

- Data generated from 6 different settings.
  1. Gaussian
  2. t-1 with 4 degrees of freedom
  3. t-5 with 4 degrees of freedom
  4. skew t-1 with 4 degrees of freedom ($\alpha = 3$)
  5. skew t-5 with 4 degrees of freedom ($\alpha = 3$)
  6. Half-Gaussian, Half t-1 (spatial range = 0.4)
- Spatial settings
  - $\mathbf{s} \in [0, 1] \times [0, 1]$
  - Exponential covariance with range: 0.1
  - Ratio of spatial to nugget error: 0.9

Simulation study methods

- 5 different analysis methods
  1. Gaussian
  2. Skew t-1 ($T = 0$)
  3. Skew t-1 ($T = 0.9$)
  4. Skew t-5 ($T = 0$)
  5. Skew t-5 ($T = 0.9$)
- All methods use a Matérn covariance structure except for method 5 which uses an exponential covariance ($\nu = 0.5$)

Simulation results

Still rerunning a few of the MCMC runs because quantile scores for 4. and 5. are unusually high for method 5

Future work

- Comparison with extreme value analysis methods
- Including time in the model
  - AR(1): $Y_t(\mathbf{s}) = X_t(\mathbf{s})\beta + \phi Y_{t-1}(\mathbf{s}) + \alpha z_t(\mathbf{s}) + v_t(\mathbf{s})$

Questions

- Any questions?
- Thank you for your attention.

References

- Cooley, D., Cisewski, J., Erhardt, R. J., Mannshardt, E., Omolo, B. O. and Sun, Y. (2012) A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects. *REVSTAT*, **10**, 135–165.
- 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
- Zhang, H. and El-Shaarawi, A. (2010) On spatial skew-Gaussian processes and applications. *Environmetrics*, **21** 33–47.