# Spatiotemporal Modeling of Extreme Events

**Brian Reich and Samuel Morris**

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 ways to incorporate this spatial correlation when estimating the tails of the distribution.

# Introduction to extremes

- Max-stable processes (Cooley et al., 2012):
  - Consider a spatial process $x_t(\mathbf{s})$, $t = 1, \ldots, T$.
  - Let $M_T(\mathbf{s}) = \left\{ \bigvee_{t=1}^T x_t(\mathbf{s}_1), \ldots, \bigvee_{t=1}^T x_t(\mathbf{s}_n) \right\}$
  - If there exists normalizing sequences $a_T(\mathbf{s})$ and $b_T(\mathbf{s})$ such that for all sites, $\mathbf{s}_i$, $i = 1, \ldots, d$,

$$a_T^{-1}(\mathbf{s}) \{ M_T(\mathbf{s}) - b_T(\mathbf{s}) \} \xrightarrow{d} Y(\mathbf{s})$$

    which has a non-degenerate distribution, then $Y(\mathbf{s})$ is a max-stable process.

# General approaches

- Two standard approaches:
  1. Block maxima:
     - Uses yearly maxima
     - Discards many observations
  2. Peaks-over-threshold
     - Incorporates more data than block maxima
     - Autocorrelation may be an issue between observations (e.g. flood levels don't dissipate overnight)

# Univariate distribution functions

- Generalized extreme value distribution (GEV):

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

- 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 representations

- Multivariate distributions:
  - Assume common standardized max-stable marginal, like unit-Fréchet

$$\Pr(Z < z) = exp(-z^{-1})$$

  - The multivariate representation for the GEV is

$$\Pr(\mathbf{Z} \leq \mathbf{z}) = G^*(\mathbf{z}) = \exp(-V(\mathbf{z}))$$

$$V(\mathbf{s}) = d \int_{\Delta_d} \bigvee_{i=1}^{d} \frac{w_i}{z_i} H(\mathrm{d}w)$$

where

- $\Delta_d = \{\mathbf{w} \in \mathcal{R}_+^d \mid w_1 + \cdots + w_d = 1\}$
- $H$ is a probability measure on $\Delta_d$
- $\int_{\Delta_d} w_i H(\mathrm{d}w) = 1/d$ for $i = 1, \ldots, d$.

- Although $V(\mathbf{s})$ can be flexible, the number of parameters is unwieldy.
  - For an asymmetric logistic dependence, we have $2^d - d - 1$ free parameters
- New model builds on existing geostatistical methods.

# Skew-normal distribution

- Assume observed data $Y_t(\mathbf{s})$ come from a skew-normal (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
  - $v_t(\mathbf{s}) \sim \text{MVN}(0, \Sigma)$ where $\Sigma_t$ is a spatial covariance matrix with variance $\sigma_t^2$

# Skew-normal distribution

- By conditioning on $\alpha$ and $z_t$

$$Y_t(\mathbf{s}) \mid \alpha, z_t \sim \text{MVN}(\mu_t(\mathbf{s}), \Sigma_t)$$

  where $\mu_t(\mathbf{s}) = X_t(\mathbf{s})\beta + \alpha z_t$

- When $\sigma_t^2 \overset{iid}{\sim} \text{IG}(a, b)$, then this becomes a multivariate $t$-distribution after marginalizing over the $\sigma_t^2$ terms.
- Can use standard geostatistical methods to fit this model.
- Predictions can be made through kriging.

# Random daily partition

- 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 $\mathbf{s} \in P_{tk}$, the model becomes
  - $\mu_t(\mathbf{s}) = X_t(\mathbf{s})\beta + \alpha z_{tk}$
  - $\sigma_t^2 = \sigma_{tk}^2$
- After marginalizing over the $\sigma_{tk}^2$ terms, $Y_t(\mathbf{s})$ is a multivariate $t$-distribution within the partition.

# Thresholding data

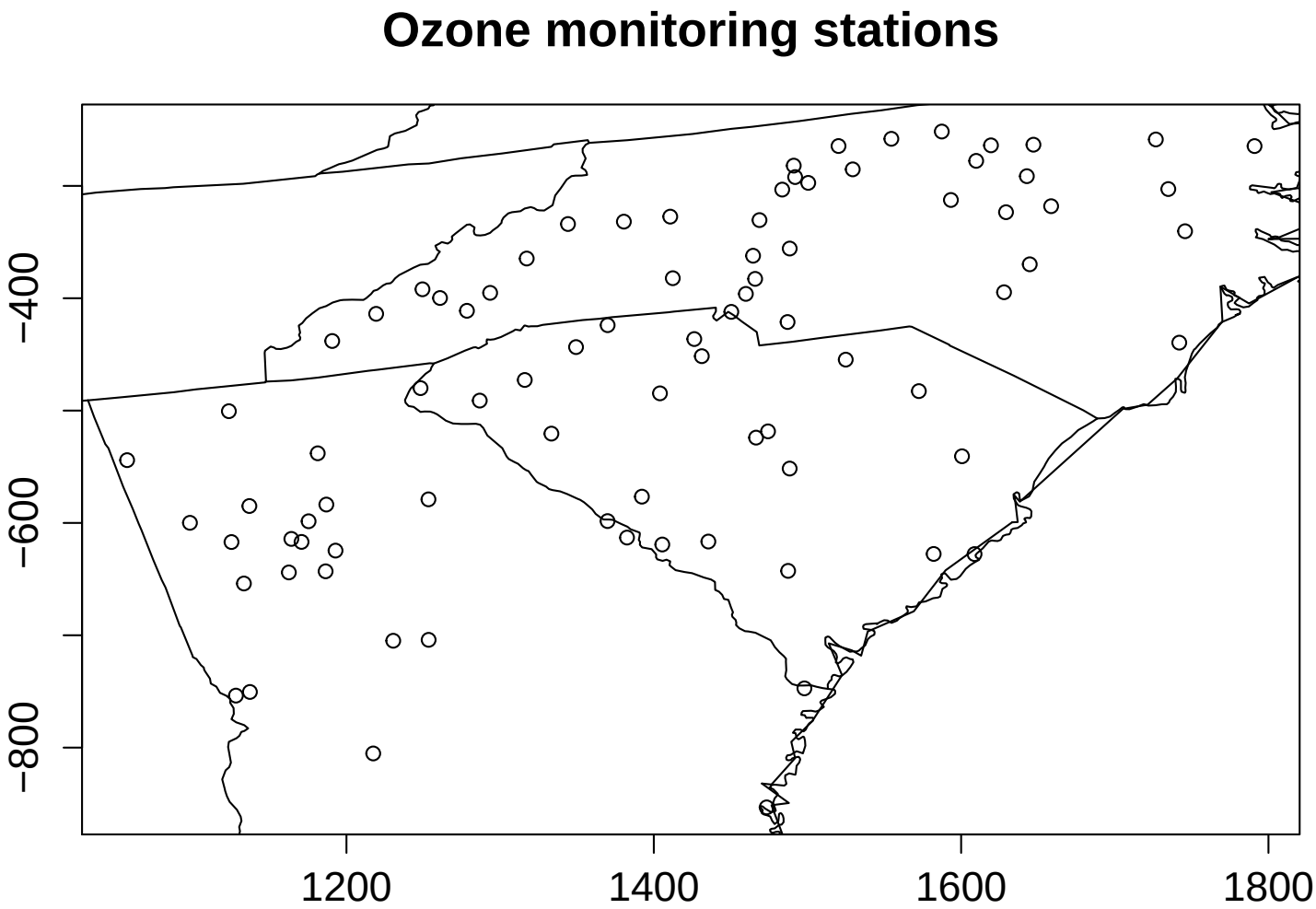
- Tails of the distribution should speak for themselves.
- We threshold the observed data at a suitably high threshold $T$.
- Observations below the threshold are imputed during MCMC.

# 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

# Data analysis

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



# Model comparisons

- 9 different analysis methods
  1. Gaussian
  2. $t$, $K = 1$, $T = 0$
  3. $t$, $K = 5$, $T = 0$
  4. $t$, $K = 1$, $T = 0.9$
  5. $t$, $K = 5$, $T = 0.9$
  6. skew-$t$, $K = 1$, $T = 0$
  7. skew-$t$, $K = 1$, $T = 0.9$
  8. skew-$t$, $K = 5$, $T = 0$
  9. skew-$t$, $K = 5$, $T = 0.9$
- Compare quantile scores using 5-fold cross validation

# 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.
- Zhang, H. and El-Shaarawi, A. (2010) On spatial skew-Gaussian processes and applications. *Environmetrics*, 33–47.