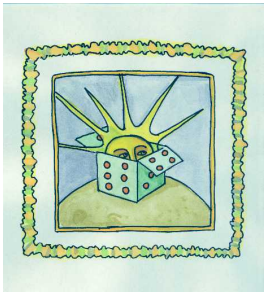
An introduction and case study using extremes

Douglas Nychka, Eric Gilleland, Uli Schnieder
National Center for Atmospheric Research

- Some examples
- Extreme value theory and methods (GEV,GPD)
- Space-time model for ozone
- Spatial model for ozone extremes



Supported by the National Science Foundation DMS

Extreme Events

"Man can believe the impossible. But man can never believe the improbable."

- Oscar Wilde

Some Examples

Precipitation

Extremes are used to determine flood potential for urban areas and also for dam specifications.

Ecological models are often driven by meteorology.

Typically extremes are described by the return period: "A 100 or 500 year event".

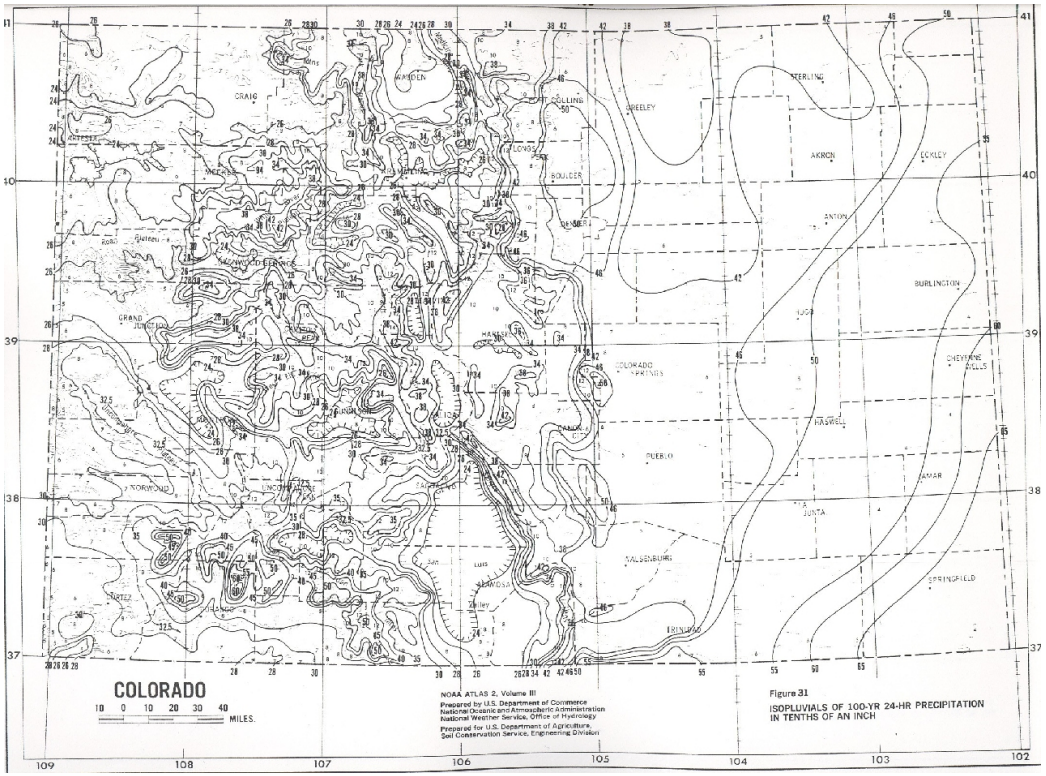
How does one determine this from 50 years of data?

Fort Collins Flood, July 1997

Heaviest rains ever documented over an urbanized area in Colorado (10 inches in 6 hours).

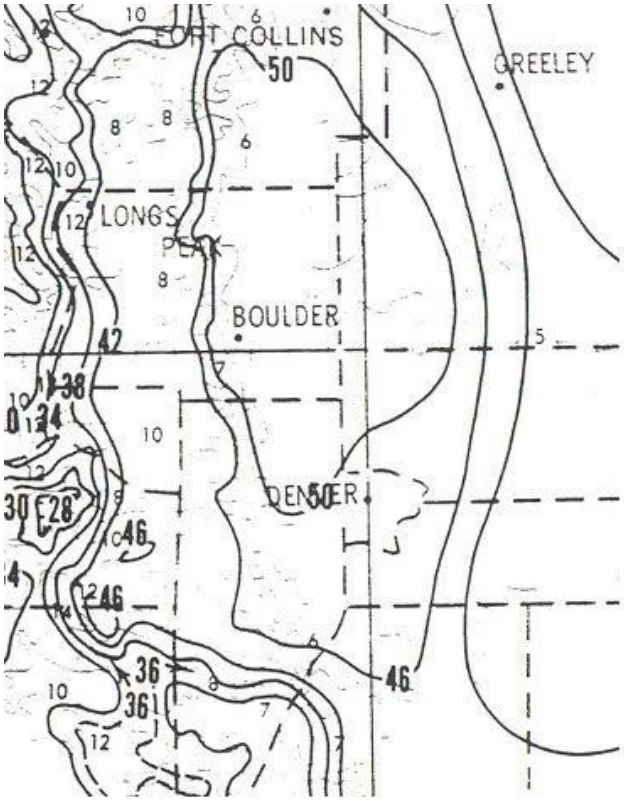
5 dead, 54 injured, 200 homes destroyed, 1,500 structures damaged.

Data product for high rain fall rates in CO is the precipitation atlas.

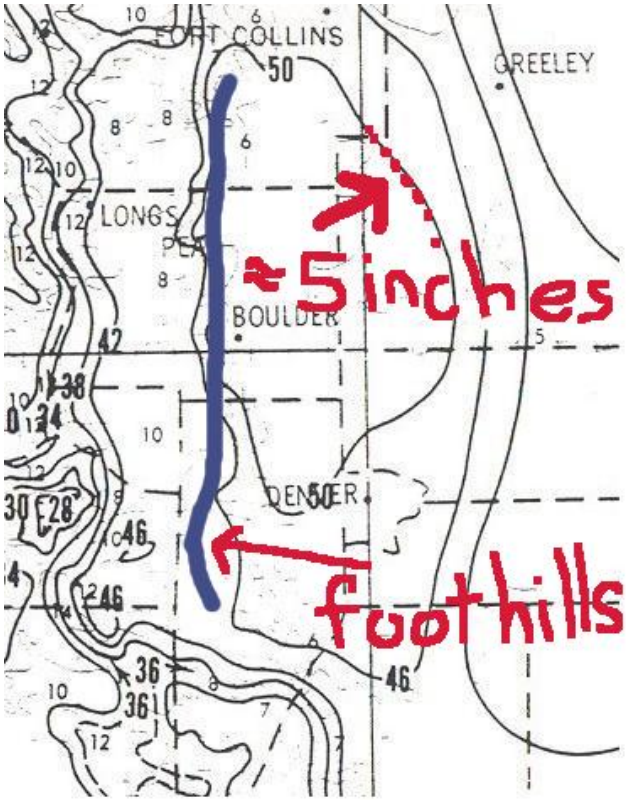


No quantification of uncertainty!

Area around Boulder



Area around Boulder



100 year event: $\approx$ 5.0 inches of rain in 24 hours

Surface level ozone

Ozone(O_3) is formed from volatile organic compounds (both pollutants and naturally occurring compounds) emitted into the atmosphere and NO_x in the presence of sunlight.

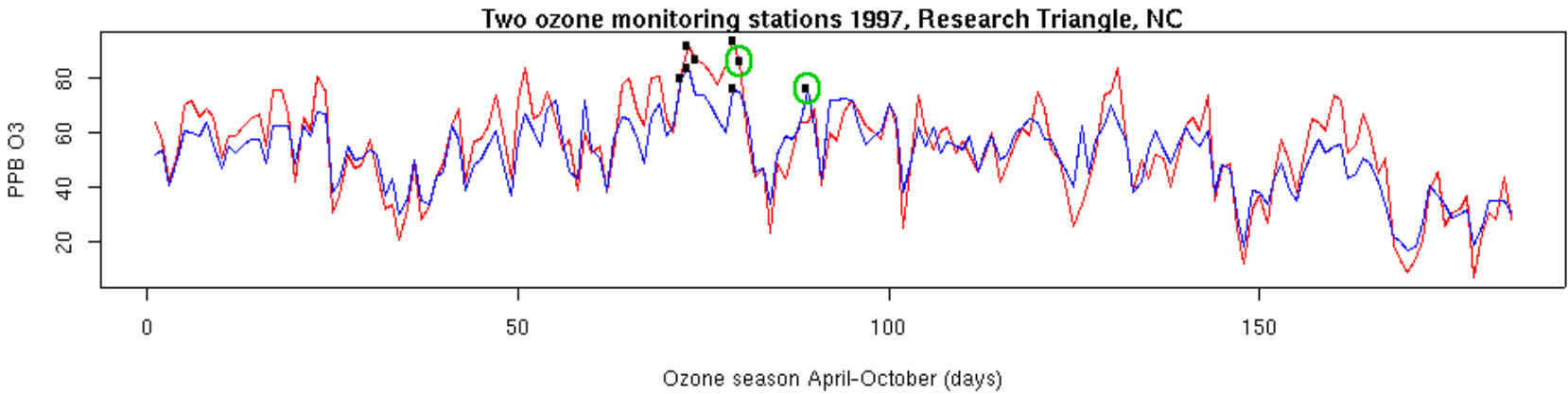
High ozone concentrations cause respiratory problems and sustained lower levels damage vegetation.

US EPA regulations based on the Clean Air Act:

A suggested ozone pollutant standard is based on the fourth *highest* (max) 8-hour daily average (FHDA) recorded during the year. A region is in attainment if the three year average is less than 80 PPB.

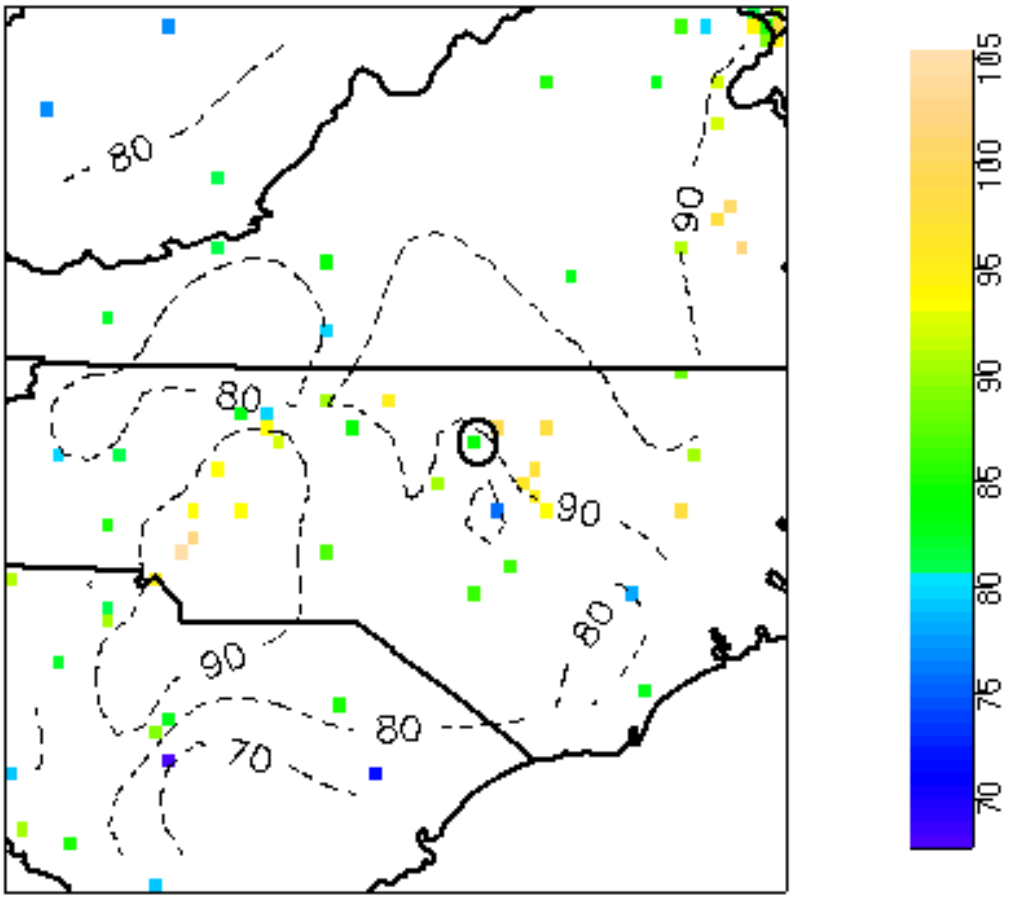
Example of calculating the FHDA statistic

The ozone "season" is 184 days. $\approx$ May-September.



$$FHDA = (1 - \tfrac{4}{184})\ \textit{quantile of the ozone distribution}$$

RTP study region with FHDA's for 1997



Spatial extremes: our problem today

Based on both of these examples:

Given a spatial process $Z(\boldsymbol{x})$, what can we say about

$$P(Z(\boldsymbol{x}) > z_0)$$

when z_0 is large?

Spatial extremes: our problem today

Based on both of these examples:

Given a spatial process $Z(\boldsymbol{x})$, what can we say about

$$P(Z(\boldsymbol{x}) > z_0)$$

when z_0 is large?

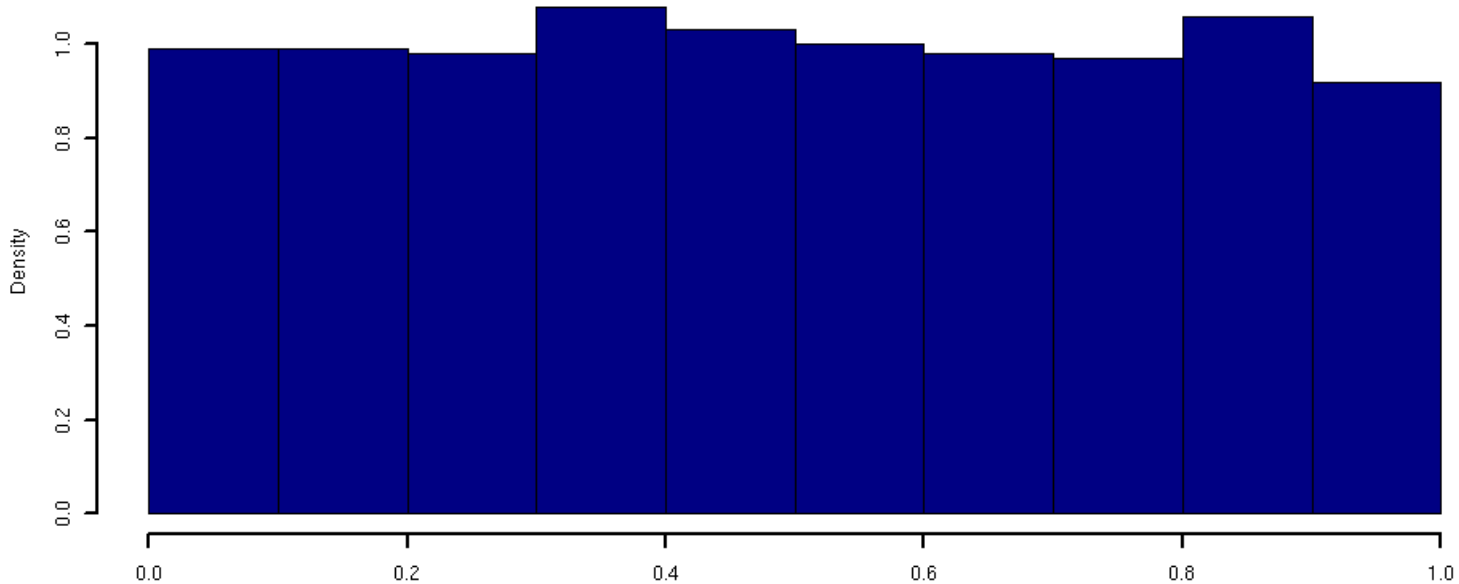
Note:

This is not about dependence between $Z(\boldsymbol{x})$ and $Z(\boldsymbol{x}')$ – this is another topic!

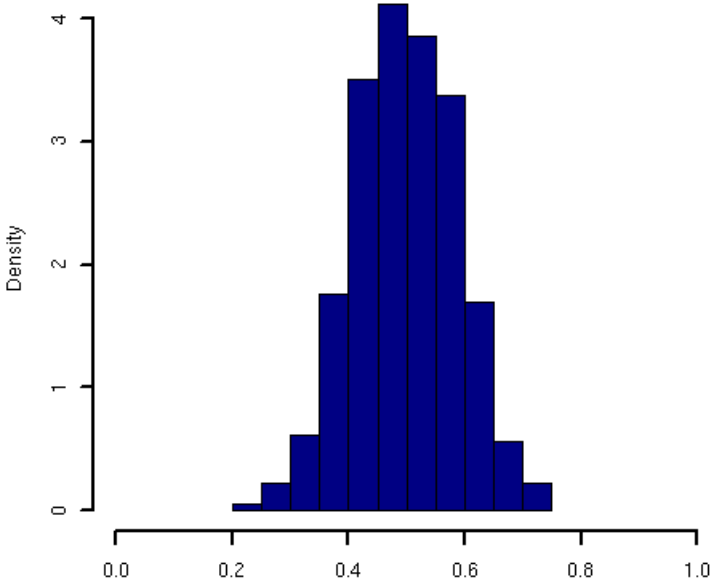
Extreme value distributions

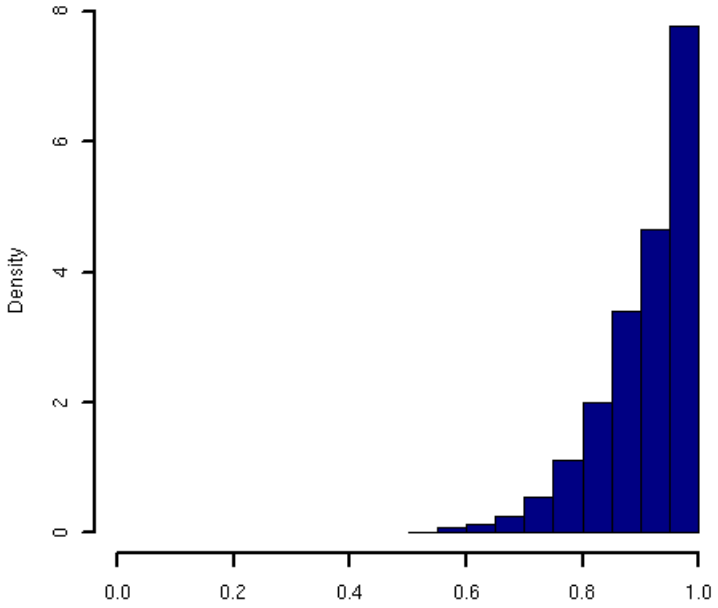
An example

Random sample 1000 Uniforms

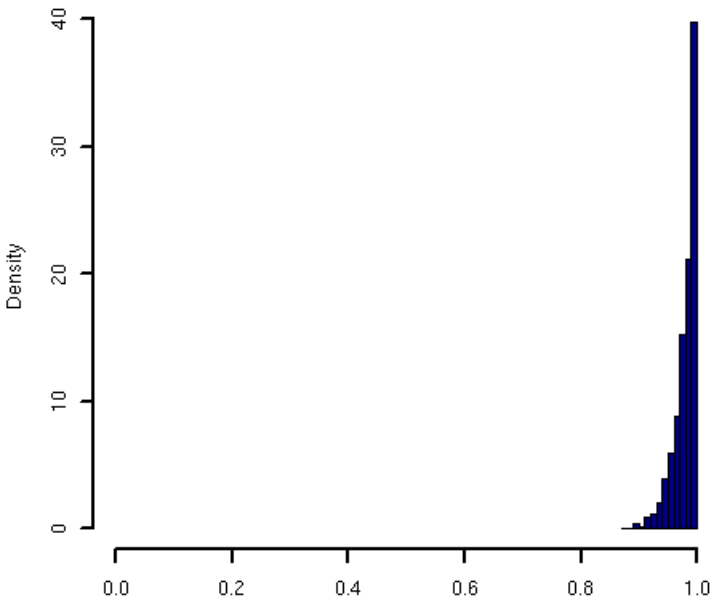


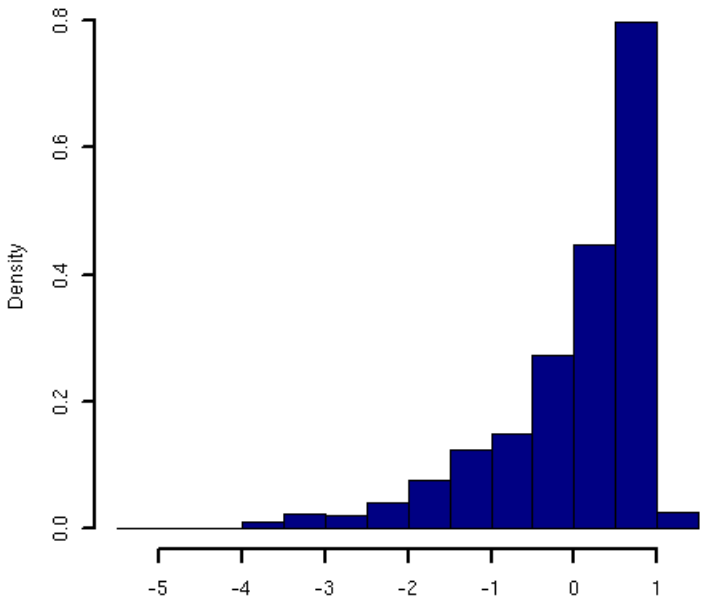
Random sample 1000, *mean* of 25 uniforms, *max* of 25



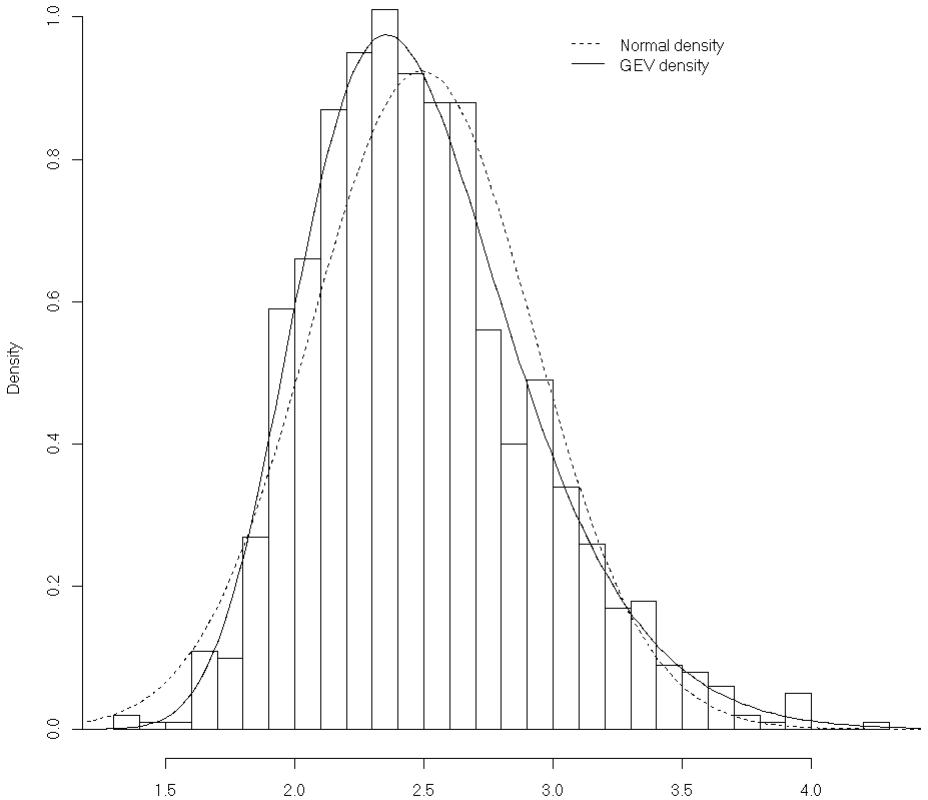


Random sample 1000, max of 100 , standardized



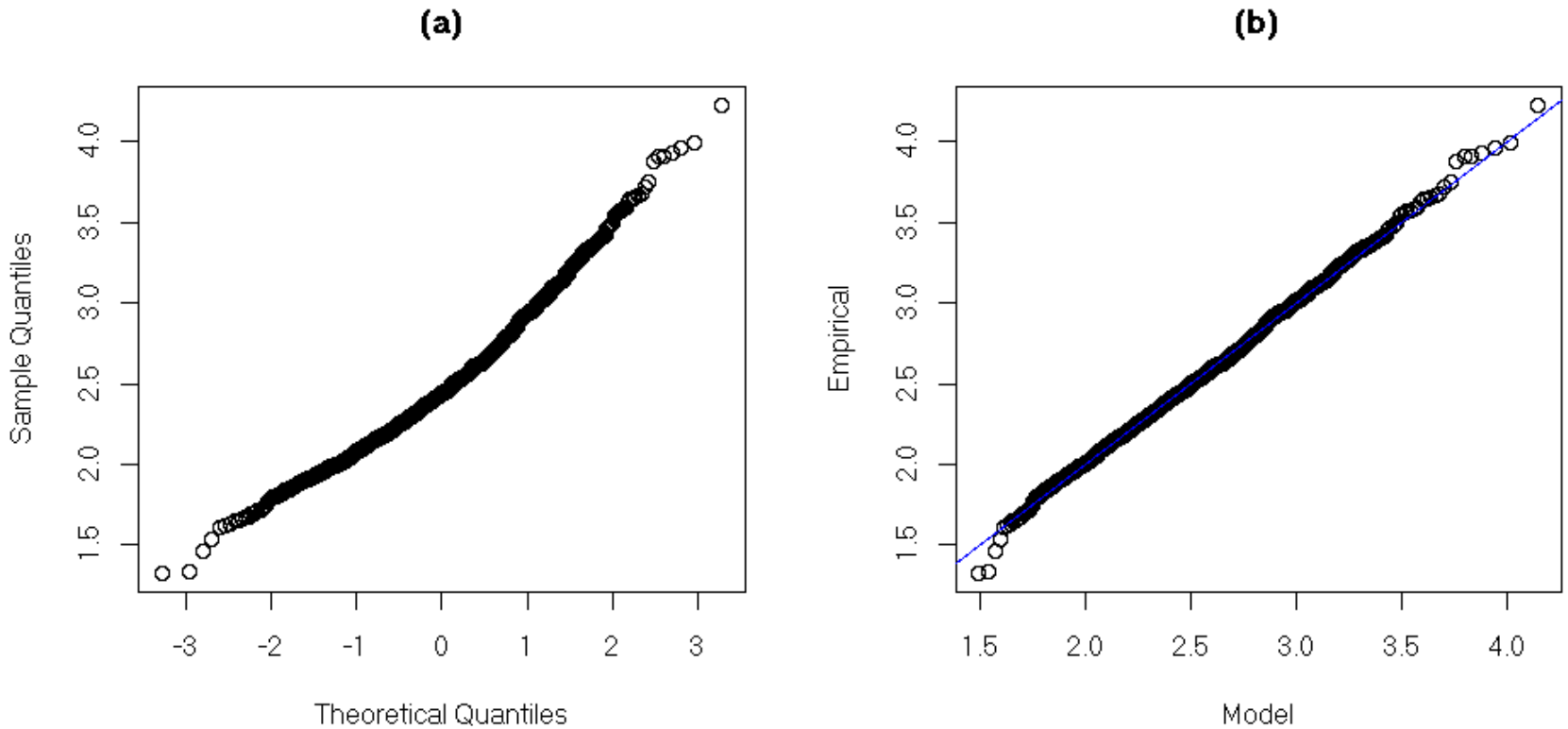


Example with Normal, max of 100 Normals



A Q-Q diagnostic plot

The maxima against a normal (a) and a GEV (b).



Generalized extreme value distribution (GEV)

Clearly, it was not serendipity that the GEV distribution was chosen as an example for the second Q-Q plot!

GEV cumulative distribution function:

$$F_{GEV}(z) = e^{-(1+\xi(z-\mu)/\sigma)^{-1/\xi}}.$$

Parameters: location (μ), scale (σ) and shape (ξ).

Some points about the GEV

- **When the shape is negative, the GEV distribution is zero above $\mu - \sigma/\xi$.**
- **M_n is the maximum of n random variables,**

$$P(M_n < z) \approx F_{GEV}(z),$$

- **Special case for the normal, the *Gumbel***

$$F_{\mathbf{Gumbel}}(z) = 1 - e^{-e^{-(z-\mu)/\sigma}}.$$

- **GEV includes the exponential and Wiebul**

Generalized Pareto Distribution (GPD)

Exceedence over threshold model

X is random and u is a (large) threshold

Conditional distribution

The probability of X exceeding x given that X is greater than u is

$$P(X > x|X > u) = [X > x|X > u] = \frac{1 - F(x)}{1 - F(u)}$$

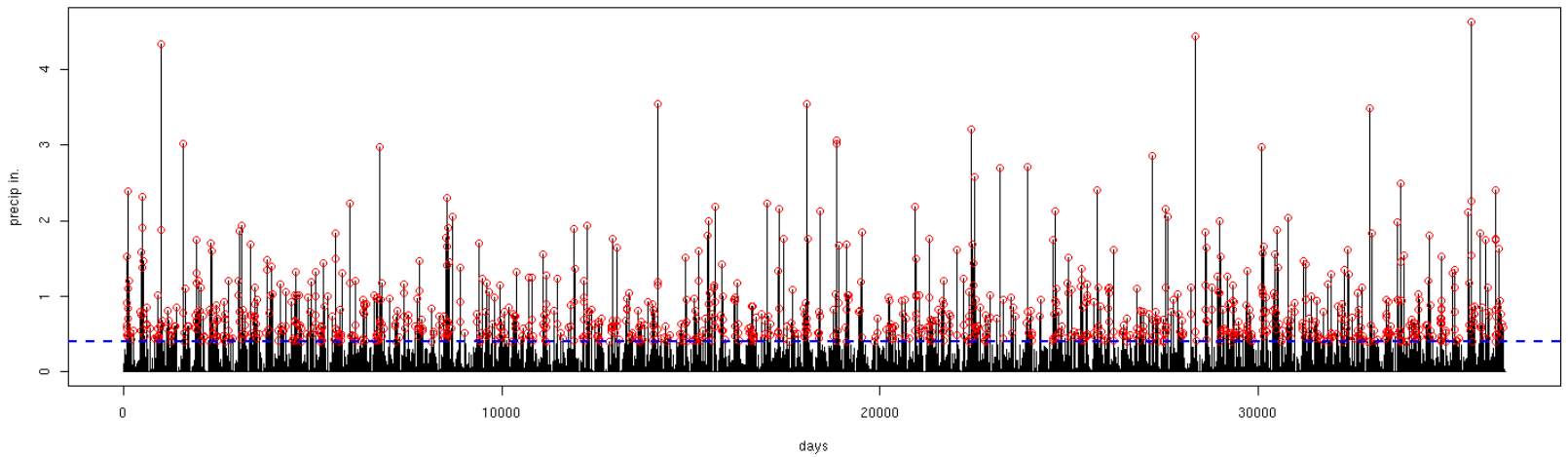
The GPD:

$$\frac{1 - F(x)}{1 - F(u)} \approx (1 + \xi(x - u)/(\sigma))^{-1/\xi}$$

for x and u "large".

Fort Collins Precip Series

Daily precipitation with threshold of .395



Quantiles and return levels

Given a GEV or GPD model for the tail of a distribution, a useful transformation is the quantile function.

Quantile

F a distribution function, the p-quantile is $x_p = F^{-1}(p)$. E.g. the probability of *exceeding* x_p is $1 - p$.

Return levels

- **For *annual* maxima the 100 year return level is the (1- (1/100))= .99 quantile.**

- **For a GPD fit to *daily* exceedences the return level is the $1 - 1/(365 * 100)$ quantile.**

The GEV and GPD have simple forms for their quantiles, but they are nonlinear in the parameters.

Fitting the GEV and GPD models to data

Use Maximum likelihood

The likelihood is easy to evaluate and maximize. Using a Bayes method is more difficult ... .

Choosing the threshold for the GPD

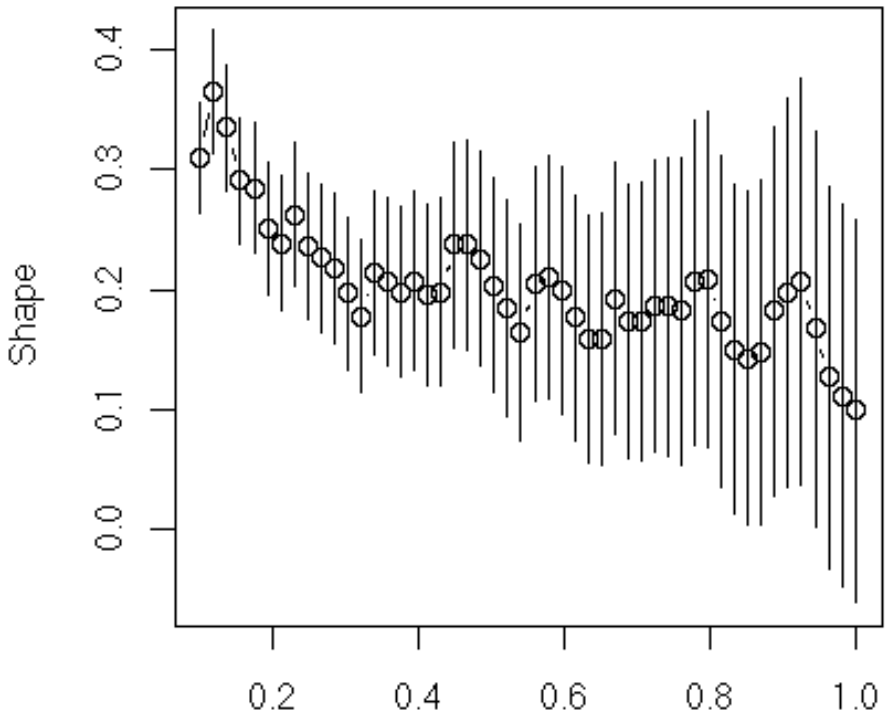
Find a range for u where the GPD parameters do not vary much.

Profile likelihoods are useful for inference on return levels

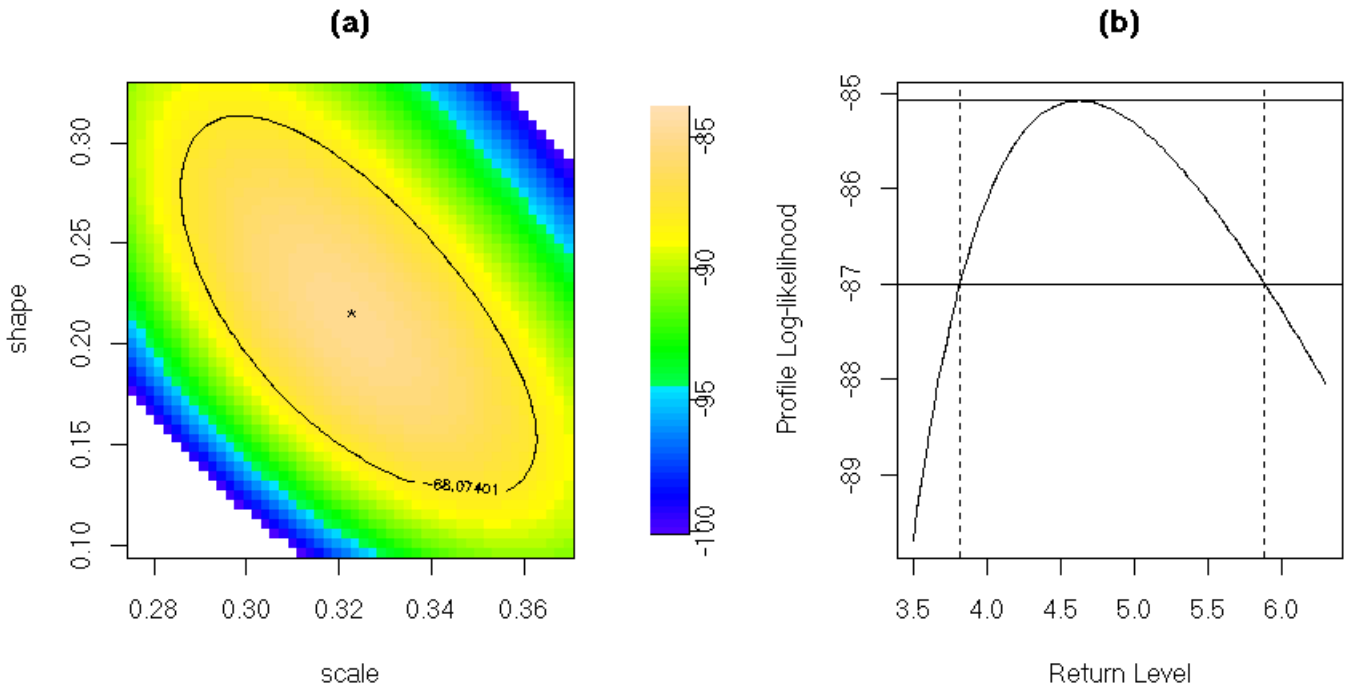
Approximate confidence intervals and sets are based on the likelihood surface and a χ^2 critical value.

Fort Collins precip example

Variation of GPD parameters with threshold – the shape, ξ

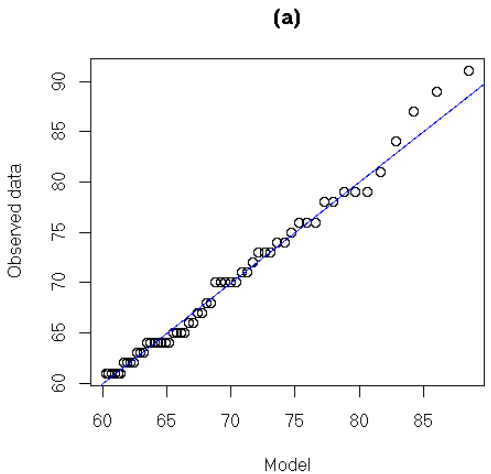


Suggests a threshold of 0.395**.**



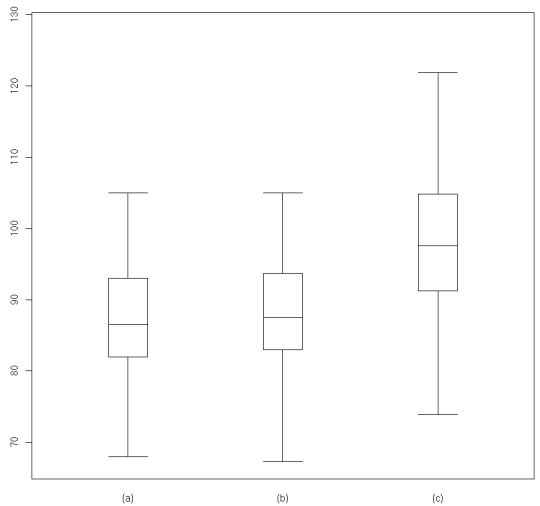
(a) confidence set for σ and ξ and (b) the profile likelihood for the 50 year return level.

Fitting a daily ozone series u = 60 PPB



Estimates across station network.

To get an estimate of FHDA substitute $\hat{\sigma}$ and $\hat{\xi}$ into the GPD quantile function (with p = 1- 4/184).



(a) observed FHDA, (b) estimated FHDA and (c) five-year return levels

A modest *space-time* model for ozone

Key idea: Conditional Simulation

For unmonitored locations find the conditional distribution of the FHDA. The distribution of the fields does not have a closed form and so we just generate samples from it.

Space-time model for daily values

In order to simulate from the FHDA field one needs a model for the temporal and spatial dependence of daily ozone. Transformed and scaled daily ozone follows an autoregressive model with spatially correlated shocks.

Model components

Transformation: $y(\boldsymbol{x}, t) =$ **8-hour ozone at location $\boldsymbol{x}$ and time t.**

$$u(\boldsymbol{x}, t) = \frac{y(\boldsymbol{x}, t) - \mu(\boldsymbol{x}, t)}{\sigma(\boldsymbol{x})}$$

Autoregression: $u(\boldsymbol{x}, t) = \rho(\boldsymbol{x})u(\boldsymbol{x}, t) + e(\boldsymbol{x}, t)$

Spatial dependence: $e(\boldsymbol{x}, t)$ **uncorrelated over time and stationary over time.**

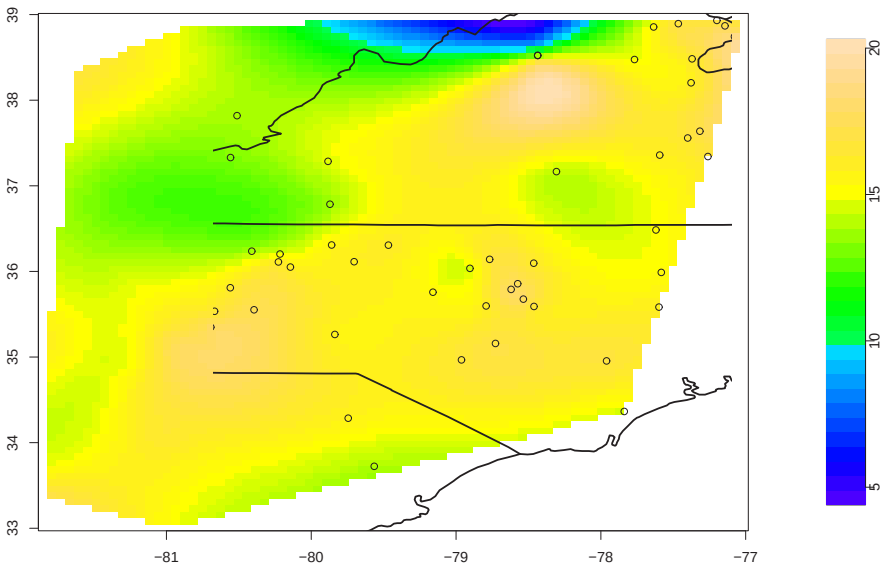
$$COV(e(\boldsymbol{x}, t), e(\boldsymbol{x}', t)) = (1 - \rho(\boldsymbol{x})^2)k(||\boldsymbol{x} - \boldsymbol{x}'||)$$

Under the assumption of multivariate normality one can generate fields of daily ozone conditional on the observed values.

Transformation

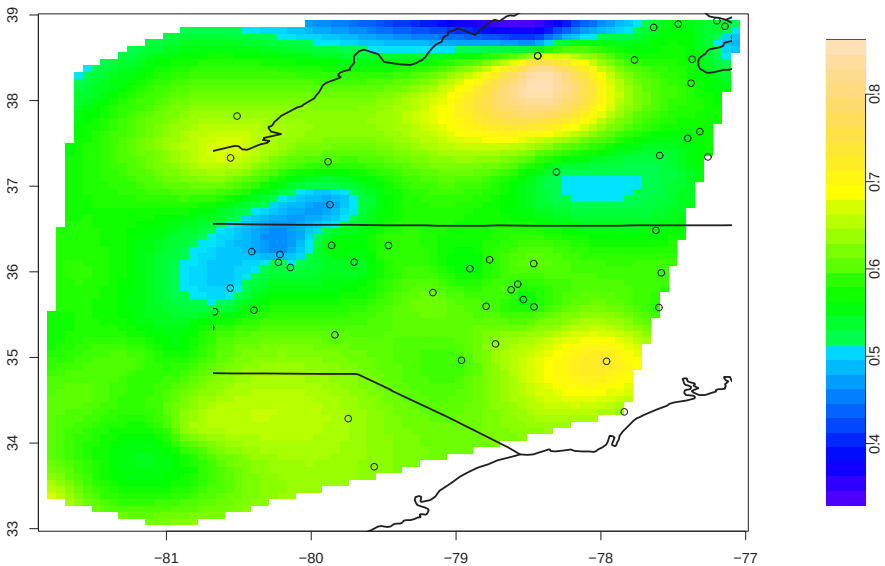
$\mu(\boldsymbol{x}, t)$ fit for each station location with a sine/cosine expansion and then smoothed over space using PC applied to regression coefficients. Extrapolation to unmonitored locations using interpolating thin plate splines (TPS) .

$\sigma(\boldsymbol{x})$ **also based on TPS interpolation of station estimates.**



Autoregressive model

$\rho(\boldsymbol{x})$ found from autoregression on transformed station data and then extrapolated using TPS

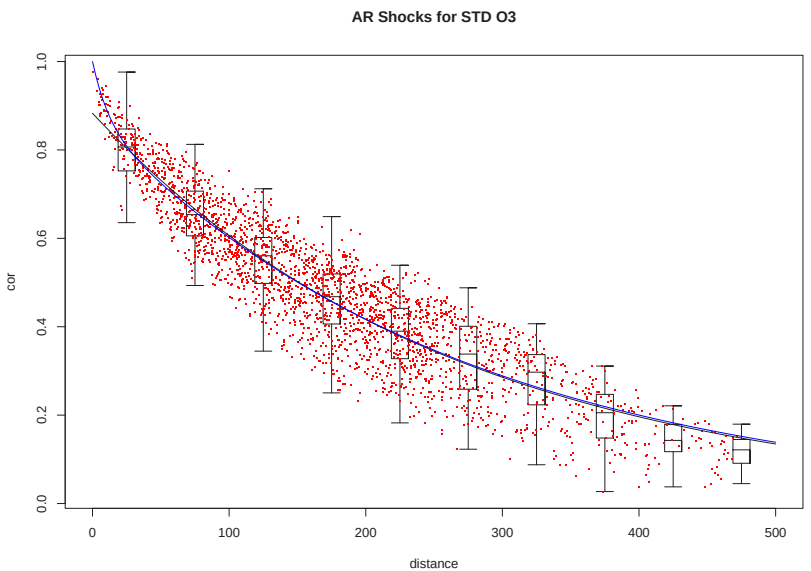


Spatial dependence

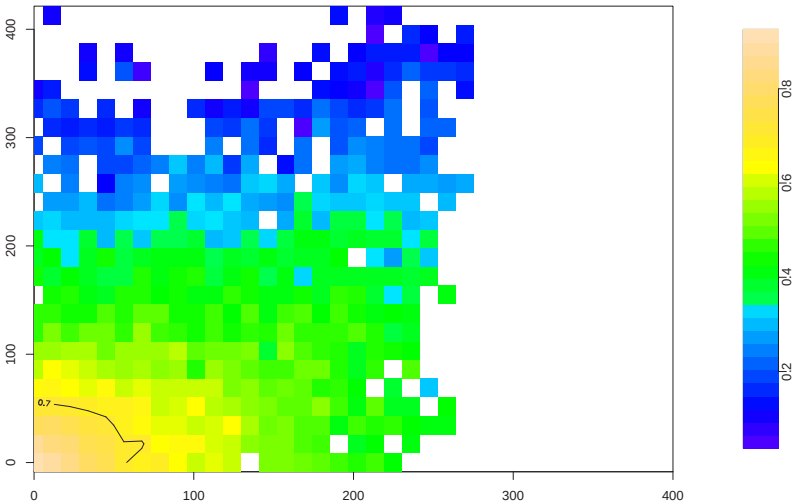
Correlogram of shocks suggests a mixture of exponential covariances

$$k(d) = \alpha e^{-d/\theta_1} + (1-\alpha)e^{-d/\theta_2}$$

with $\alpha = .09$**,** $\theta_1 = 18$ **(miles) and** $\theta_2 = 270$ **(miles)**



Anisotropy?



Inference for FHDA where we don't measure it.

First discretize this problem
y_t daily ozone values on a grid and including the stations locations.

$$\boldsymbol{o} = \begin{pmatrix} \boldsymbol{y}^s \\ \boldsymbol{y}^g \end{pmatrix} \tag{1}$$

Generating one year

Starting with an initial field: $\boldsymbol{y}_0$

1. ***spatial shock*** **sample from** $[\boldsymbol{e}_t^g | e_t^s]$
2. ***propagate*** $\boldsymbol{u}_t = \boldsymbol{\rho u}_{t-1} +$ **conditional shocks**
3. ***back transform*** $\boldsymbol{y}_t = \boldsymbol{u}_t \boldsymbol{\sigma} + \boldsymbol{\mu}$

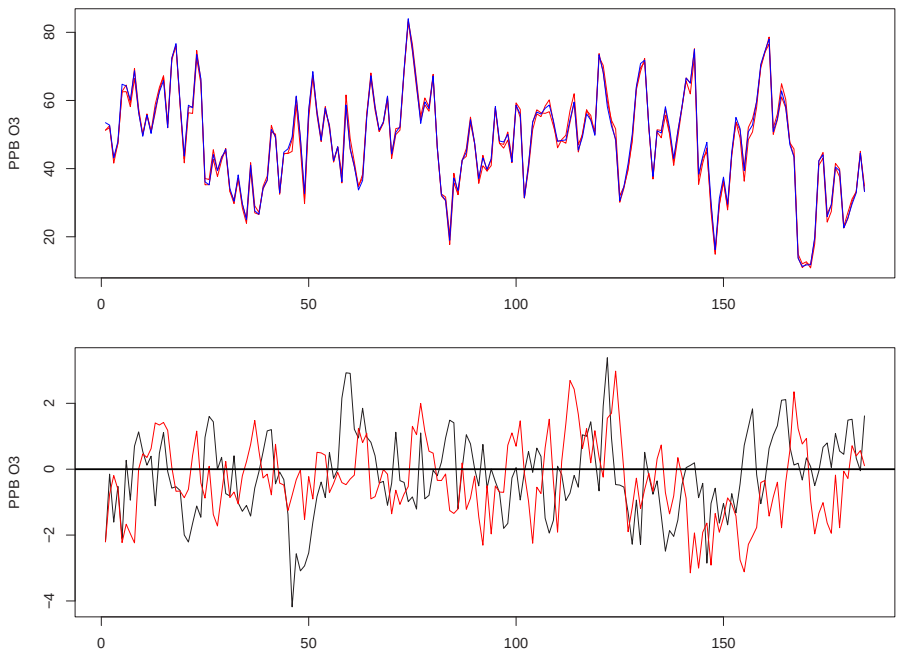
Repeat for entire season.

For each location compute FHDA from the seasons results.

Do this a "1000" times.

Two samples from conditional dist. near RTP, NC

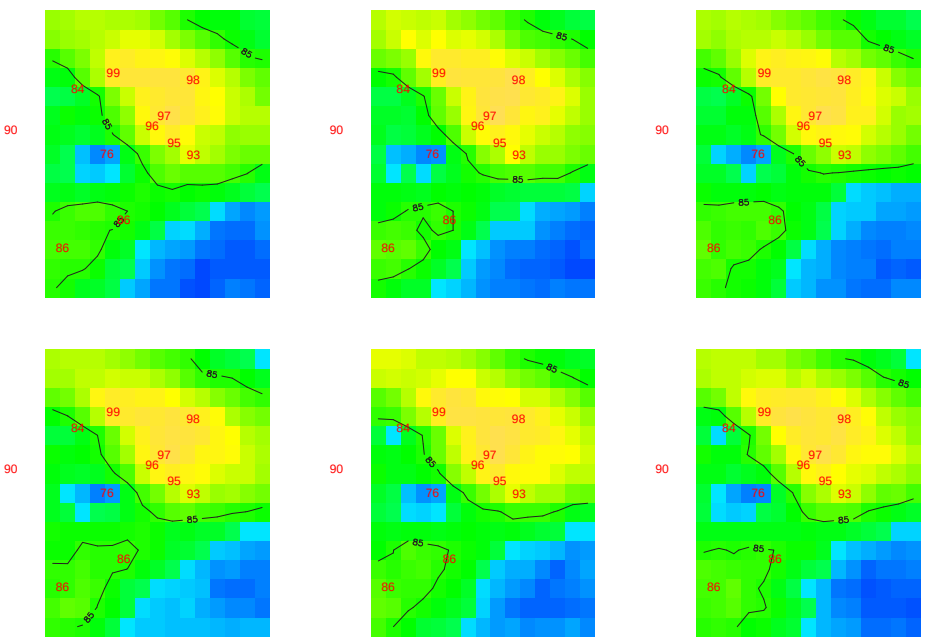
***Top*: Mean (blue) samples (red)**



***Below:* Differences from mean**

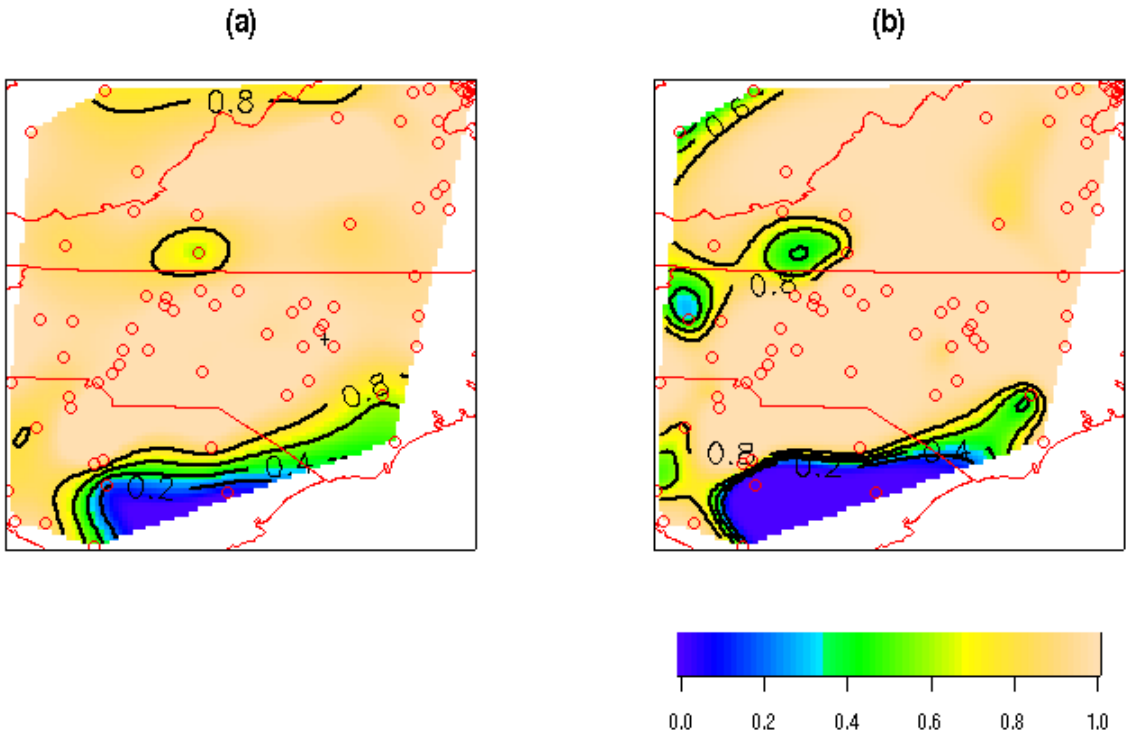
Inference/Posterior

Repeat simulations of year to accumulate a distribution of FHDA



Probability of exceeding the ozone standard

(a) Analysis from space-time model (b) Using GPD



A modest *spatial* model for ozone extremes

Predict extremes in ozone at locations where there are not monitoring stations.

Using the GPD model this is can be done by predicting σ and ξ at arbitrary locations.

A hierarchical spatial model

Observation model:

$y(\boldsymbol{x}, t)$ **surface ozone at location** $\boldsymbol{x}$ **and time** t**.**

$$[y(\boldsymbol{x}, t)|\sigma(\boldsymbol{x}), \xi(\boldsymbol{x}), u, y > u]$$

A Model for the spatial Process

$$[\sigma(\boldsymbol{x}), \xi(\boldsymbol{x}), u|\boldsymbol{\theta}]$$

Prior for hyperparameters

$$[\boldsymbol{\theta}]$$

Joint pdf for data and parameters

Assume that the extreme observations are conditionally independent.

$$\Pi_{i,t}[y(\boldsymbol{x}_i, t)|\sigma(\boldsymbol{x}), \xi(\boldsymbol{x}), u, y > u] \quad [\sigma(\boldsymbol{x}), \xi(\boldsymbol{x}), u|\boldsymbol{\theta}] \quad [\boldsymbol{\theta}].$$

t indexes days and i indexes station locations.

A Bayes analysis

For a formal Bayesian analysis, the specification of the joint is a complete recipe for inference on the parameters. Using Bayes Theorem, the posterior for $\sigma(\boldsymbol{x})$ and $\xi(\boldsymbol{x})$ given the data ($[\sigma(\boldsymbol{x}), \xi(\boldsymbol{x})|y(\mathrm{x}, t), \boldsymbol{\theta}]$) can, in principle, be computed.

Posterior mode

A useful summary of the posterior is to find the parameters that maximize the posterior. This can also be done by maximizing the joint density (because it is proportional).

Some shortcuts and assumptions

ξ, the shape, is constant over space

Justified by univariate fits.

Spatial model for $\sigma(x)$

Assume that $\sigma(\boldsymbol{x})$ is a Gaussian process with an isotropic Matern covariance function.

Parameters for the Matern

Fix the smoothness of the Matern at $\nu = 2$ and let the range be very large. The only remaining parameter λ is the sill, or variance of the process.

This is a Bayesian description of a thin plate spline model.

More about $\sigma(\boldsymbol{x})$

λ is the only hyperparameter and we just put an uninformative prior on it.

$$\sigma(\boldsymbol{x}) = P(\boldsymbol{x}) + e(\boldsymbol{x})$$

where P is a linear function of space and e is a smooth process.
As $\lambda \to \infty$ the surface will tend toward just the linear function.

As $\lambda \to 0$ the posterior surface will fit the data more closely.

log of the joint distribution

$X\beta$ **is a linear function over space.**

$$\sum_{i=1}^{M} l_{GPD}(\boldsymbol{Y}_i, \sigma(\boldsymbol{x}_i), \xi) -$$

log of the joint distribution

$X\beta$ **is a linear function over space.**

$$\sum_{i=1}^{M} l_{GPD}(\boldsymbol{Y}_i, \sigma(\boldsymbol{x}_i), \xi) -$$

$$\lambda(\boldsymbol{\sigma} - \boldsymbol{X\beta})^T (K^{-1})(\boldsymbol{\sigma X\beta})/2 - \log(|\lambda K|)$$

log of the joint distribution

$X\beta$ is a linear function over space.

$$\sum_{i=1}^{M} l_{GPD}(\boldsymbol{Y}_i, \sigma(\boldsymbol{x}_i), \xi) -$$

$$\lambda(\boldsymbol{\sigma} - \boldsymbol{X\beta})^T (K^{-1})(\boldsymbol{\sigma X\beta})/2 - \log(|\lambda K|)$$

$$+C.$$

log of the joint distribution

$X\beta$ is a linear function over space.

$$\sum_{i=1}^{M} l_{GPD}(\boldsymbol{Y}_i, \sigma(\boldsymbol{x}_i), \xi) -$$

$$\lambda(\boldsymbol{\sigma} - \boldsymbol{X\beta})^T (K^{-1})(\boldsymbol{\sigma X\beta})/2 - \log(|\lambda K|)$$

$$+C.$$

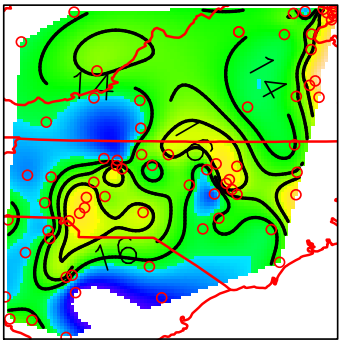
K is the thin-plate spline like covariance for the prior on σ. at the observations.

This is also a penalized likelihood.

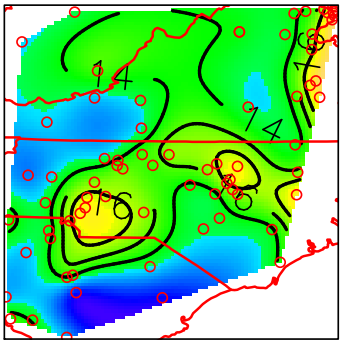
The penalty on σ is due to the covariance and the smoothing parameter, λ.

The profile likelihood surfaces for σ

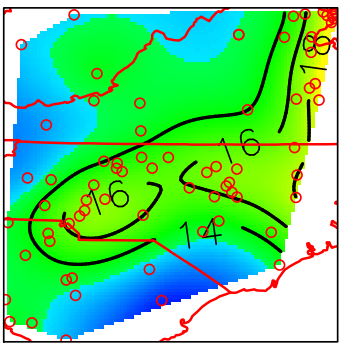
(a) lambda= 0



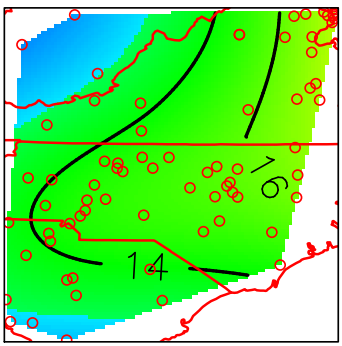
(b) lambda= 1e−6



(c) lambda= 1e−4

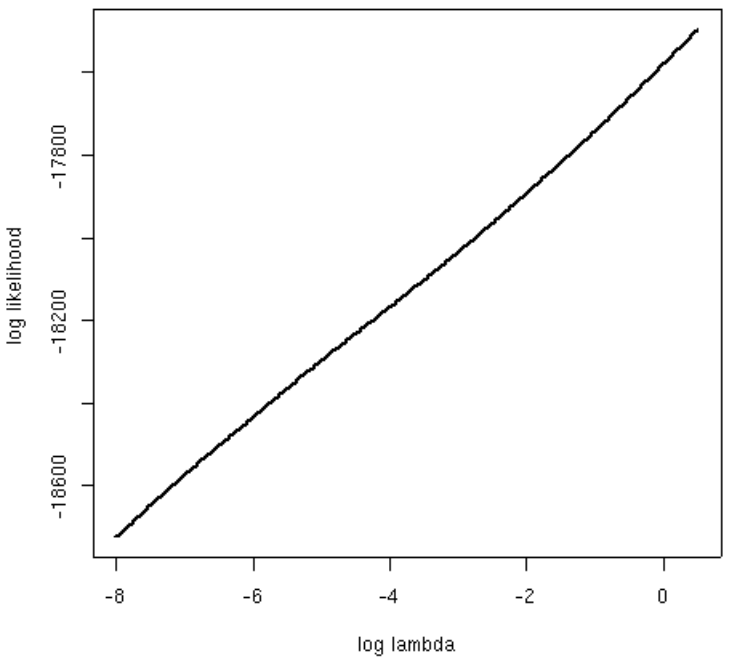


(d) lambda= 1e−2



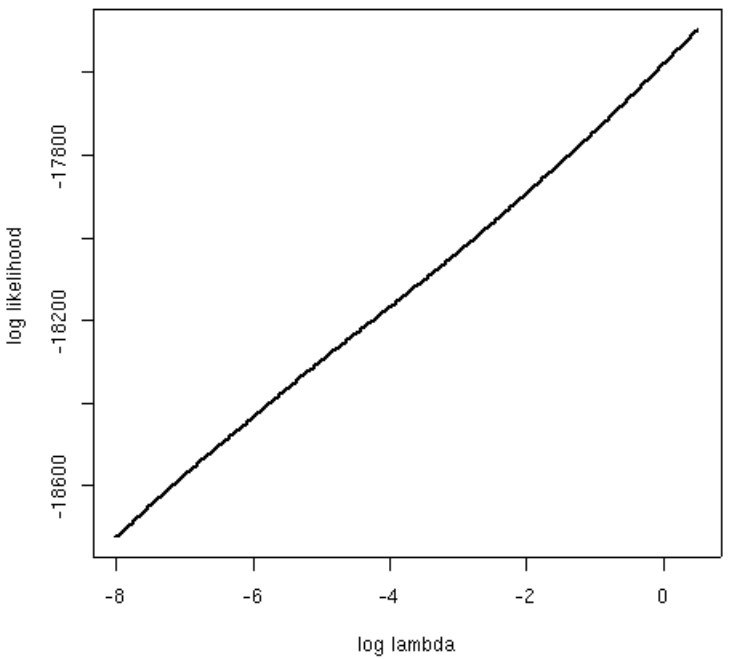
Inference for λ

Profile likelihood for λ



Inference for λ

Profile likelihood for λ



Oops!

Another try

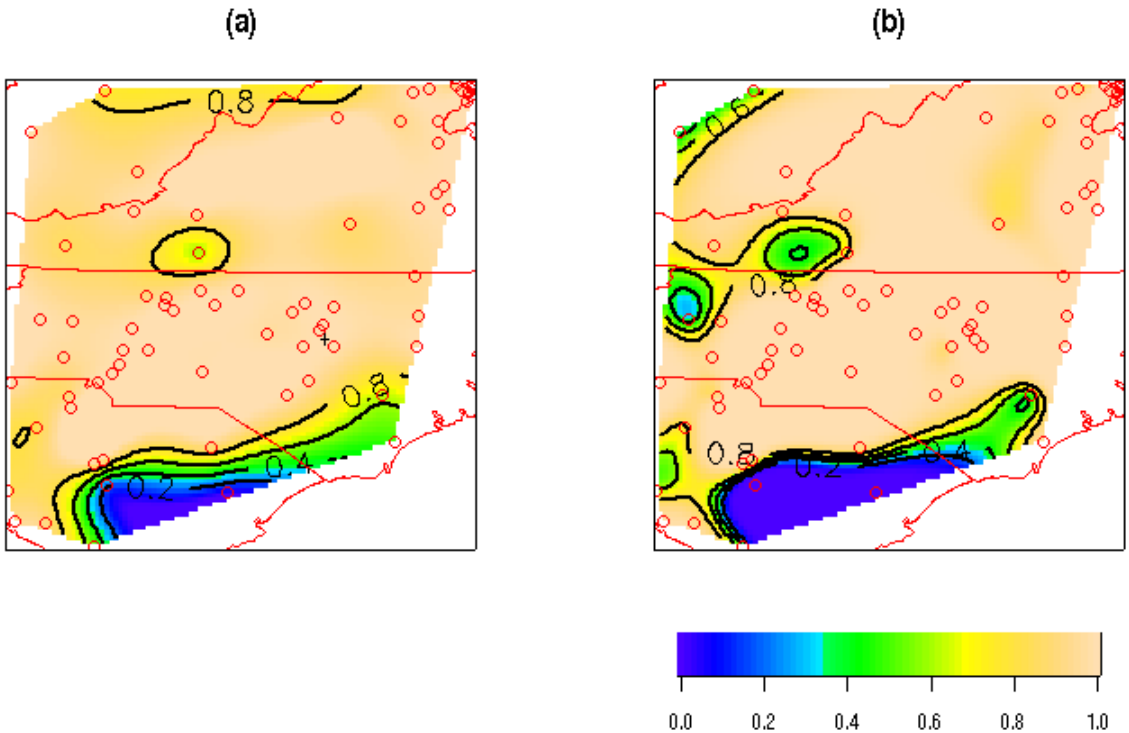
Fit a thin-plate spline to the MLE from the univariate fitting.
Determine λ by cross-validation:

- **For fixed λ omit each location, fit a spline to the remaining points**
- **predict the value at the omitted location.**
- **Compare this to the observed value. (difference is called the cross-validated residual)**
- **Choose the value of λ that minimizes the sum of squares of the CV residuals**

When this is done one gets a picture like (c).

Probability of exceeding the ozone standard

(a) Analysis from space-time model (b) Using GPD



Discussion

- Extreme value distributions offer an alternative to modeling tails of a distribution.
- Using two very different approaches the ozone case study ends with a surprisingly similar analysis.
- A more elaborate model would be possible by including more data, both over space and time. Some possible extensions: nonstationary covariances, link function for the GPD parameters and including some temporal dependence.