Extremes and Atmospheric Data

Eric Gilleland
Research Applications Laboratory
National Center for Atmospheric Research

2007-08 Program on Risk Analysis, Extreme Events and Decision Theory,
opening workshop 16-19 September, North Carolina

Extremes

- Interest in making inferences about large, rare, extreme phenomena.
- Given certain properties (e.g., independence), extremes follow EVD.
- Tricky because atmospheric data are often spatially (and temporally) correlated, even in their extreme values.

Atmospheric Data

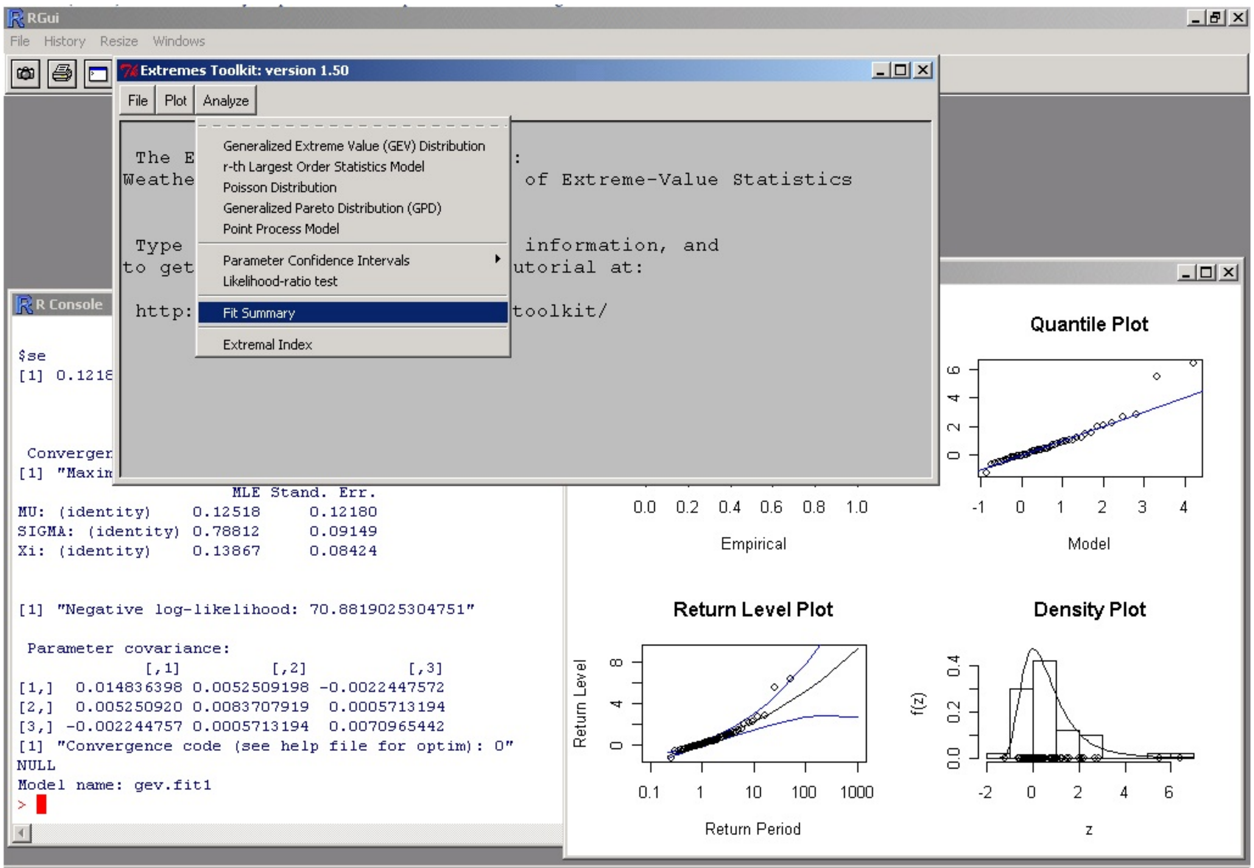
- Point Data
  - Air Quality Monitoring (e.g., Ozone concentrations)
  - METARs stations (ground-based weather observations)
  - Radiosondes (weather balloons, airplanes)
  - ...
- Gridded Data
  - Global NCAR/NCEP Reanalysis
    All available observational data are synthesized with a static data assimilation process.
  - Remote sensing (e.g., satellite)
  - Model Output (e.g., Weather/Climate models)

Software

- R:
  - statistical programming language
    http://www.r-project.org
  - Free software environment for statistical computing and graphics.
  - Compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

Software

- extRemes:
  - Weather and Climate Applications of Extreme Value Statistics



Software

- spatial extension to R package `extRemes` (*in dev*)
- Many other extreme-value software packages (e.g., Stephenson and Gilleland, 2005)
- Many spatial statistics package too (e.g., in R: `fields`, `sp`, ...)

Questions

- Air Quality Standards

  High values of a spatial process (e.g., "... if the three-year average of the fourth-highest maximum daily 8-hour ozone concentration exceeds 80 ppb ...")

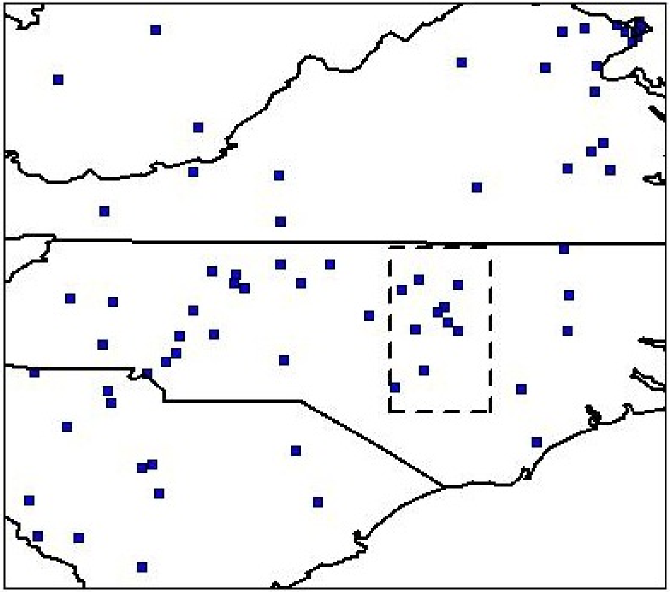
  What is the coverage at an unobserved location?

- Severe Weather and Climate Change

  Will the frequency and intensity of severe weather increase in the future?

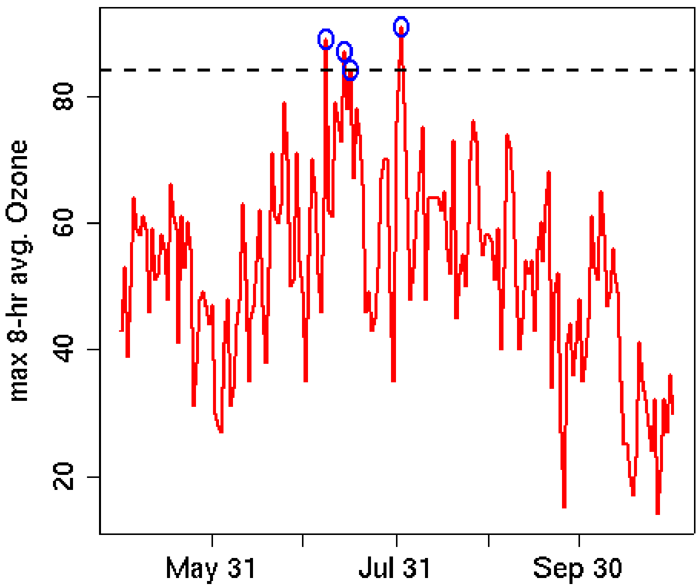
Daily Ozone Concentrations

- Number of years: 5 ozone "seasons" (1995 - 1999)
- Ozone "season": 184 days (April - October)
- Daily values: Maximum of 8-hr block averages (ppb)
- Distribution: Continuous, *Normal*
- Spatial Locations: 72 stations centered on North Carolina



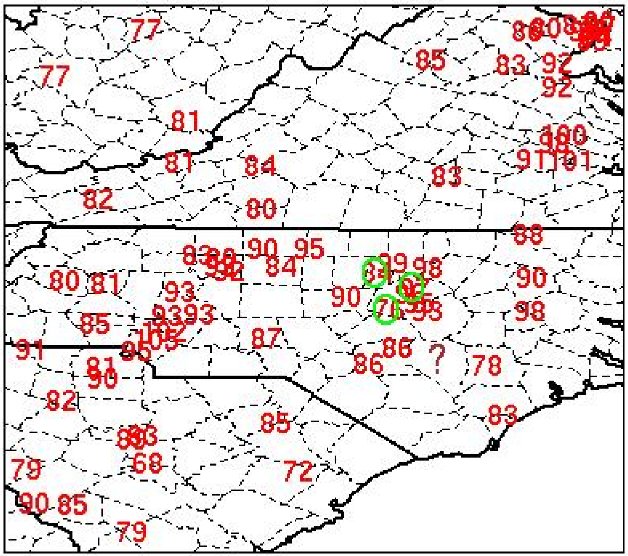
NAAQS for Ground-level Ozone

In 1997, the U.S. EPA changed the NAAQS for regulating ground-level ozone levels to one based on the *fourth-highest daily maximum 8-hr. averages (FHDA)* of an ozone season (184 days). Compliance is met when the FHDA over a three year (season) period is below 84 ppb.

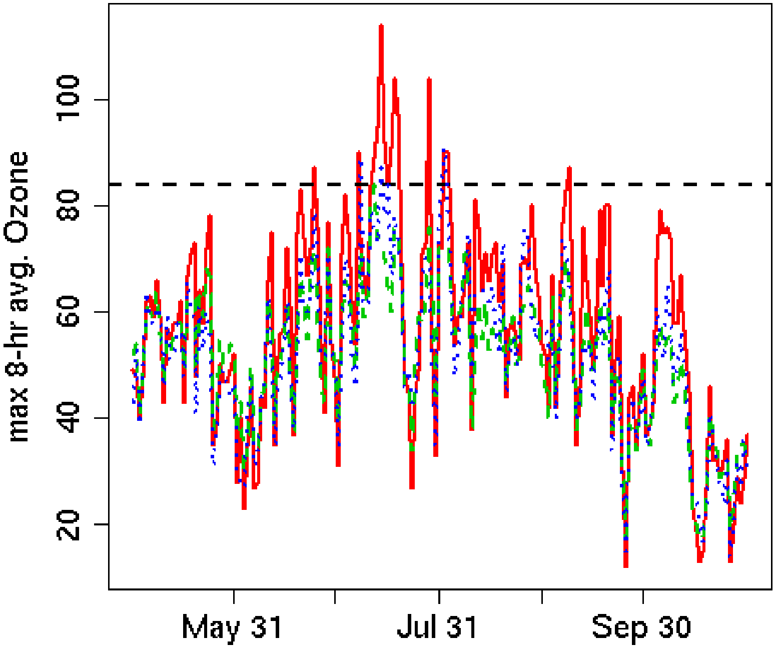


1997 FHDA Observed Data

North Carolina



Three sites



Goal

Spatial Inference for the Standard

What can be deduced about FHDA at unobserved locations?

Characterize a spatial field of fourth-highest order statistics

- Distribution: Is it Gaussian (extreme value)?
- What does the covariance structure look like?

Some Strategies

Strategies Using Spatial Extension of `extRemes`

- Spatial and Space-Time Approaches (e.g., Gilleland and Nychka, 2005)
  - Daily Model Approach
  - Seasonal Model Approach (and simple thin plate spline)
  - Comparison of Daily and Seasonal Models
- Spatial Extreme Value Model (e.g., Gilleland *et al*, 2006)

Other approaches

- Alter loss function: (e.g., Craigmile *et al.*, 2006a)
- EVD using Bayesian hierarchical model: (e.g., Cooley *et al*, *in press*, 2006a)
- Madograms: (e.g., Cooley *et al*, 2006b)

Space-Time Approach: Daily Model

Space-Time model

One strategy is to model daily ozone using a space-time model (spatial AR(1)) to determine the distribution of the standard.

Simulation

Use Monte Carlo simulations to generate samples of the standard using the space-time model. That is, simulate a season of ozone data using the space-time model and take the fourth-highest of them. Do this several times to obtain a sample from the FHDA distribution.

Space-Time Approach: Daily Model

The Daily Model

Let $Y(\mathbf{x}, t)$ denote the daily 8-hr max ozone for m sites over n time points. Consider,

$$Y(\mathbf{x}, t) = \mu(\mathbf{x}, t) + \sigma(\mathbf{x})u(\mathbf{x}, t),$$

where $u(\mathbf{x}, t)$ is a de-seasonalized zero mean, unit variance space-time process.

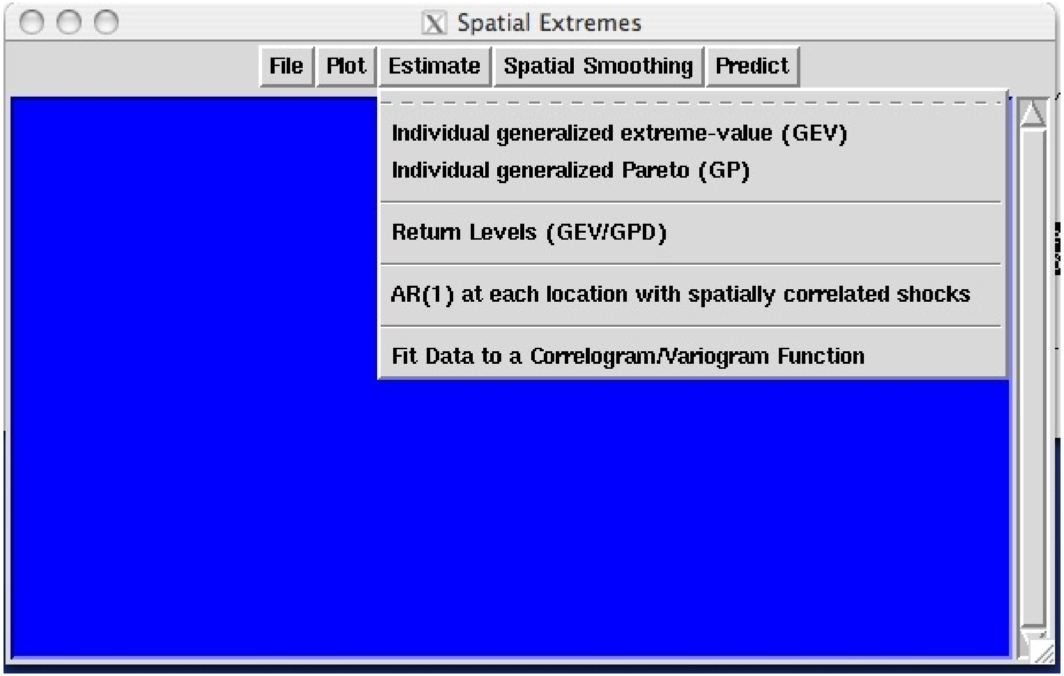
Explicitly modeling daily observations spatially in order to obtain simulated samples of FHDA

Space-Time Approach: Daily Model

Space-Time Process

After de-seasonalizing/standardizing, for each spatial site, $\mathbf{x}$, fit an AR(1) model over time for $u(\mathbf{x}, t)$.

$$u(\mathbf{x}, t) = \rho(\mathbf{x})u(\mathbf{x}, t-1) + \varepsilon(\mathbf{x}, t)$$



Space-Time Approach: Daily Model

Leads to the spatio-temporal covariance

$\text{Cov}(u(\mathbf{x},t),u(\mathbf{x}',t-\tau)) =$

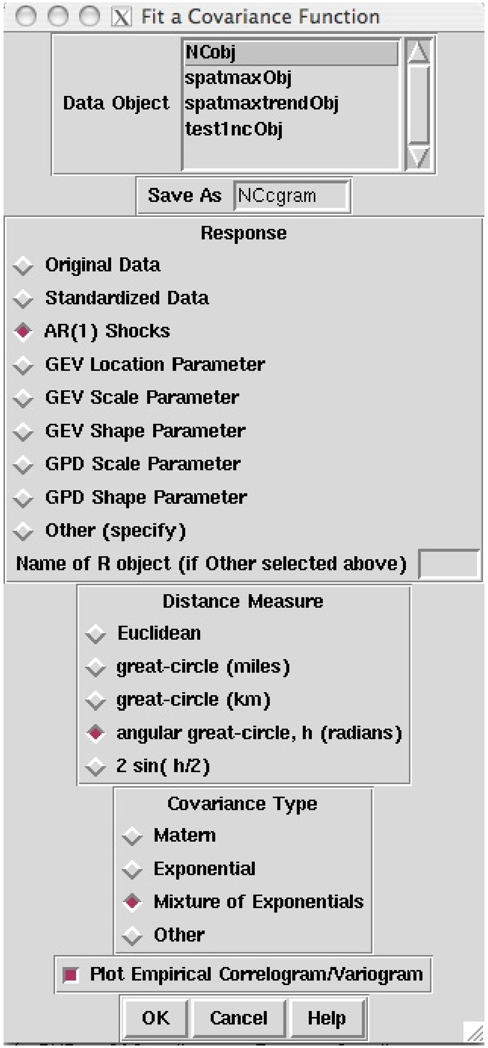
$$\frac{(\rho(\mathbf{x}))^{\tau}\sqrt{1-\rho^2(\mathbf{x})}\sqrt{1-\rho^2(\mathbf{x}')}}{1-\rho(\mathbf{x})\rho(\mathbf{x}')}\cdot\psi(d(\mathbf{x},\mathbf{x}')),$$

for $\tau \geq 0$.

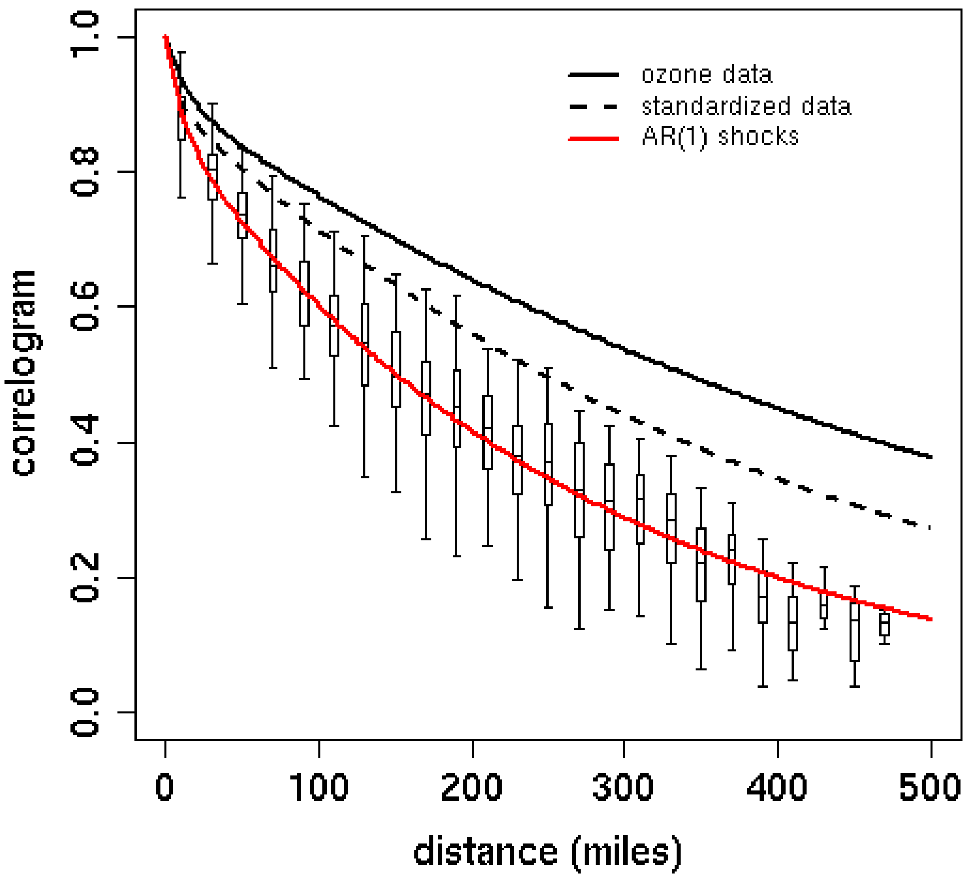
If $\rho(\mathbf{x}) = \rho$, then $\text{Cov}(u(\mathbf{x},t),u(\mathbf{x}',t-\tau))$ simplifies to

$$\rho^{\tau}\cdot\psi(d(\mathbf{x},\mathbf{x}'))$$

Space-Time Approach: Daily Model



Correlogram of AR(1) shocks, $\hat{\varepsilon}(\mathbf{x}, t)$



Space-Time Approach: Daily Model

Space-Time Process: Correlogram of AR(1) shocks

Correlogram plots for spatial shocks suggest a *mixture of exponentials* covariance model is appropriate.

Let $h = d(\mathbf{x}, \mathbf{x}')$,

$$\psi(h) = \alpha e^{-\mathbf{h}/\theta_1} + (1-\alpha) e^{-\mathbf{h}/\theta_2}$$

Fitted values are $\widehat{\alpha} \approx 0.13$ (± 0.02), $\widehat{\theta}_1 \approx 11$ miles (± 3.37 miles) and $\widehat{\theta}_2 \approx 272$ miles (± 16.89 miles).

Space-Time Approach: Daily Model

The Goal

Spatial inference for the NAAQS for ground-level ozone.

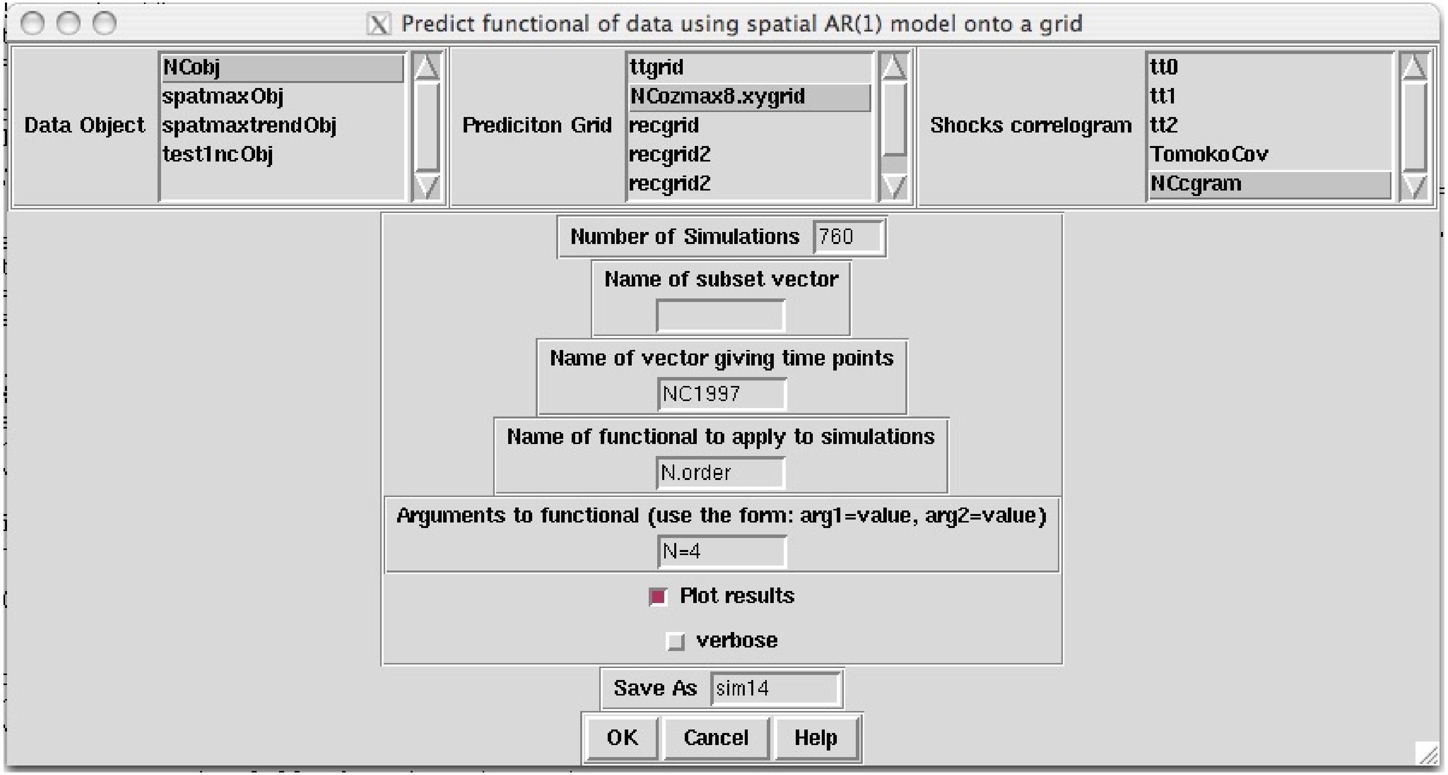
That is, what can be deduced about FHDA at unobserved locations?

Want to sample from [FHDA|daily data].

Space-Time Approach: Daily Model

Algorithm to predict FHDA at unobserved location, $\mathbf{x}_0$.

1. Simulate data for an entire ozone season



Space-Time Approach: Daily Model

Algorithm to predict FHDA at unobserved location, $\mathbf{x}_0$.

1. Simulate data for an entire ozone season

   (a) Interpolate spatially from $u(\mathbf{x}, 1)$ to get $\widehat{u}(\mathbf{x}_0, 1)$.

   (b) Also interpolate spatially to get $\widehat{\rho}(\mathbf{x}_0)$, $\widehat{\mu}(\mathbf{x}_0, \cdot)$ and $\widehat{\sigma}(\mathbf{x}_0)$.

   (c) Sample shocks at time t from $[\varepsilon(\mathbf{x}_0, t) | \varepsilon(\mathbf{x}, t)]$.

   (d) Propagate AR(1) model.

   (e) Back transform $\widehat{Y}(\mathbf{x}_0, t) = \widehat{u}(\mathbf{x}_0, t)\widehat{\sigma}(\mathbf{x}_0) + \widehat{\mu}(\mathbf{x}_0, t)$

Space-Time Approach: Daily Model

Algorithm to predict FHDA at unobserved location, $\mathbf{x}_0$.

1. Simulate data for an entire ozone season.
2. Take fourth-highest value from Step 1.
3. Repeat Steps 1 and 2 many times to get a sample of FHDA at unobserved location.

Space-Time Approach: Daily Model

Distribution for the AR(1) shocks

$[\varepsilon(\mathbf{x}_0, t)|\varepsilon(\mathbf{x}, t)]$ (Step 1c) given by

$$\text{Gau}(\mathbf{M}, \boldsymbol{\Sigma})$$

with

$$\mathbf{M} = \mathbf{k}'(\mathbf{x}_0, \mathbf{x})\mathbf{k}^{-1}(\mathbf{x}, \mathbf{x})\varepsilon(\mathbf{x}, t)$$

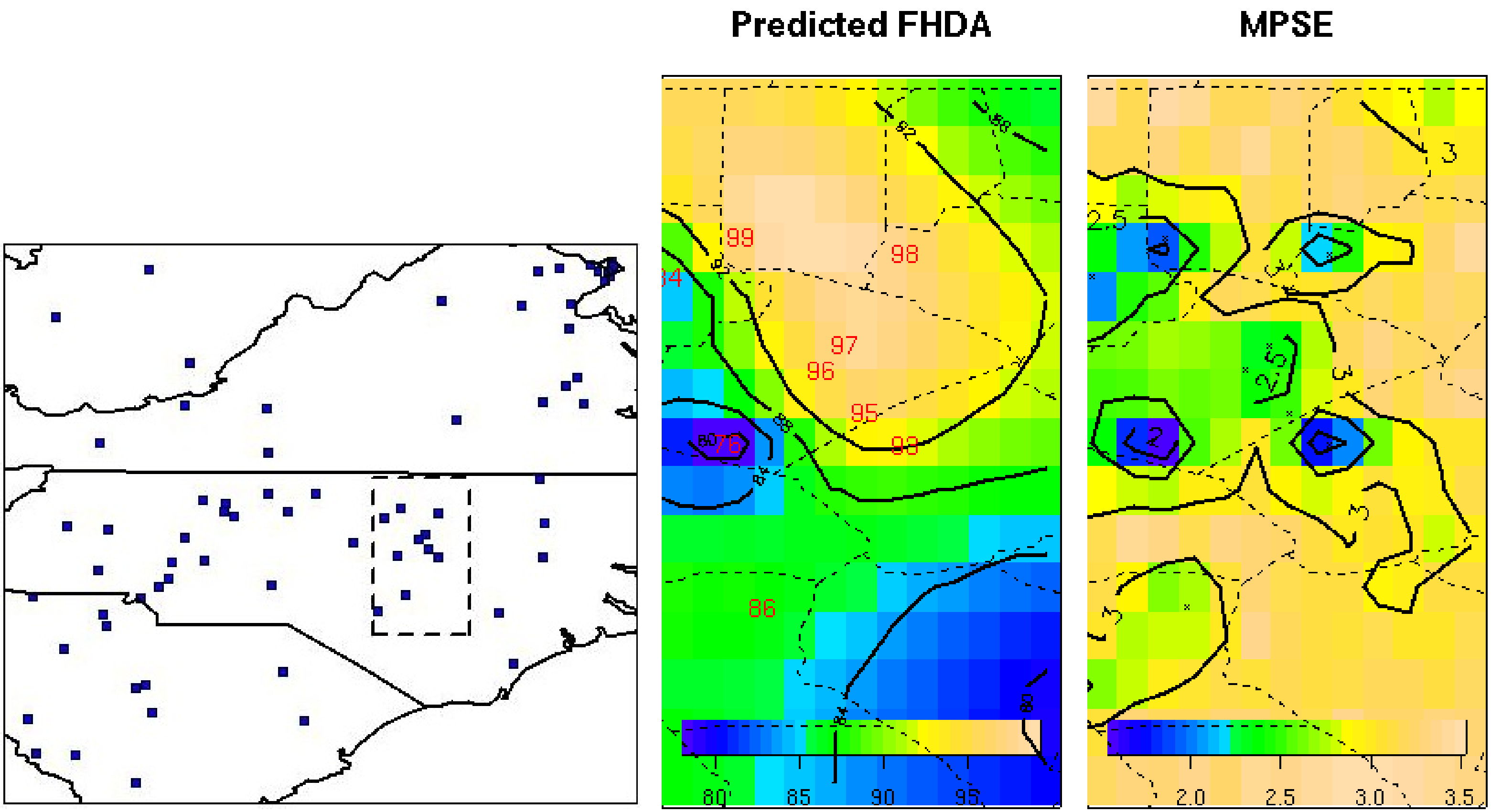
and

$$\boldsymbol{\Sigma} = \mathbf{k}'(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{k}'(\mathbf{x}_0, \mathbf{x})\mathbf{k}^{-1}(\mathbf{x}, \mathbf{x})\mathbf{k}(\mathbf{x}, \mathbf{x}_0),$$

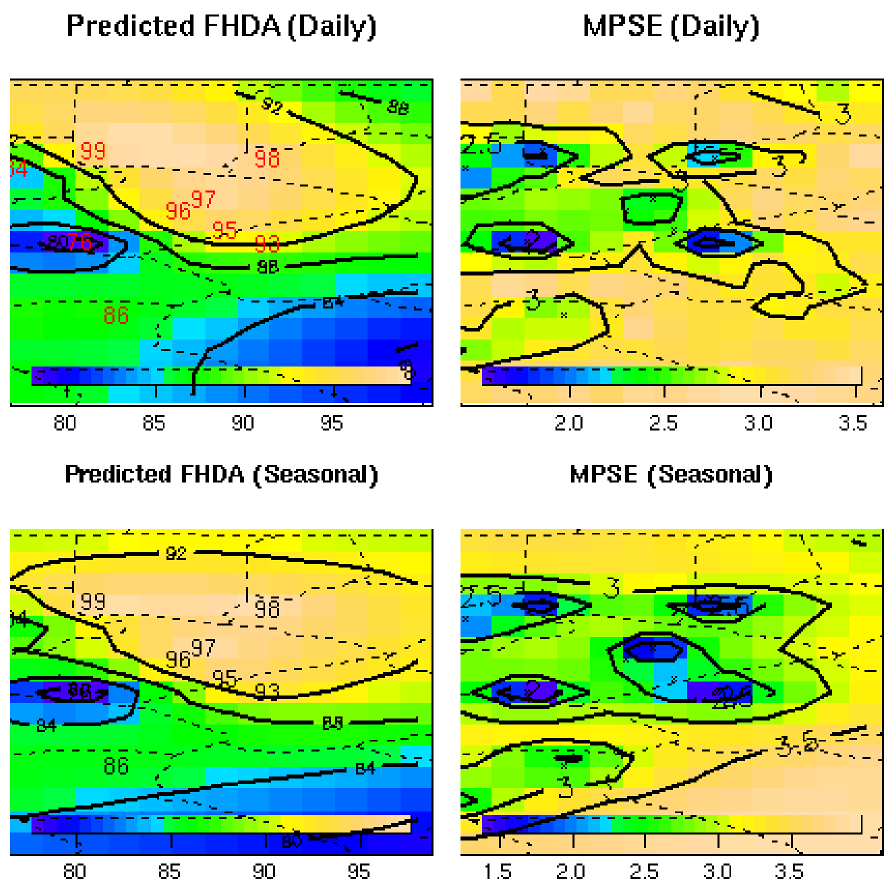
where $\mathbf{k}(\mathbf{x}, \mathbf{y}) = [\psi(\mathbf{x}_i, \mathbf{y}_j)]$ the covariance matrix for two sets of spatial locations.

Space-Time Approach: Daily Model

Results of predicting FHDA spatially with daily model (1997)



Comparing the Daily and Seasonal models



Comparing the Daily and Seasonal models

- Simplicity of the seasonal model approach is desirable.
- Daily model yields consistently lower MSE from leave-one-out cross validation.
- Daily model can account for "complicated" spatial features without resorting to non-standard techniques.
- Daily MPSE is consistently too optimistic.

Another Approach: Spatial Extremes

Given a spatial process, $Z(\mathbf{x})$, what can be said about

$$\Pr\{Z(\mathbf{x}) > z\}$$

when z is large?

Spatial Extremes

Given a spatial process, $Z(\mathbf{x})$, what can be said about

$$\Pr\{Z(\mathbf{x}) > z\}$$

when z is large?

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

Spatial Extremes

Given a spatial process, $Z(\mathbf{x})$, what can be said about

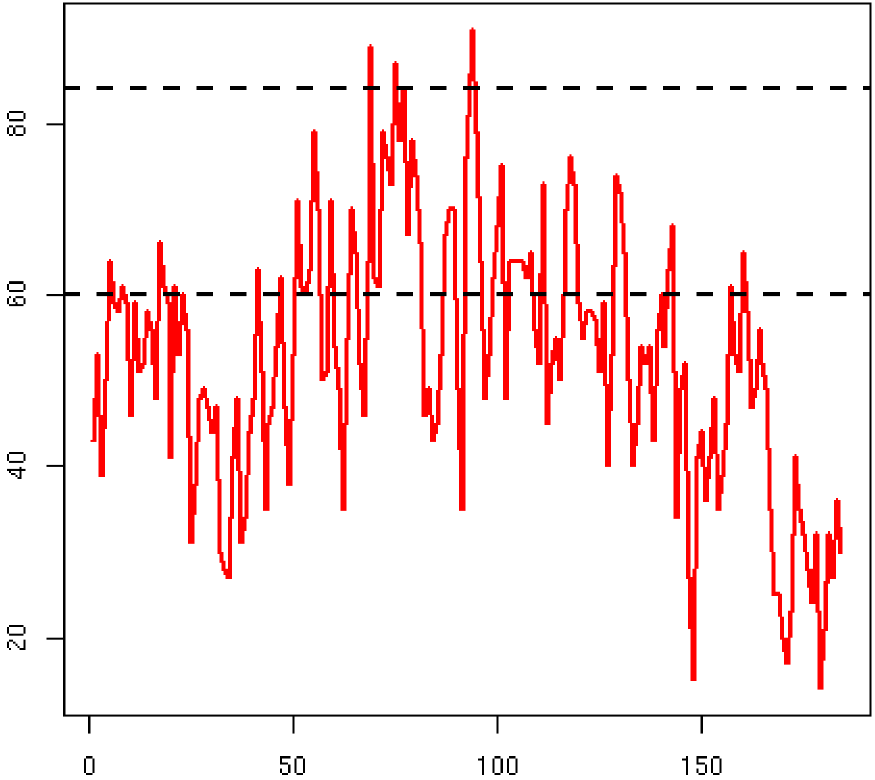
$$\Pr\{Z(\mathbf{x}) > z\}$$

when z is large?

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

Spatial structure on parameters of distribution (not FHDA).

Extreme Value Distributions: GPD



For a (large) threshold u, the GPD is given by

$$\Pr\{X > x | X > u\} \approx [1 + \frac{\xi}{\sigma}(x-u)]^{-1/\xi}$$

A Hierarchical Spatial Model

Observation Model:

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

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

Spatial Process Model:

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

Prior for hyperparameters:

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

A Hierarchical Spatial Model

Assume extreme observations to be *conditionally independent* so that the joint pdf for the data and parameters is

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

t indexes time and i stations.

Shortcuts and Assumptions

- Assume threshold, u, fixed.
- $\xi(\mathbf{x}) = \xi$ (i.e., shape is constant over space). *Justified by univariate fits.*
- Assume $\sigma(\mathbf{x})$ is a Gaussian process with isotropic Matérn covariance function.
- Fix Matérn smoothness parameter at $\nu = 2$, and let the range be very large–leaving only λ (ratio of variances of nugget and sill).

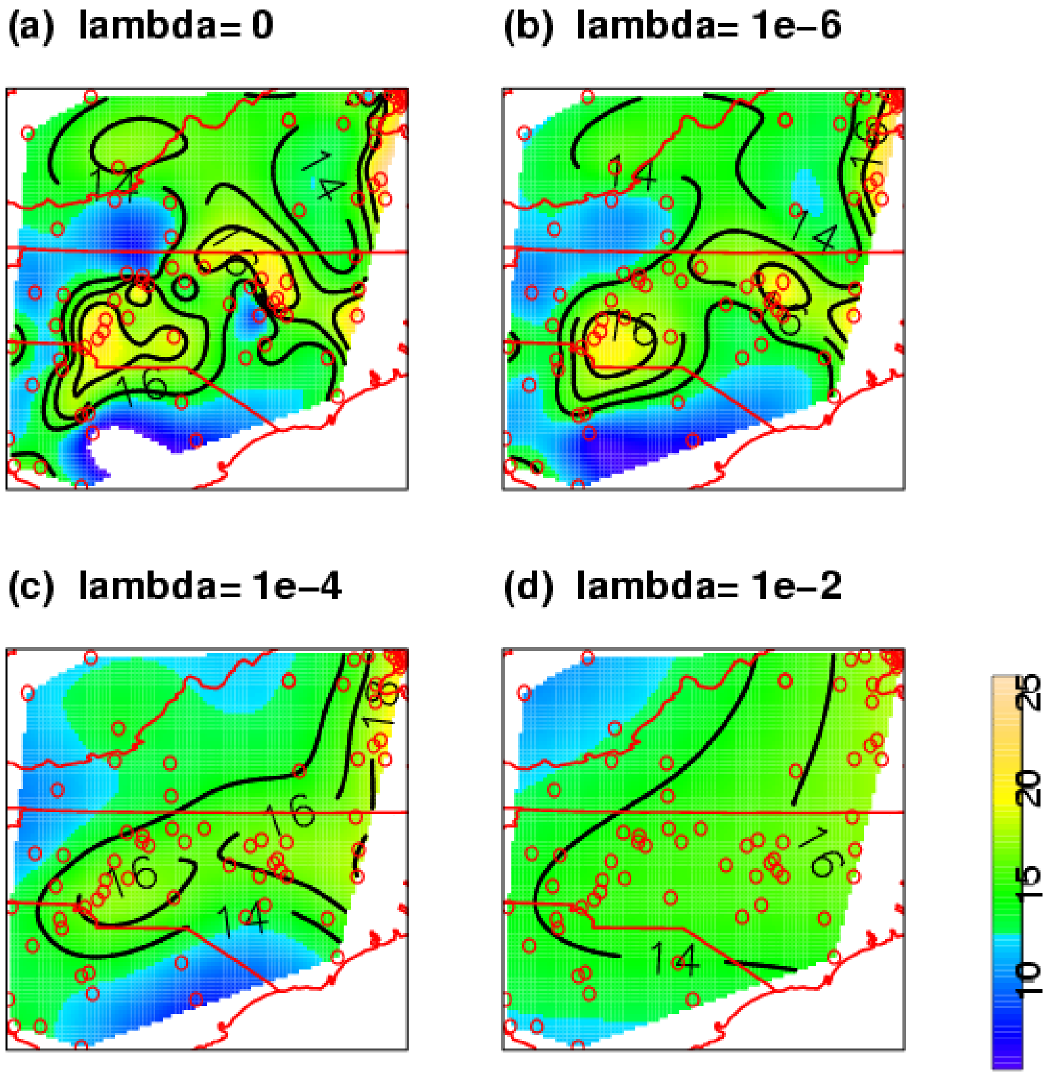
More on $\sigma(\mathbf{x})$

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

with P a linear function of space, e a smooth spatial process, and η white noise (nugget).

λ is the only hyper-parameter

- As $\lambda \longrightarrow \infty$, the posterior surface tends toward just the linear function.
- As $\lambda \longrightarrow 0$, the posterior surface will fit the data more closely.



(a) lambda= 0
(b) lambda= 1e−6
(c) lambda= 1e−4
(d) lambda= 1e−2
14
16
18
10
15
20
25

log of joint distribution

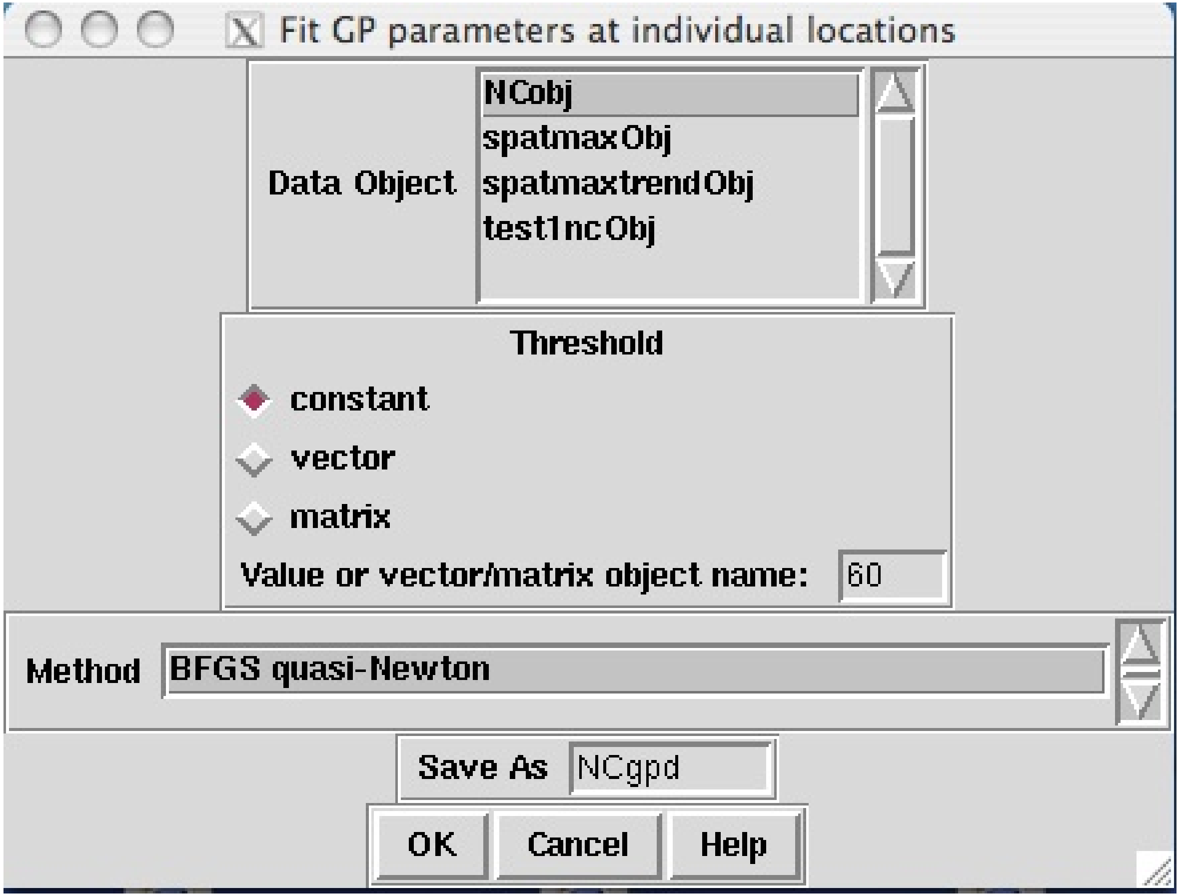
$$\sum_{i=1}^{n} \ell_{\mathsf{GPD}}(y(\mathbf{x}_i, t), \sigma(\mathbf{x}_i), \xi) -$$

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

K is the covariance for the prior on σ at the observations.

This is a penalized likelihood:
The penalty on $\boldsymbol{\sigma}$ results from the covariance and smoothing parameter λ.

Spatial extension for `extRemes`

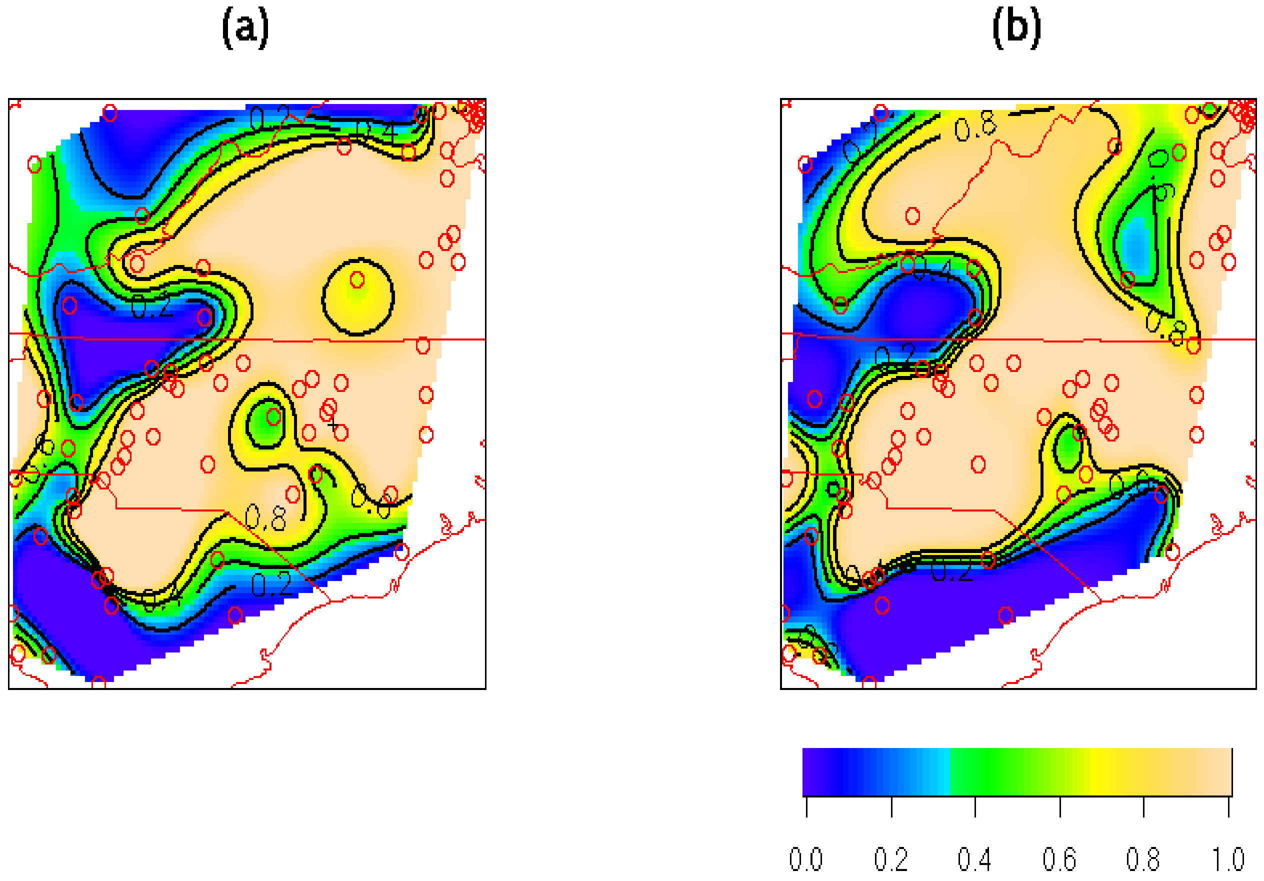


Spatial extension for extRemes

```
[1] "NCozmax8 fit to Generalised Pareto Distribution (GPD) at individual locations

(grid points)."
                   scale      shape
 N              72.000000 72.000000
 mean           16.371782 -0.291973
 Std.Dev.        2.902333  0.073854
 min            10.355406 -0.490141
 Q1             14.347936 -0.346752
 median         16.259529 -0.298820
 Q3             17.984638 -0.231408
 max            23.982931 -0.152946
 missing values  0.000000  0.000000
```

Probability of exceeding the standard



Conclusions for Ozone NAAQS

- Simplicity of the seasonal model approach is desirable.
- Daily model yields consistently lower MSE from leave-one-out cross validation.
- Daily model can account for "complicated" spatial features without resorting to non-standard techniques.
- Daily MPSE is consistently too optimistic.
- Extreme value models good alternative to modelling the tail of distributions.
- Two very different approaches yield similar results

Large-scale Indicators for Severe Weather

Severe weather typically on fine scales

Historical records limited

Weak relationship with larger-scale phenomena

CAPE (J/kg) $\times$ shear (m/s) found to be indicative of conducive environments for severe weather

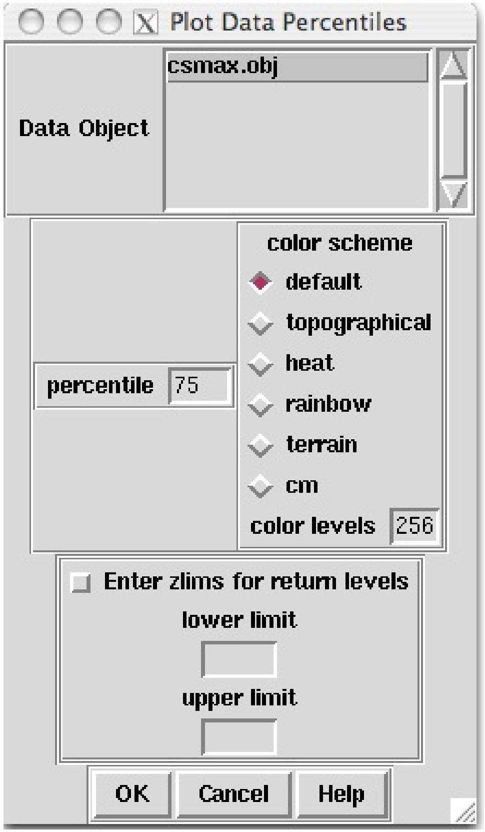
(e.g., Brooks *et al*, 2003; Pocernic *et al*, in prep)

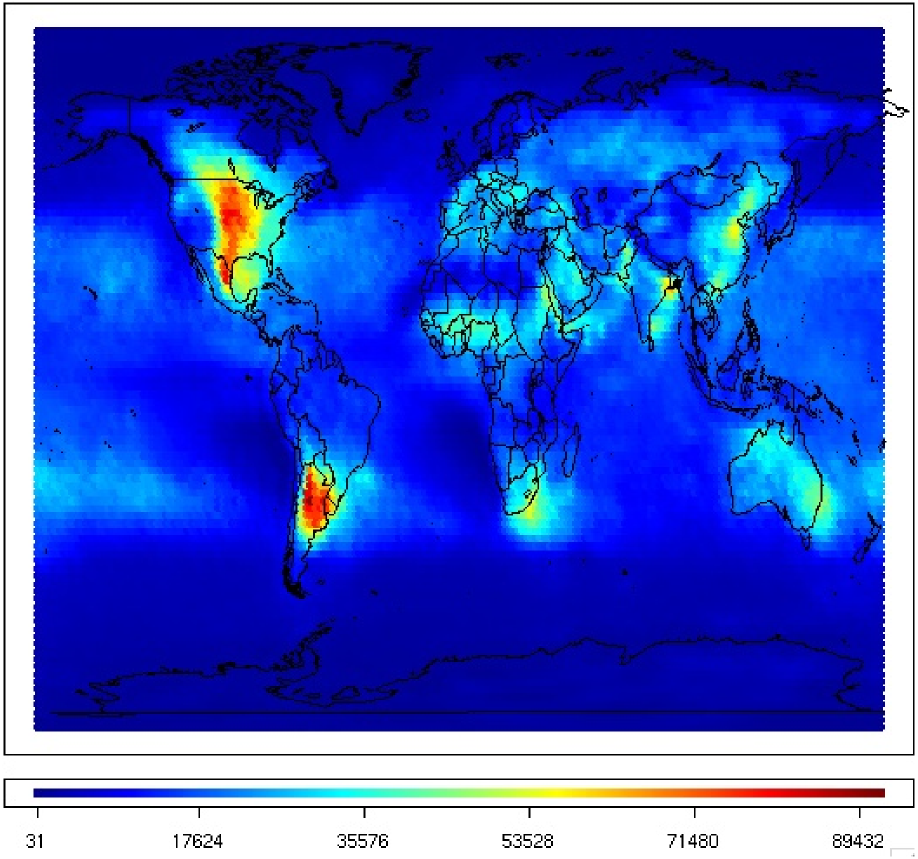
Global NCAR/NCEP Reanalysis Data

- All available observational data are synthesized with a static data assimilation process.
- Resolution $\approx 1.875^o$ longitude by 1.915^o latitude
- 17 856 grid point locations (192×94 grid)
- Temporal spacing every 6 hours
- 1958 through 1999 (42 years)
- Convective available potential energy (CAPE, J/kg)
- Magnitude of vector difference between surface and 6-km wind (shear, m/s)
- Both CAPE and shear ≥ 0 (Lots of zeros!)

Global NCAR/NCEP Reanalysis Data

Upper quartiles for (42-yr) annual maximum CAPE (J/kg) $\times$ shear (m/s)



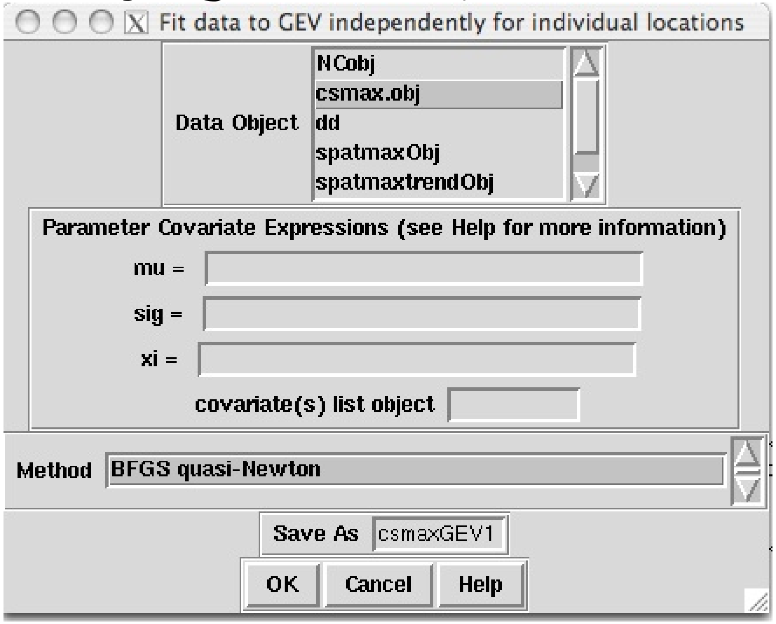


Preliminary Analysis

Extreme-value theory

- Bivariate extremes difficult because CAPE and shear tend not to be large together
- Nonstationary spatial structure

Initial analysis: Fit GEV to individual grid points without worrying about spatial structure.



Generalized Extreme-Value (GEV) distribution

The generalized Extreme Value Distribution (GEV)

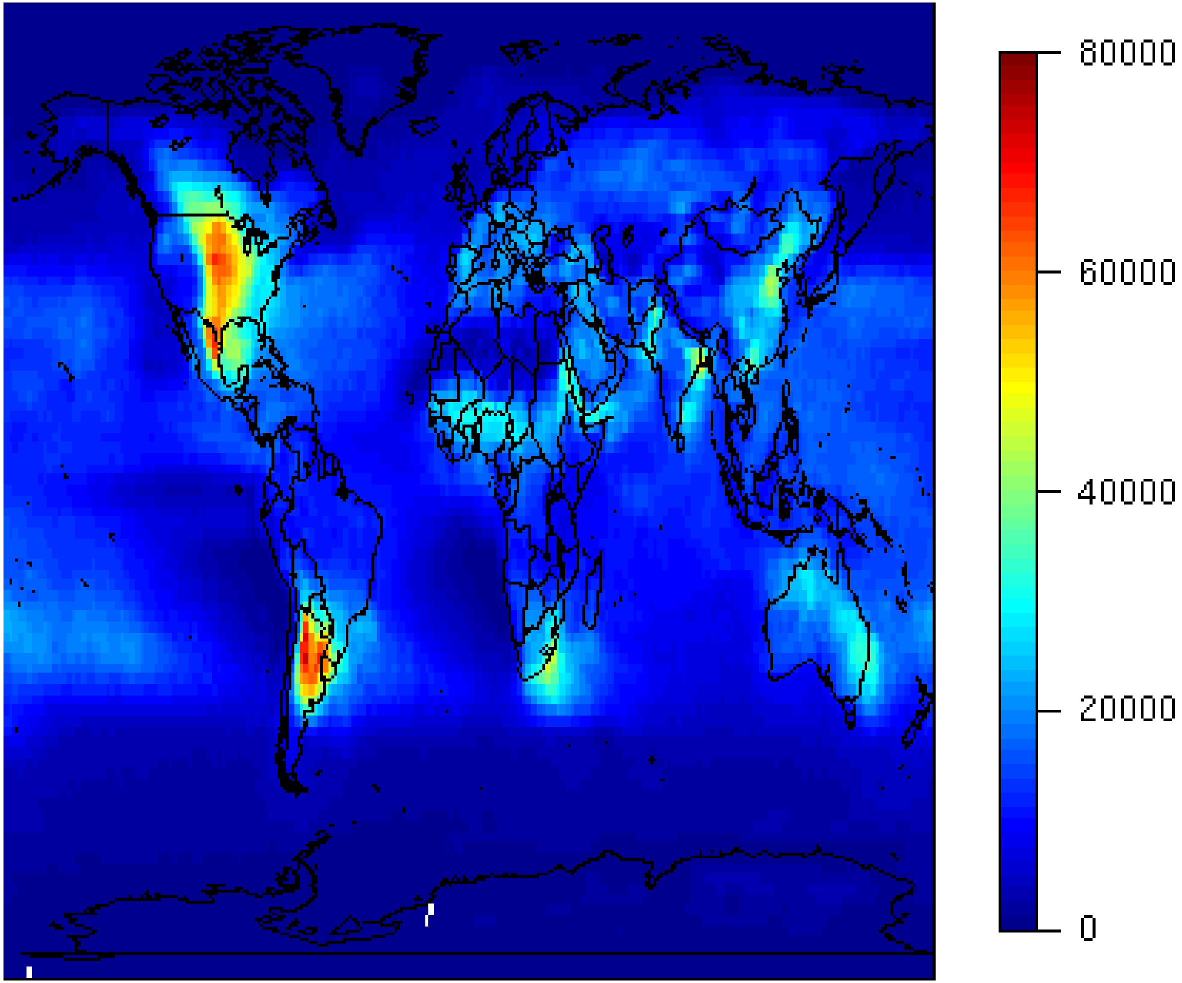
$$F_{\mathsf{GEV}}(z) = \exp\{-(1 + \frac{\xi}{\sigma}(z - \mu))_+^{-1/\xi}\}$$

Parameters: Location (μ), Scale (σ) and Shape (ξ).

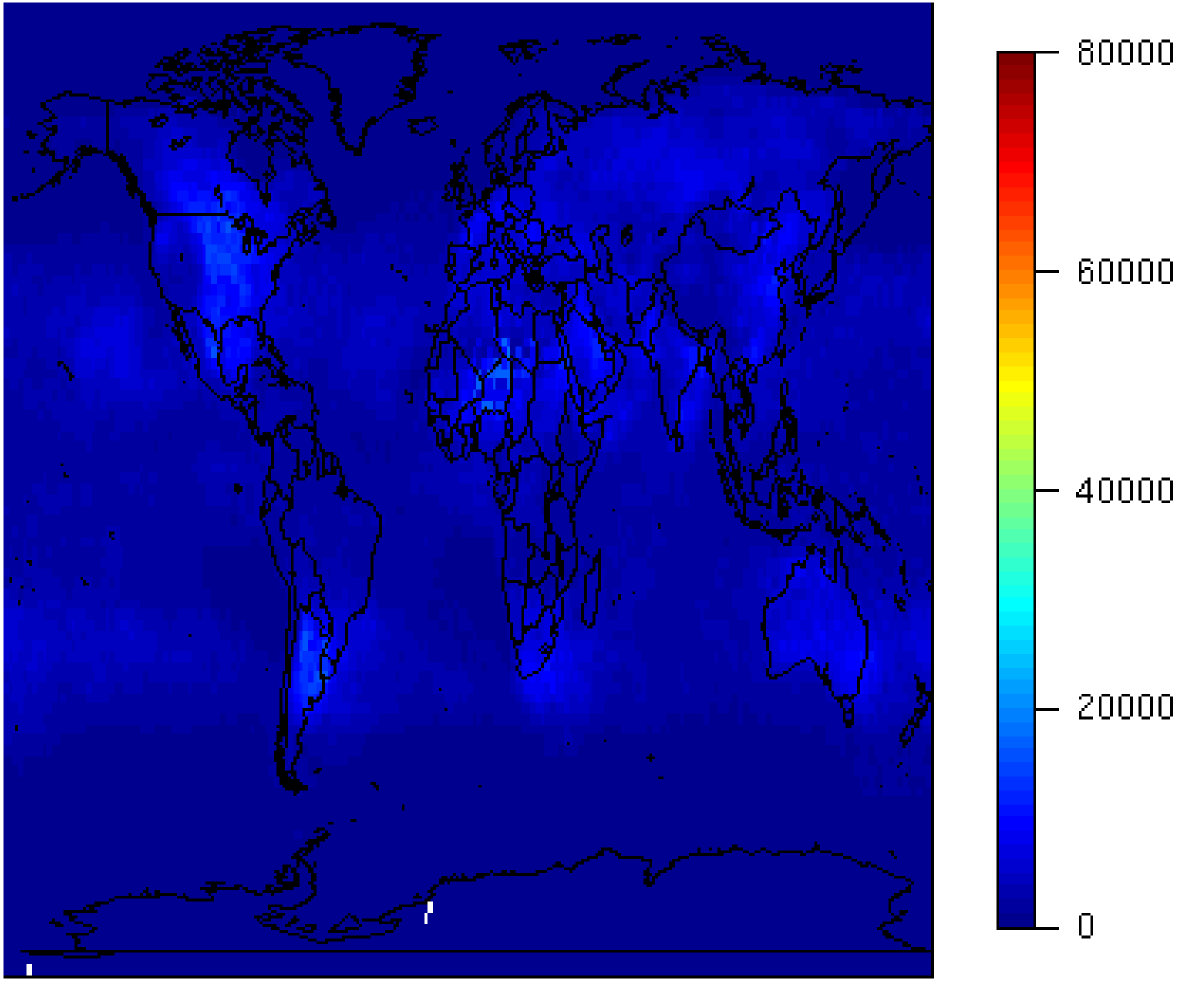
Notes about the GEV

- $\xi < 0$, Weibull, upper end-point at $\mu - \frac{\sigma}{\xi}$
- $\xi > 0$, Fréchet, lower end-point at $\mu - \frac{\sigma}{\xi}$ (heavy tail)
- $\xi = 0$, Gumbel (light tail)
- $\Pr\{\max\{X_1, \ldots, X_n\} < z\} \approx F_{\mathsf{GEV}}(z)$

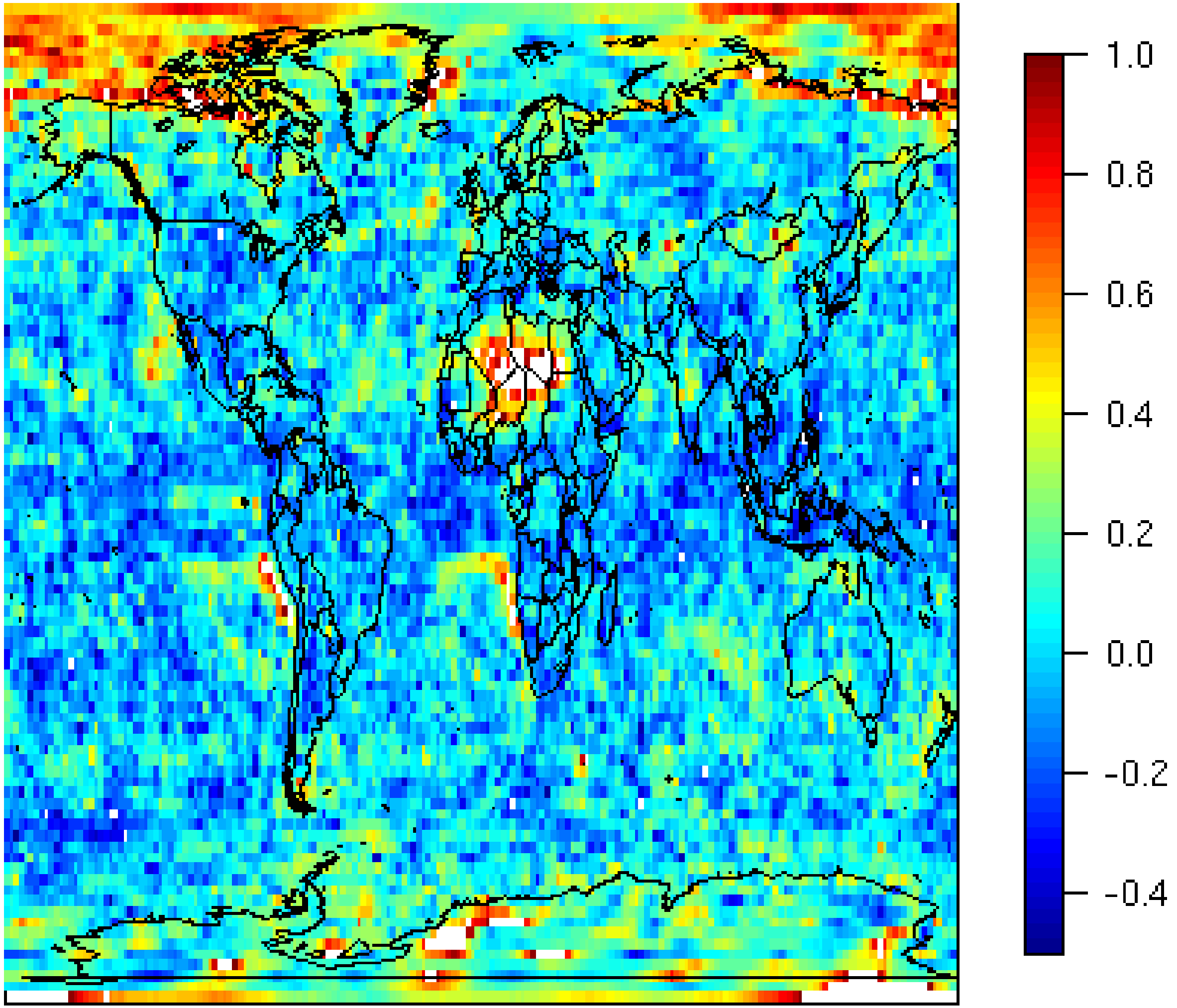
Preliminary Analysis: Location Parameter



Preliminary Analysis: Scale Parameter

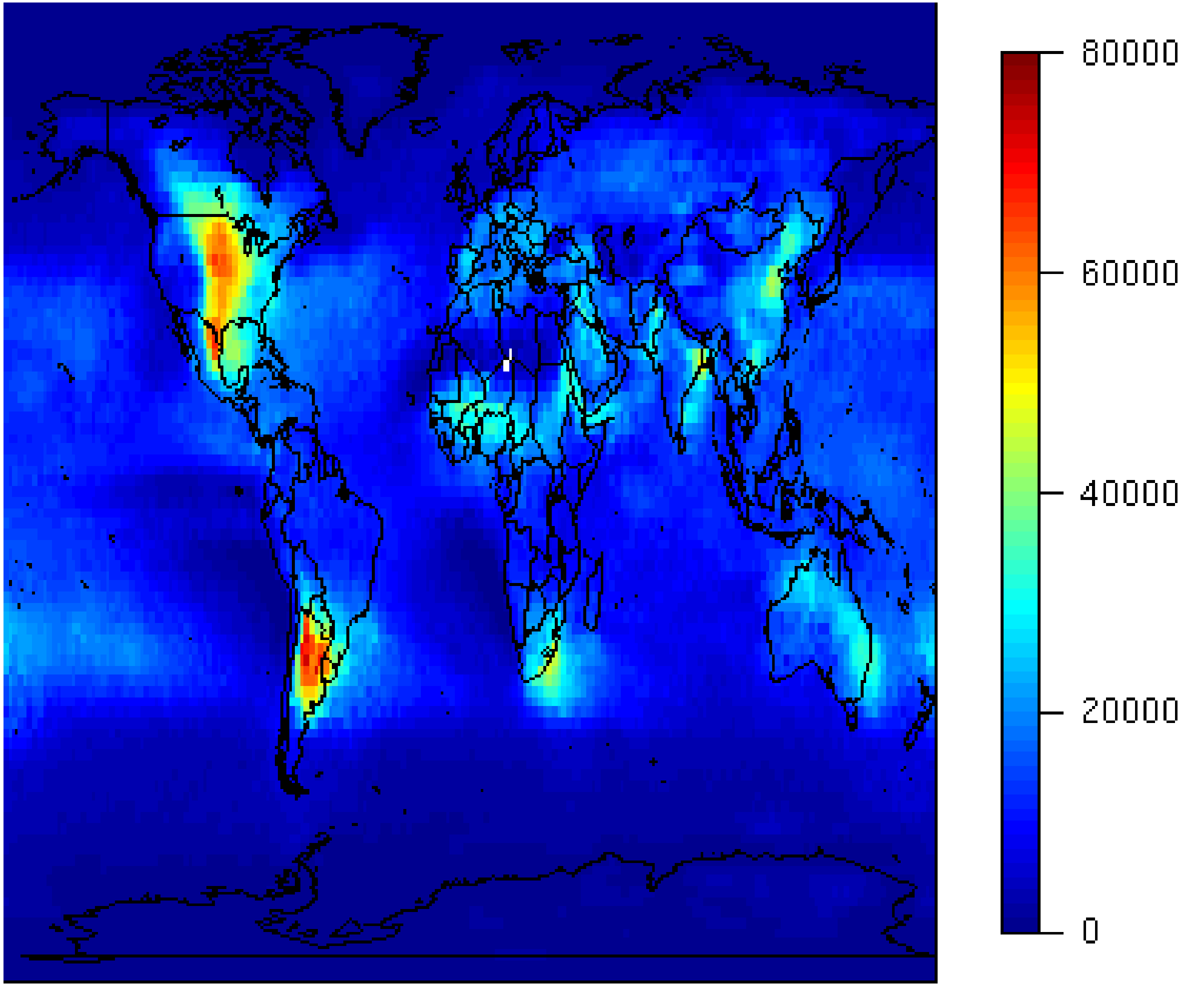


Preliminary Analysis: Shape Parameter



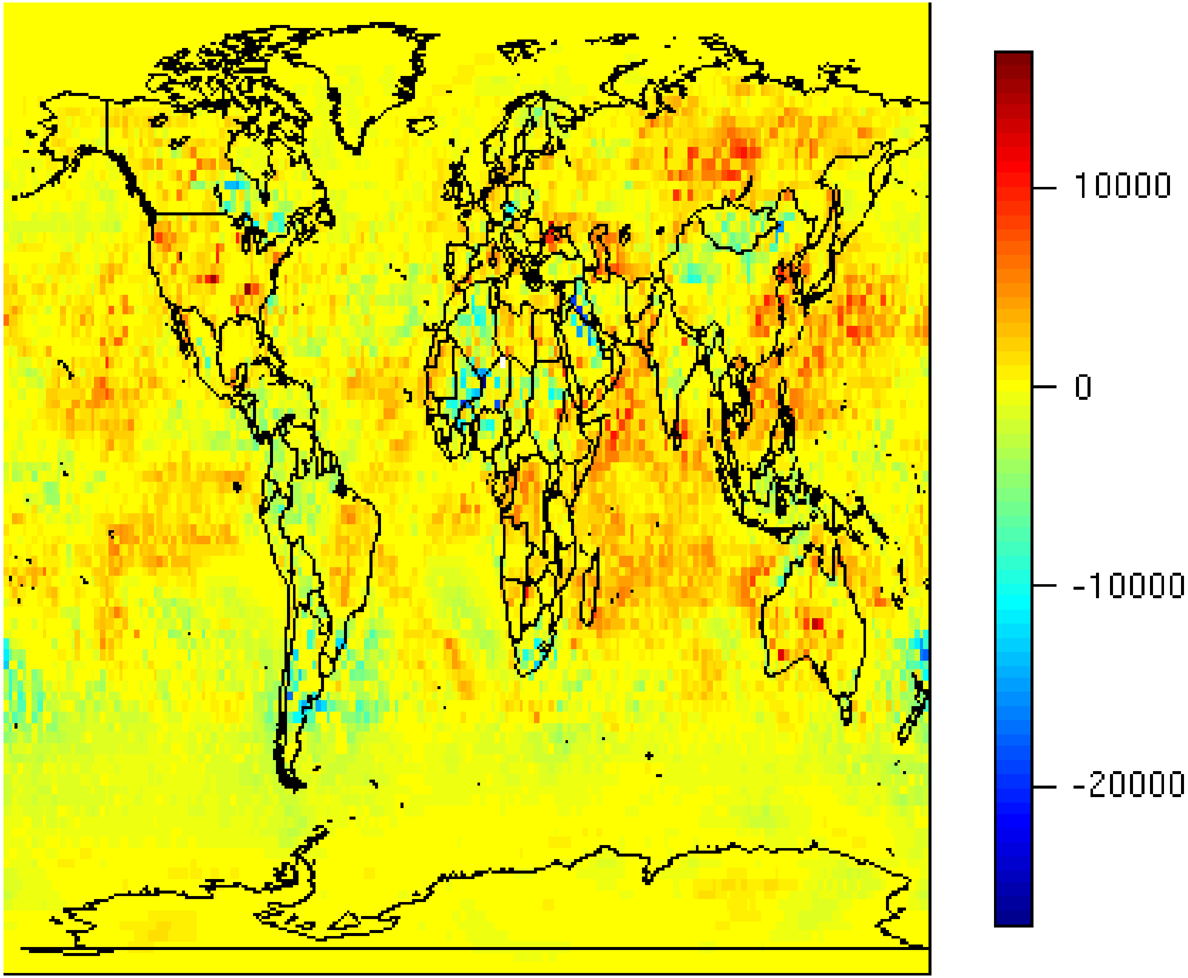
Preliminary Analysis

Trend in location parameter: $\mu(\text{year}) = \boldsymbol{\mu_0} + \mu_1 \cdot \text{year}$



Preliminary Analysis

Trend in location parameter: $\mu(\text{year}) = \mu_0 + \boldsymbol{\mu_1} \cdot \text{year}$



Preliminary Results

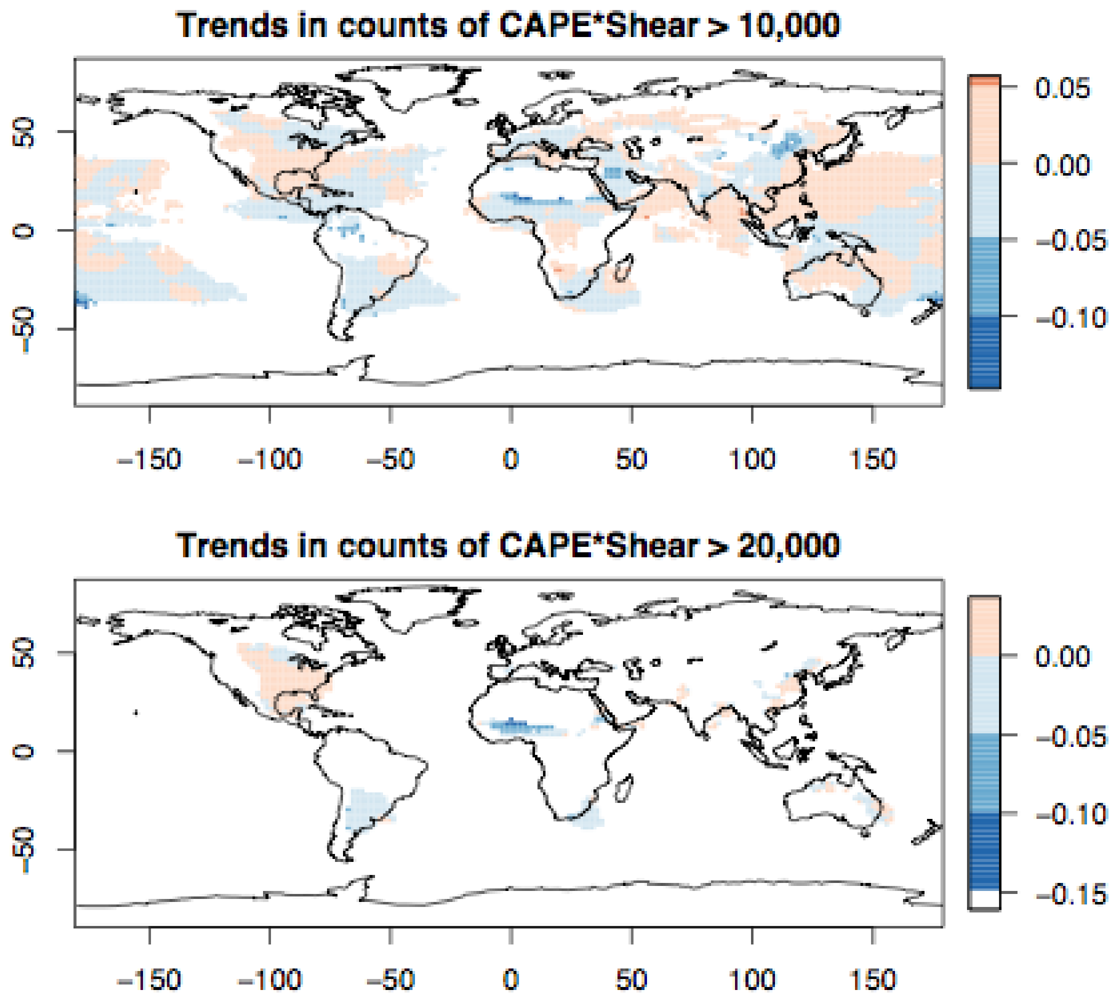
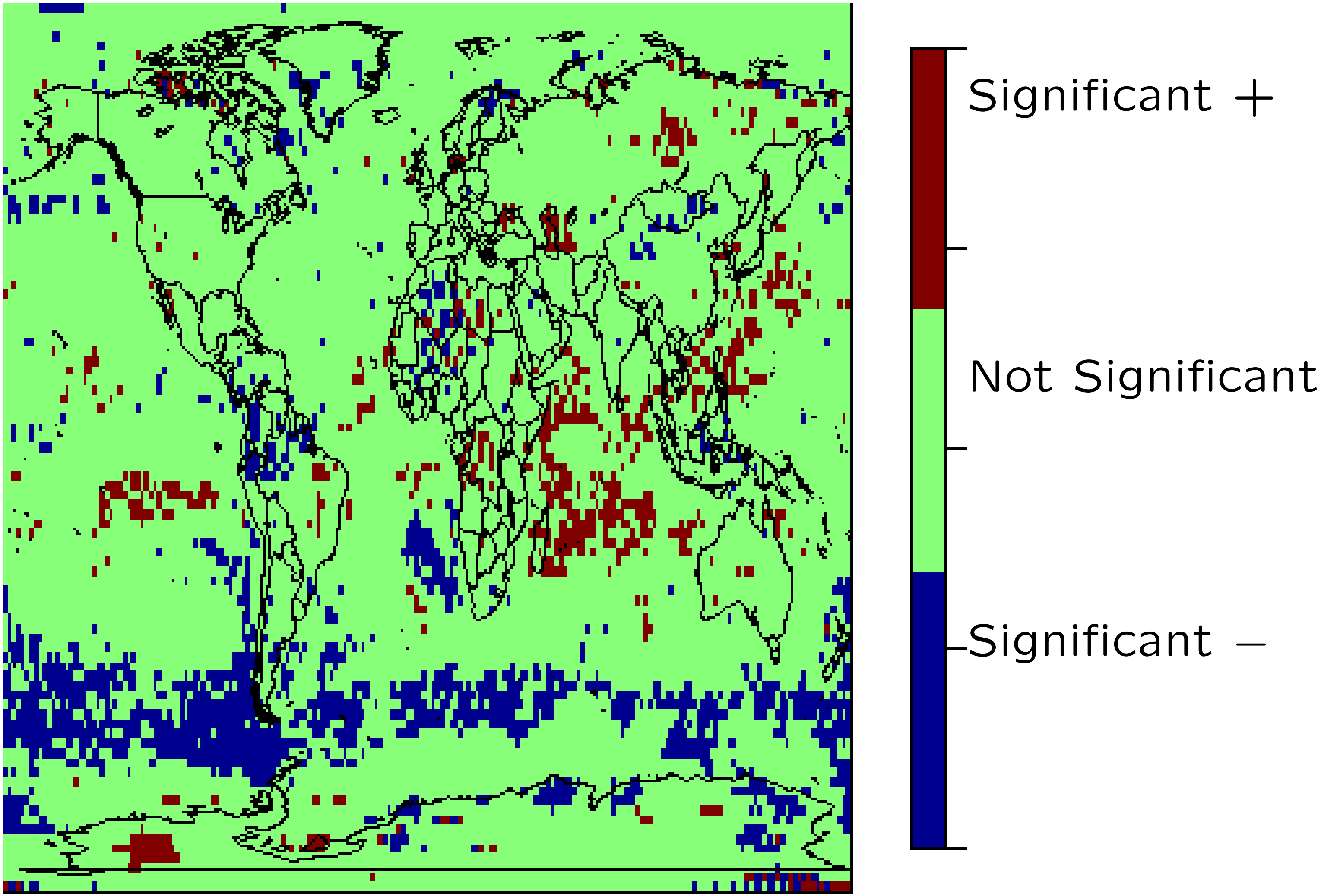


Fig. from Pocernich *et al* (in prep)

Preliminary Results

Significance?

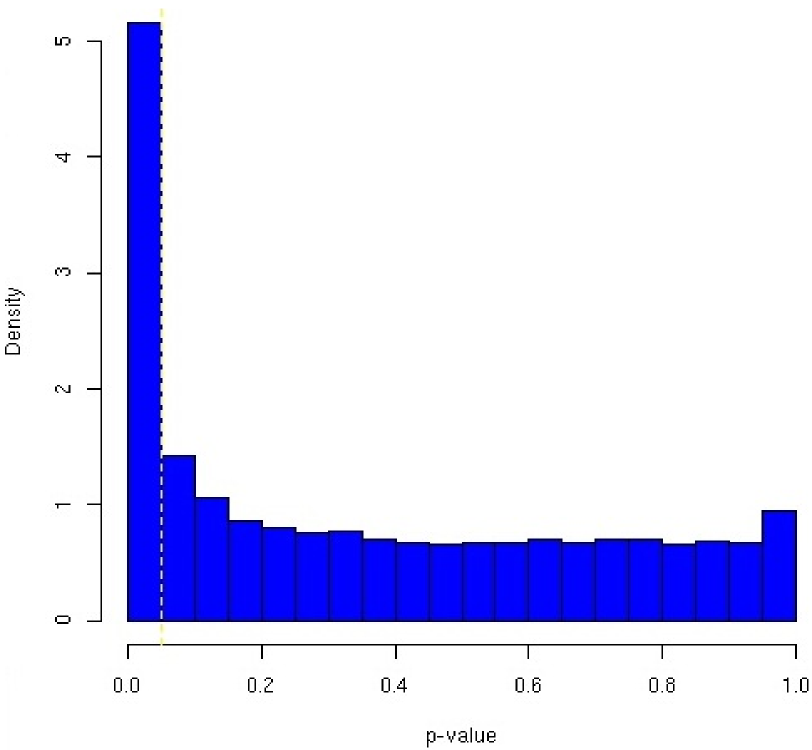


Preliminary Results

Likelihood-ratio tests for

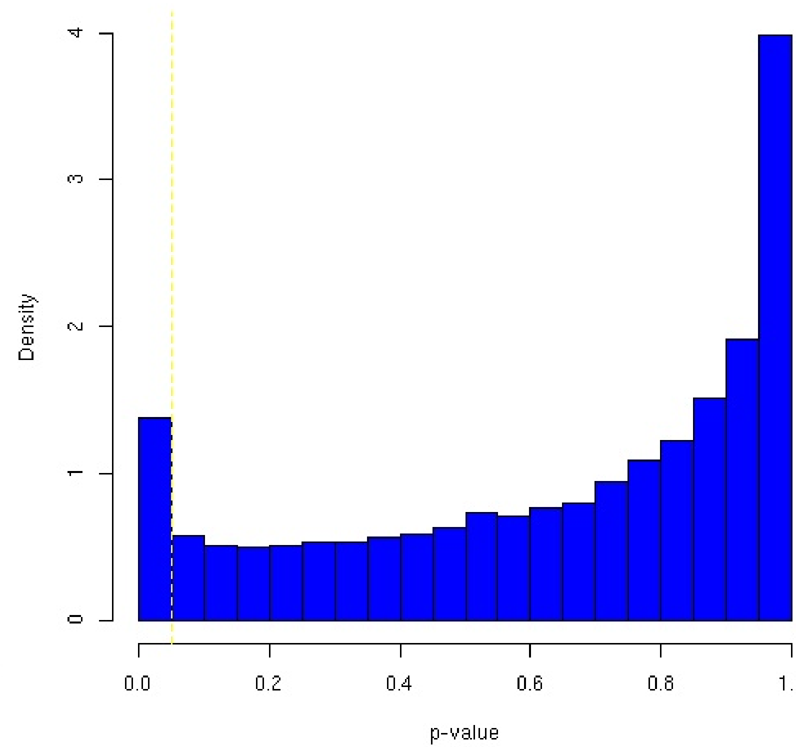
linear trend

$\mu(\text{year}) = \mu_0 + \mu_1(\text{year})$



quadratic trend

$\mu(\text{year}) = \mu_0 + \mu_1\text{year} + \mu_2\text{year}^2$

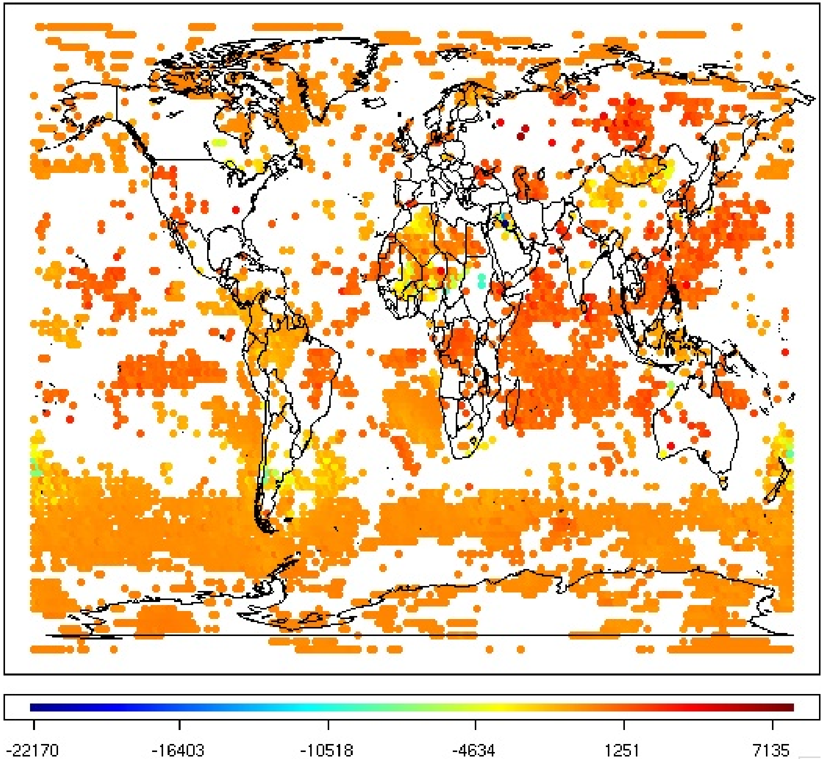


Preliminary Results

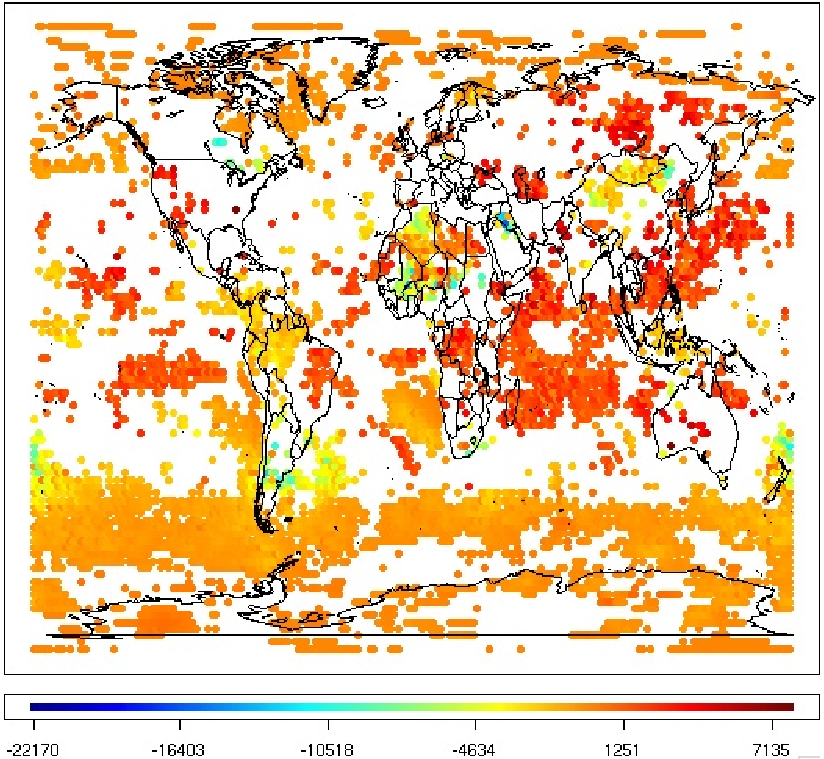
20-year return level differences for *linear/quadratic* trend

$$\mu(\text{year}) = \begin{cases} \mu_0 + \mu_1 \cdot \text{year} \\ \mu_0 + \mu_1 \cdot \text{year} + \mu_2 \cdot \text{year}^2 \end{cases}$$

$z_{1975} - z_{1962}$



$z_{1989} - z_{1962}$



Uncertainty

Likelihood-ratio test $\longrightarrow$ Model fit

Also want to know about uncertainty for *return level differences*

- δ method (shorter return periods)
- profile likelihood (longer return periods)
  not realistic for so many grid points
- Bootstrap
- Bayesian hierarchical model

Tanke wol!

Developmental version of the spatial extension of `extRemes` available at:

`http://www.isse.ucar.edu/extremevalues/evtk.html`

Ozone data available at:

`http://www.image.ucar.edu/GSP/Data/O3.shtml`

Email: `EricG @ ucar.edu`

References

Brooks HE, JW Lee, and JP Craven, 2003: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmos. Res.*, **67-8**:73–94.

Cooley D, D Nychka, P Naveau. (*in press*). Bayesian Spatial Modeling of Extreme Precipitation Return Levels. JASA.

Cooley D, P Naveau, V Jomelli, A Rabatel, D Grancher, 2006a. A Bayesian hierarchical extreme value model for lichenometry. *Environmetrics*, **17**(6):555–574.

Cooley D, P Naveau, P Poncet, 2006b. Variograms for max-stable random fields. In *Statistics for Dependent Data*: Springer Lecture Notes in Statistics No. 187.

Craigmile PF, N Cressie, TJ Santner, and Y Rao, 2006. A Loss function approach to identifying environmental exceedances. *Extremes*, **8**(3):143–159.

Gilleland E, D Nychka, and U Schneider, 2006. Spatial models for the distribution of extremes, *Computational Statistics: Hierarchical Bayes and MCMC Methods in the Environmental Sciences*, Edited by J.S. Clark and A. Gelfand. Oxford University Press.

Gilleland E and D Nychka, 2005. Statistical models for monitoring and regulating ground-level ozone. *Environmetrics* **16**: 535–546 doi: 10.1002/env.720

Pocernich M, E Gilleland, HE Brooks, and BG Brown, in prep. Identifying patterns and trends in severe storm environments using Re-analysis Data. *Manuscript in Preparation*

Stephenson A and E Gilleland, 2005. Software for the Analysis of Extreme Events: The Current State and Future Directions, *Extremes* **8**:87–109.