Climate Extremes,
What to do with 8000 histograms?

Douglas Nychka and Whitney Huang,
National Center for Atmospheric Research

Summary

- Precipitation extremes
- Regional Climate models
- Adding a spatial element
- What about *Yellowstone* and *Cheyenne*?

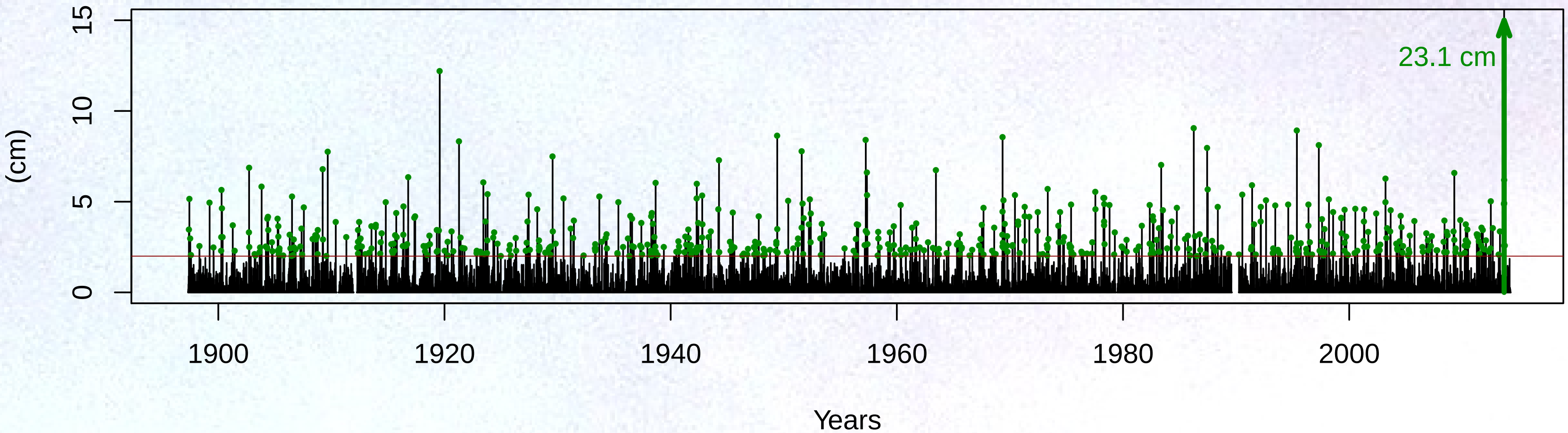
Components: Density estimates, Good and Gaskins (1971), functional data, sparse and embarrassingly parallel methods,

Credits:

Dorit Hammerling, Sophia Chen, and Nathan Lenssen.

Precipitation extremes for Boulder, CO

Daily precipitation amounts for Boulder



25 year daily return level:

In any given year daily precipitation has a 1/25 chance of exceeding this level.

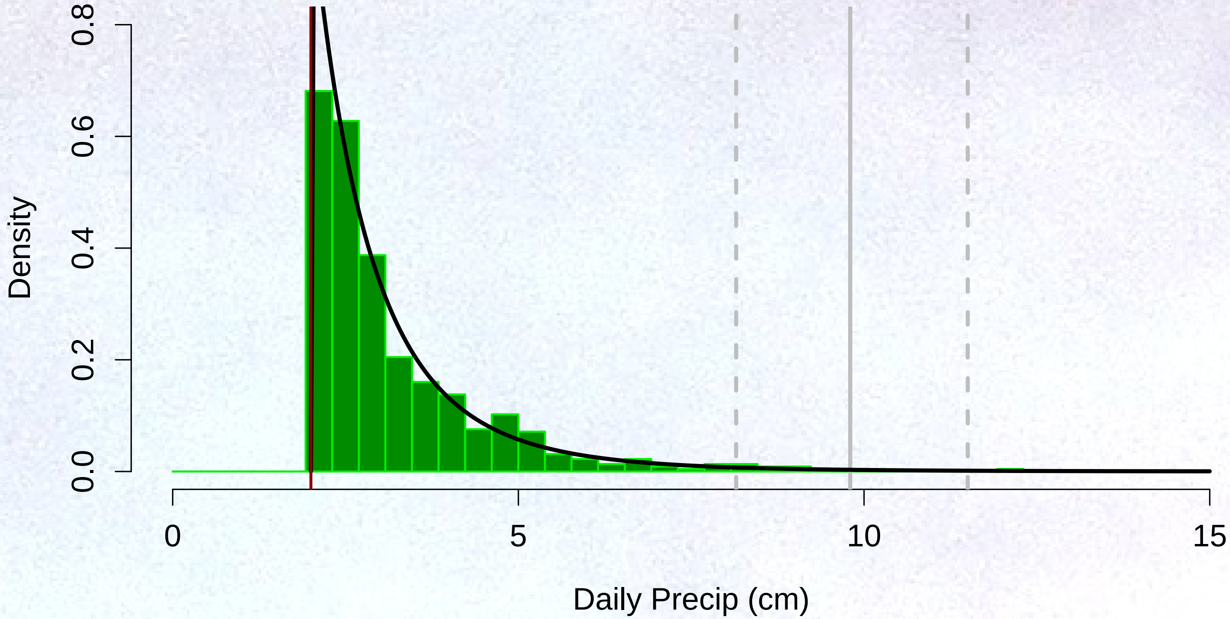
How does this vary over space?

How well does a model simulate this variable?

PART 1:
Estimates of climate extremes

- Generalized Pareto pdf
- Nonparametric density estimates

Generalized Pareto Fit:



Fit to observations > 2 cm
with 95% CI for
25 year return level

Generalized Pareto: pdf(x) depends on three parameters:

$$pdf(x) \sim \left((1 + \xi \frac{(x-\mu)}{\sigma} \right)^{-\frac{1}{\xi}} \text{ for } x \geq \mu$$

- (1) scale (σ) , (2) shape(ξ) and (3) probability of exceeding threshold ($P(Z > \mu)$) .
- With these one can find all quantiles, means and return levels.

Beyond the Pareto

Probability density function:

$$pdf(x) = e^{g(x)}$$

g is the log density function

- Estimate g as a flexible spline function

and in the scale of log precipitation.

i.e. x = log(precip)

Good and Gaskins (1971)

Given a random sample $\{Y_1, \ldots, Y_n\}$ *log penlized likelihood :*

$$max_f \sum_{i=1}^{n} log(f(Y_i)) - R(f)$$

subject to $\int f(x)dx = 1, f > 0$.

R is a roughness or other kind of penalty or the log prior density for f.

with $g = log(f)$

$$max_g \sum_{i=1}^{n} g(Y_i) - \lambda R(g)$$

subject to $\int e^{g(x)}dx = 1$

$\lambda > 0$, a smoothing parameter that controls the weight given to the penalty.

Silverman (1982) suggested:

$$max_g \sum_{i=1}^{n} g(Y_i) - \int e^{g(x)} dx - \lambda R(g)$$

Satisifies density constraint as long as for any constant α

$$R(g) = R(g + \alpha)$$

Roughness with exponential function in the null space

$$R(g) = \int \left[g'' \right]^2 dx$$

Solution is a peicewise cubic smoothing spline.

More on $\hat{g}$

- For λ large g is close to linear,
outside range of data g is *exactly* linear

- Silverman proves estimator has the "usual" optimal nonparametric convergence rates.

- But only for a nonzero density on a finite interval.

Approximate, but fast, log densities

- Apply a Possion generalized linear model to a finely binned histogram of counts

- Expected counts in bin i is $\approx ng(x_i)\delta$
δ is the histogram bin width

- Use a penalized, cubic spline smoother and estimate the smoothing parameter by approximate cross validation.

log Penalized likelihood,

$$max_g \left(\sum_{j=1}^{N} \boldsymbol{y}_j g_j - n\delta e^{g_j} \right) - \lambda J(g) + \text{constants}$$

x_j bin midpoints, $\boldsymbol{y}_j$ bin counts, $g_j = g(x_j)$

- Maximization is easy using iteratively reweighted least squares

- Flexibility in modeling — nested in the `gam` universe.

Details ...

$$R(g) = \int \left[g''\right]^2 dx$$

- Constrains g to extrapolate as a linear function

linear g $\rightarrow$ pdf has exponential tail in log(precip)

$\rightarrow$ polynomial tail precip

Off the shelf tools:

- *logspline* – Kooperberg R package, Stone et al (1997)
- Chong Gu spline density estimate
- Adapt *gam*, *mgcv* – S. Woods R packages

More on density estimates

Suppose the binning is fine enough so that only one observation is in each bin.

i. e. y_i is either 1 or 0

Penalized Poisson likelihood:

$$\left(\sum_{j=1}^{N} \boldsymbol{y}_j g_j - n\delta e^{g_j} + log(\boldsymbol{y}_j!) \right) - R(g)$$

$$\sum_{j=1}^{N} \left(\boldsymbol{y}_j g_j - n\delta e^{g_j} \right) \approx \sum_{i=1}^{n} g(Y_i) - (n) \int e^{g(x)} dx$$

So approximately maximizing:

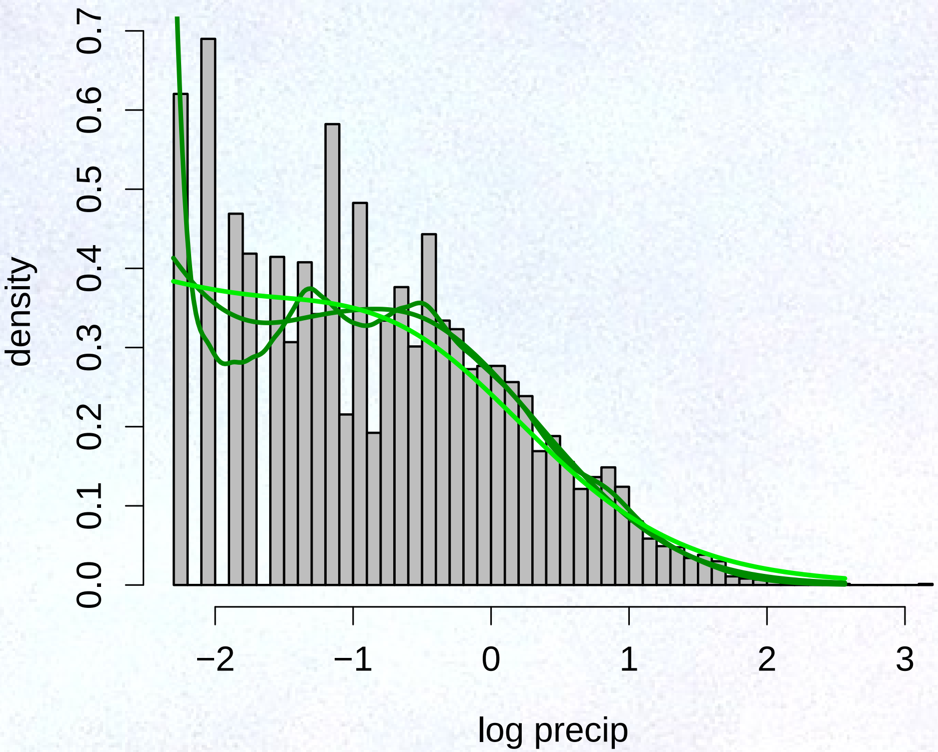
$$\sum_{i=1}^{n} g(Y_i) - (n) \int e^{g(x)} dx - R(g)$$

– exactly the logspline density estimate!

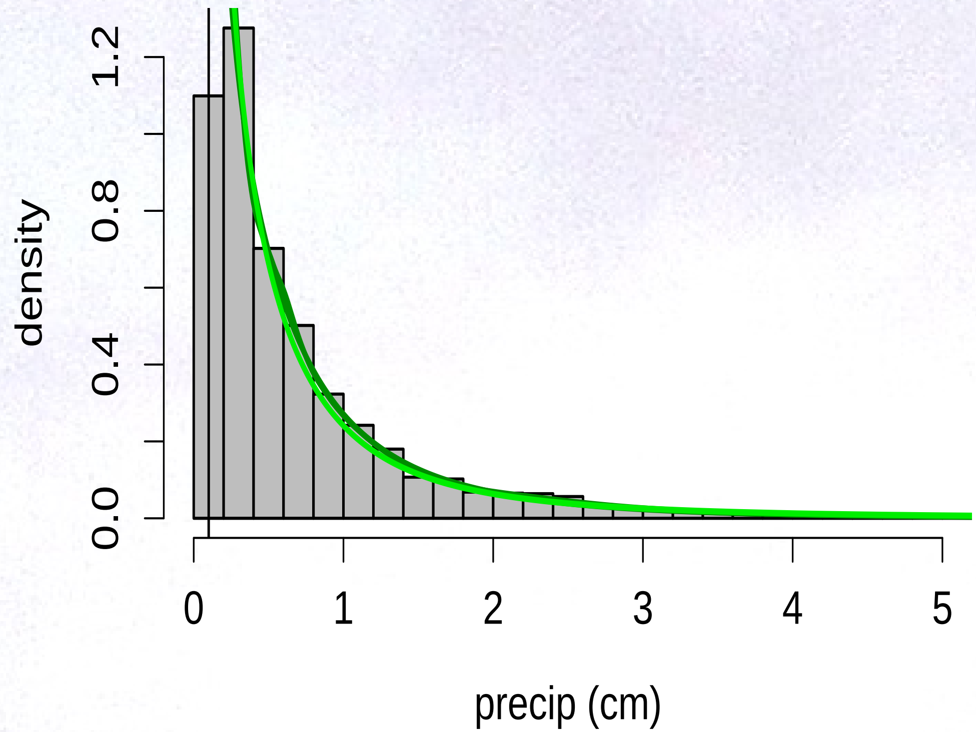
Fit to Boulder data

Three different smoothing parameters:

Log scale

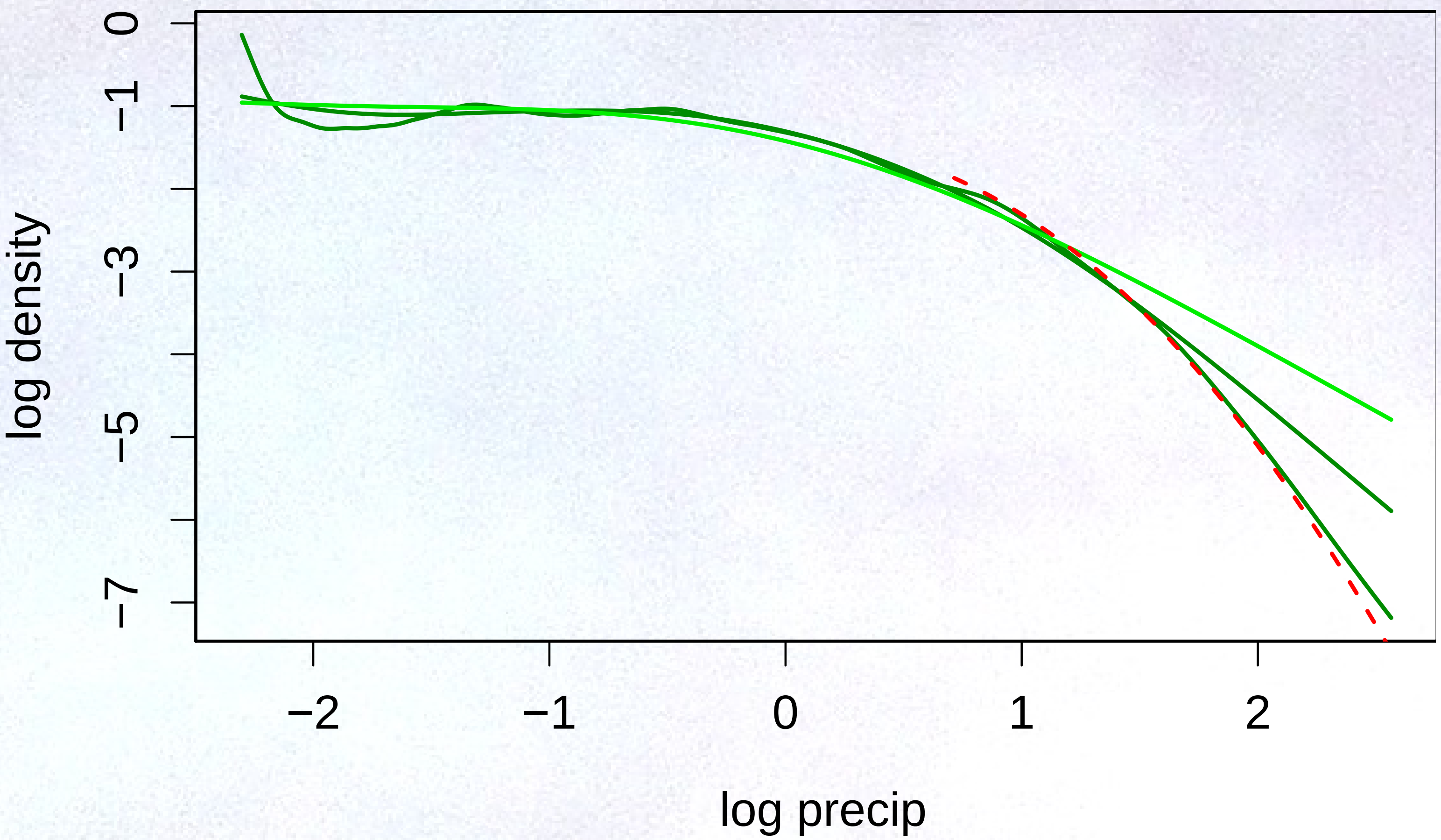


Raw scale:



Cross validation choice for λ is effected by discretization at small precipitation amounts.

log densities

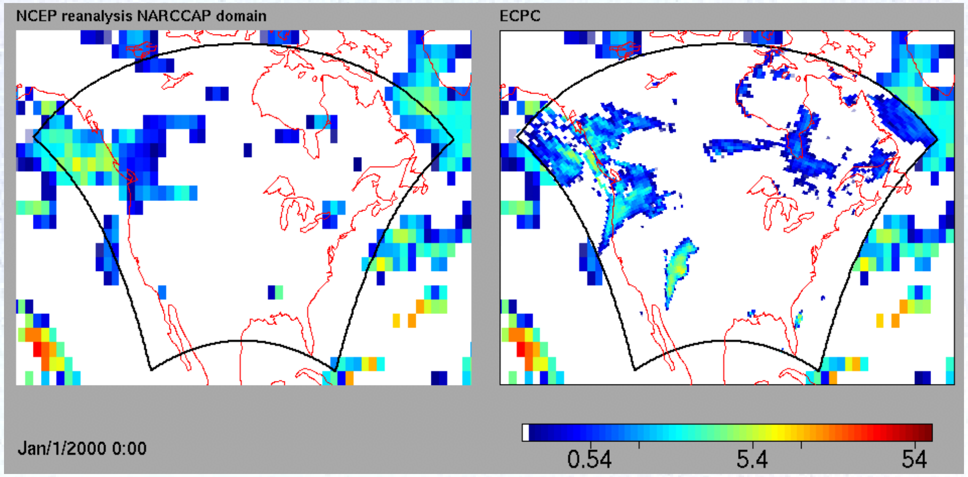


log spline rough , log spline smooth, Generalized Pareto

PART 2: Extremes from regional climate models

Modeling strategy

- Nest a fine-scale weather model in part of a global model's domain.

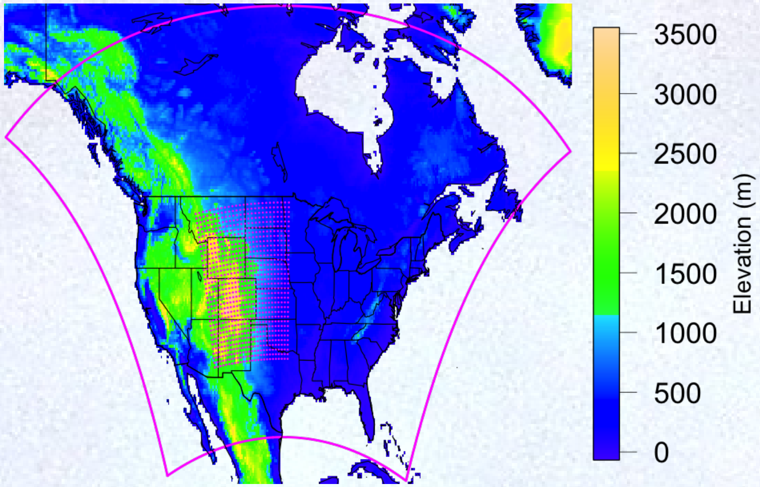


A snapshot from the 3-dimensional RSM3 model (right) forced by global observations (left)

- Consider different regional models to characterize model uncertainty.

- North American Regional Climate Change and Assessment Program (NARCCAP)
a large set of numerical experiments to explore uncertainty.

A very small part of NARCCAP



- Four regional models (MM5I, RCM3, WRFP, ECPC) that are driven by observed atmosphere at the boundaries of the NARCCAP domain.

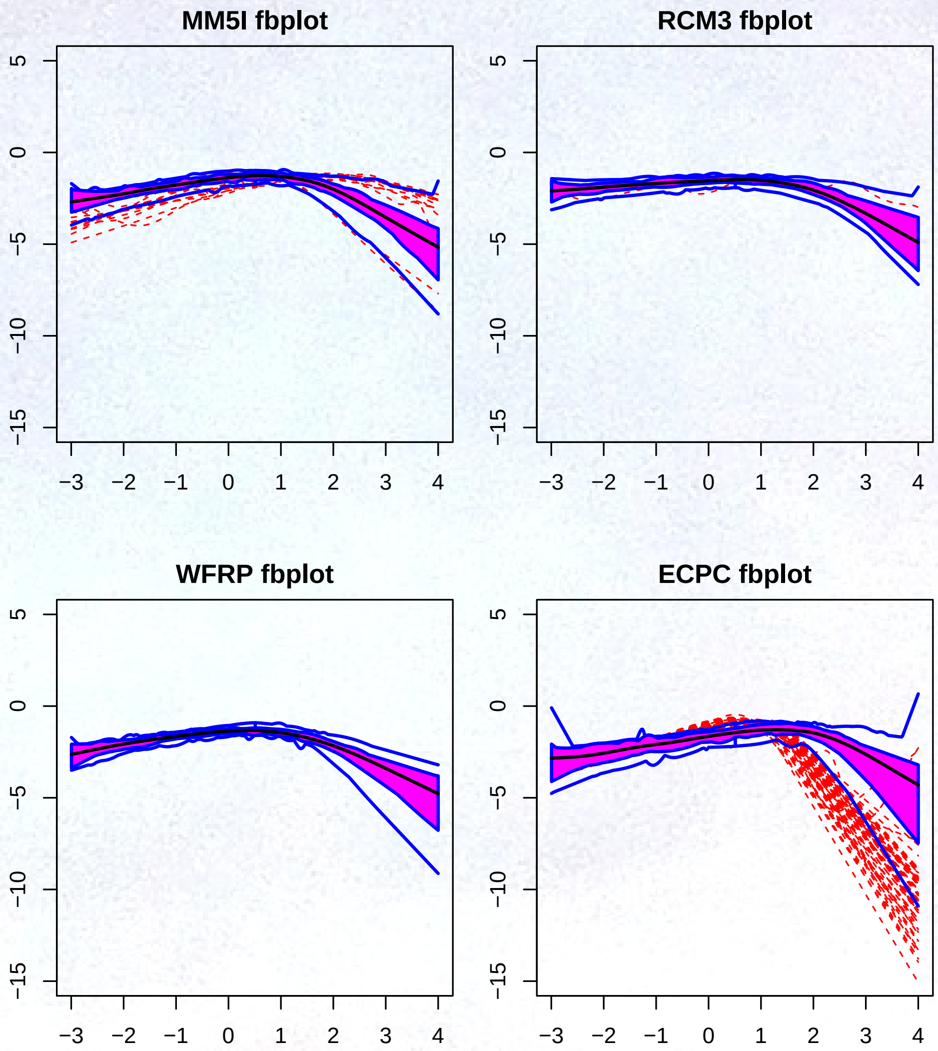
- Just look at part of Rocky Mountain region – about 800 grid points

- 20 years of daily downscaled/simulated weather for each model.

How do extremes of daily summer rainfall vary over space and and over climate models?

Functional boxplots of log densities

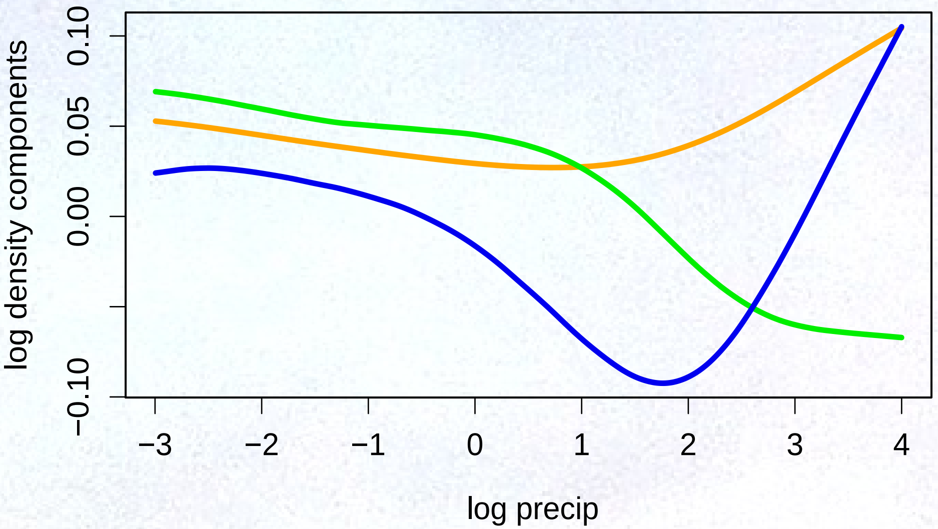
log spline pdfs for four models at all 800+ grid points



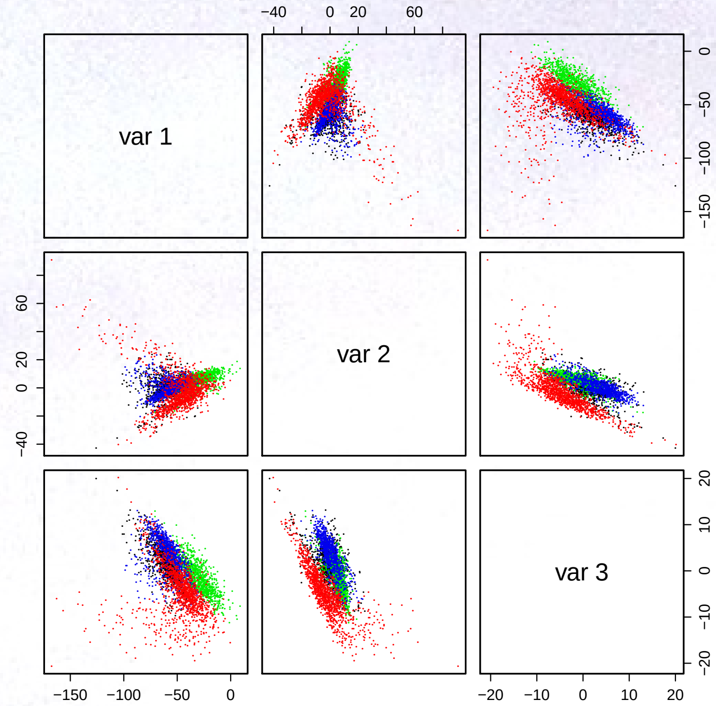
See Sun and Genton (2011) for more on functional boxplots

Principle components

First three principle components of log densities



(a_1, a_2, a_3) by model



$$g(x) = a_1\phi_1(x) + a_2\phi_2(x) + a_3\phi_3(x)$$

(a_1, a_2, a_3) vary for every grid box and every regional model $4 \times 800 \times 3$ coefficients

Use these as basis functions to refit models using standard GLM maximum likelihood

$g(x) = a_1\phi_1(x) + a_2\phi_2(x) + a_3\phi_3(x)$

- Still cheating by fitting bins counts and not normalizing as a density function

- This may be a way to introduce covariates in a simple way

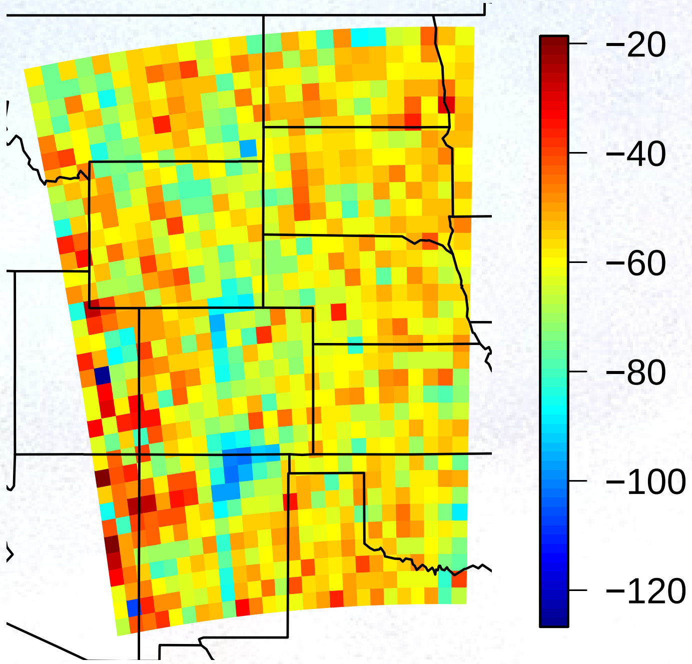
- Local likelihood fitting over space to smooth.

The spatial problem

Coefficients vary over space, are noisy and are correlated.

We have 4 Models $\times$ 3 coefficients $=$ 12 spatial fields.

First coefficient for MM5I

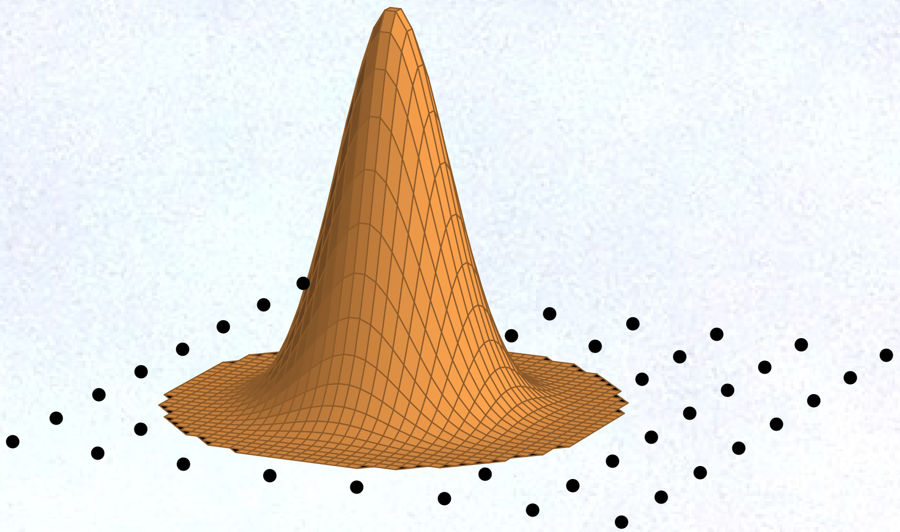


- Transform each climate models coefficients to be uncorrelated.
- Smooth transformed coefficients using spatial statistics.

PART 3:
Spatial stats for large data

LatticeKrig: spatial smoother

Representing the surface: $g(x) = \sum_j \phi_j(x) c_j$



Fix the basis and estimate the coefficients from data.

$$\boldsymbol{y} = X\boldsymbol{c} + \boldsymbol{e} \quad \boldsymbol{c} \sim N(0, Q^{-1})$$

$X_{i,j} = \phi_j(\boldsymbol{x}_i)$

More about Q

Some coefficients:

$$\begin{array}{ccccc} . & . & . & . & . \\ . & . & c_1 & . & . \\ . & c_2 & c_* & c_3 & . \\ . & . & c_4 & . & . \\ . & . & . & . & . \end{array}$$

Some weights:

$$\begin{array}{ccccc} . & . & . & . & . \\ . & . & \text{-1/4} & . & . \\ . & \text{-1/4} & \alpha & \text{-1/4} & . \\ . & . & \text{-1/4} & . & . \\ . & . & . & . & . \end{array}$$

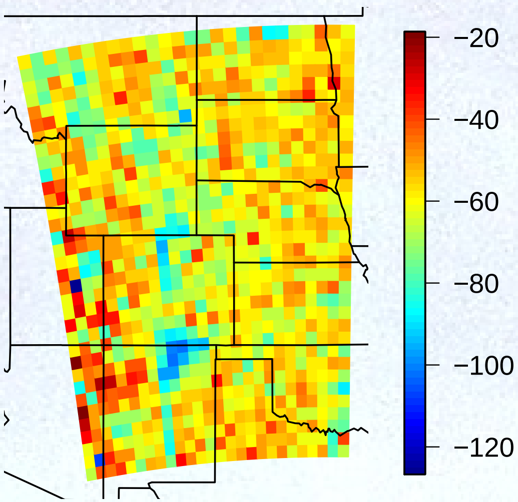
The filter:

$\alpha c_* - 1/4\,(c_1 + c_2 + c_3 + c_4) =$ white noise

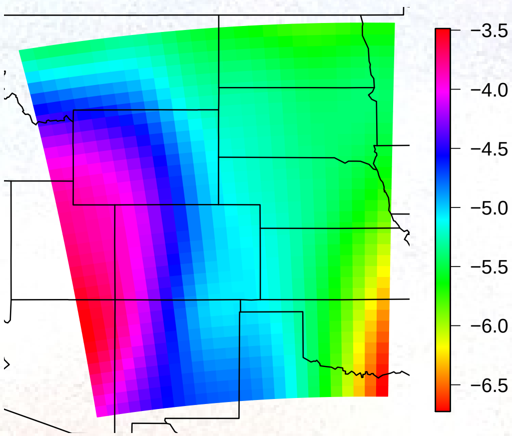
- $\alpha \geq 1$.
- Can exploit sparse linear algebra for the "Kriging" computation
- Multiresolution version approximates standard spatial covariance functions.

First coefficient for MM5I

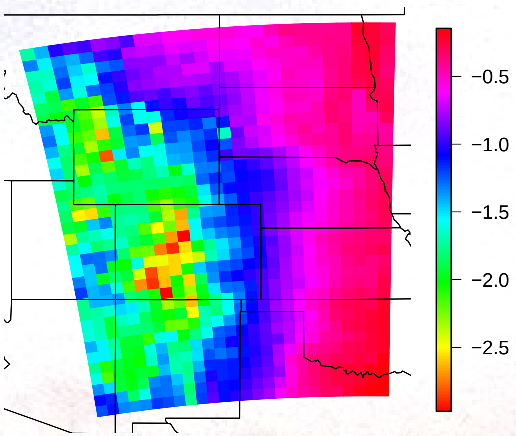
Orignal coefficients.



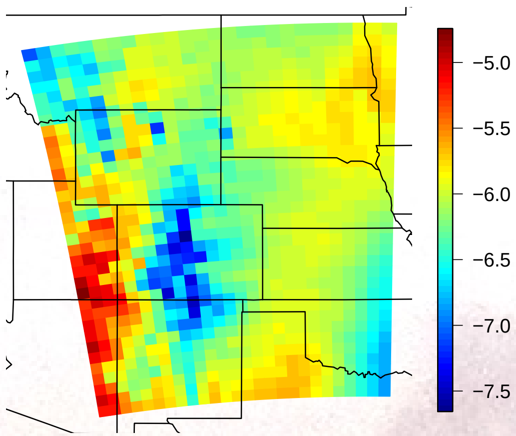
Smooth component



Elevation



Fitted values

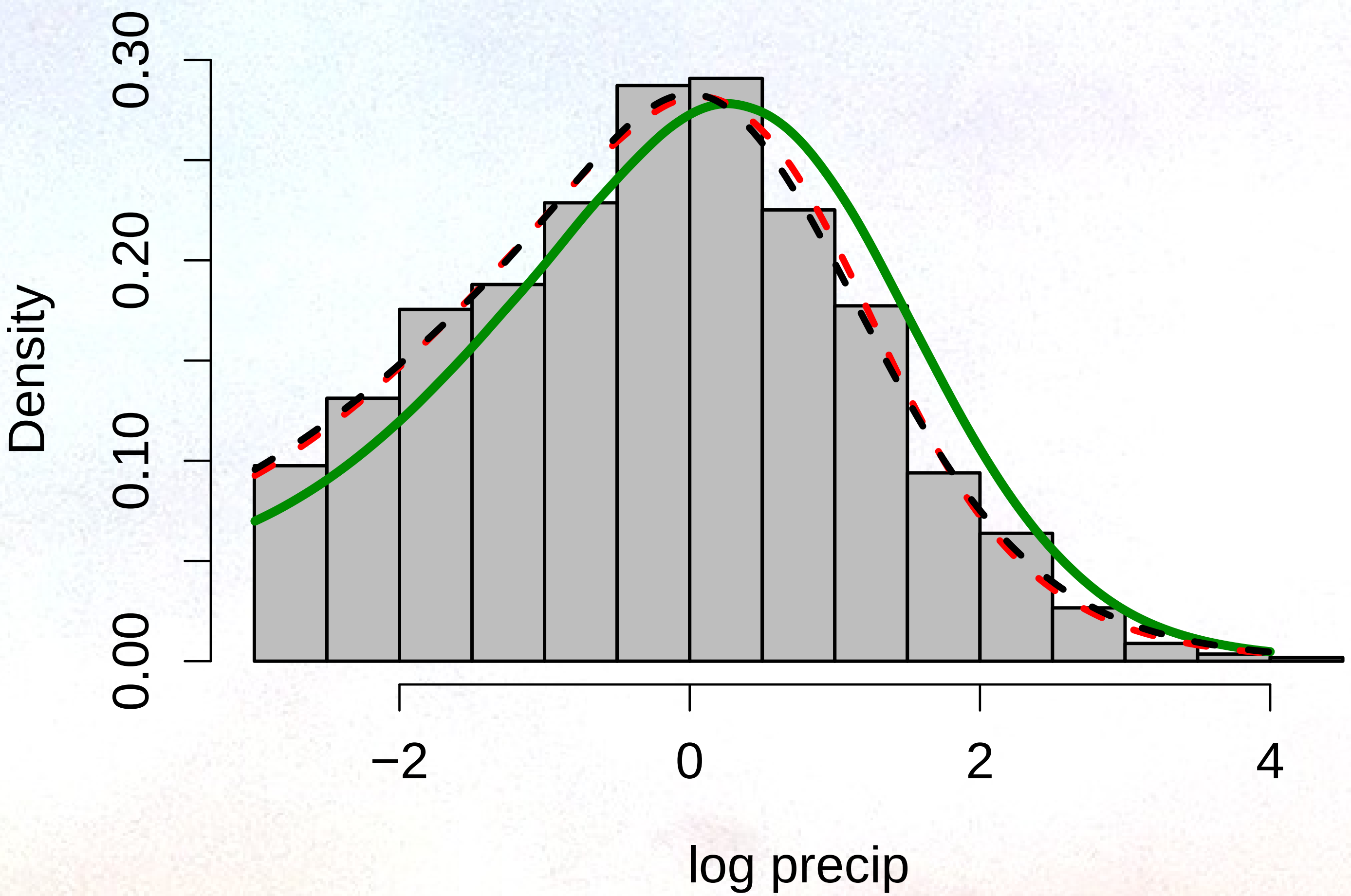


Reconstucting the Boulder grid box

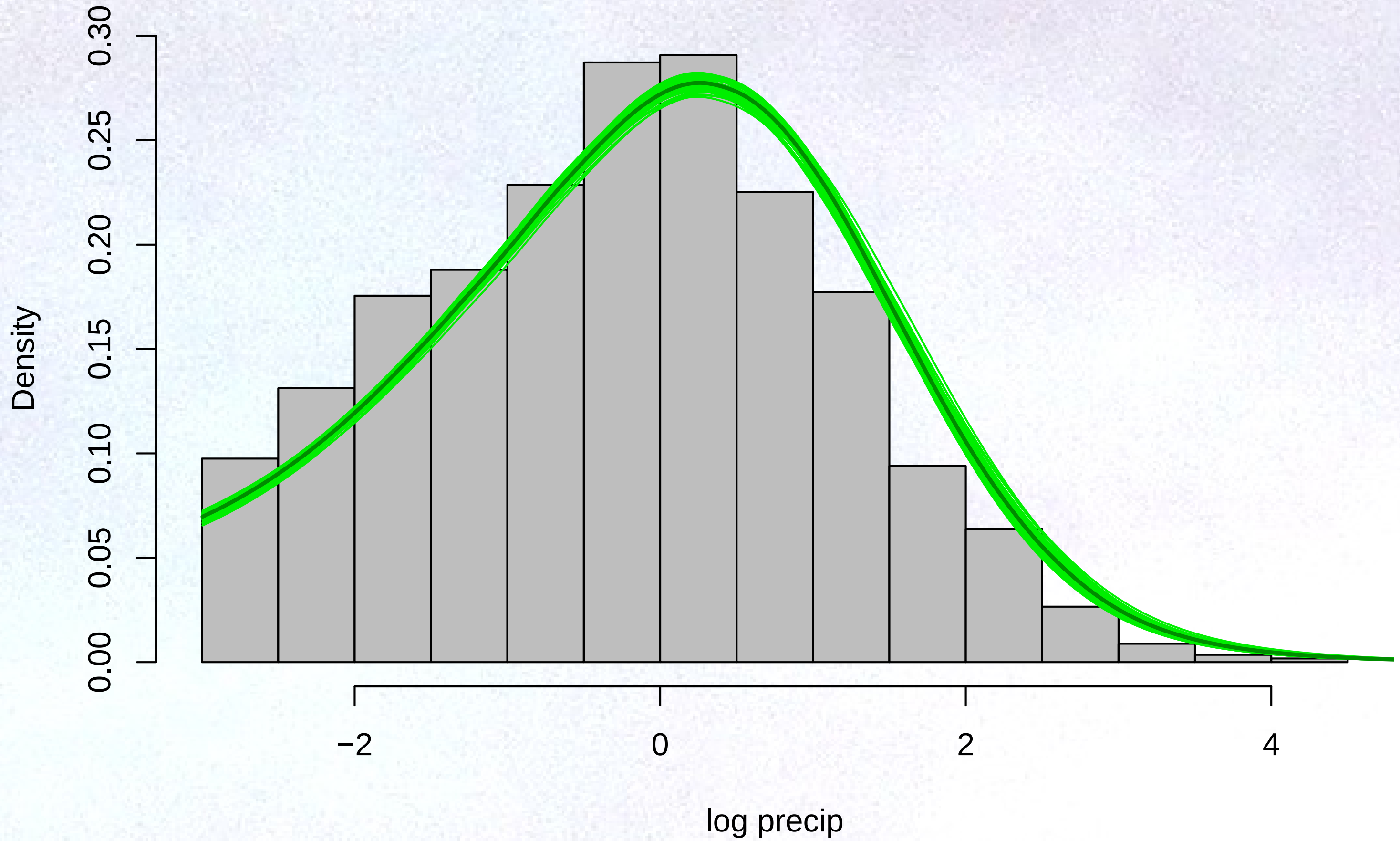
MMI5 model,

log spline , GLM with 3 basis functions,

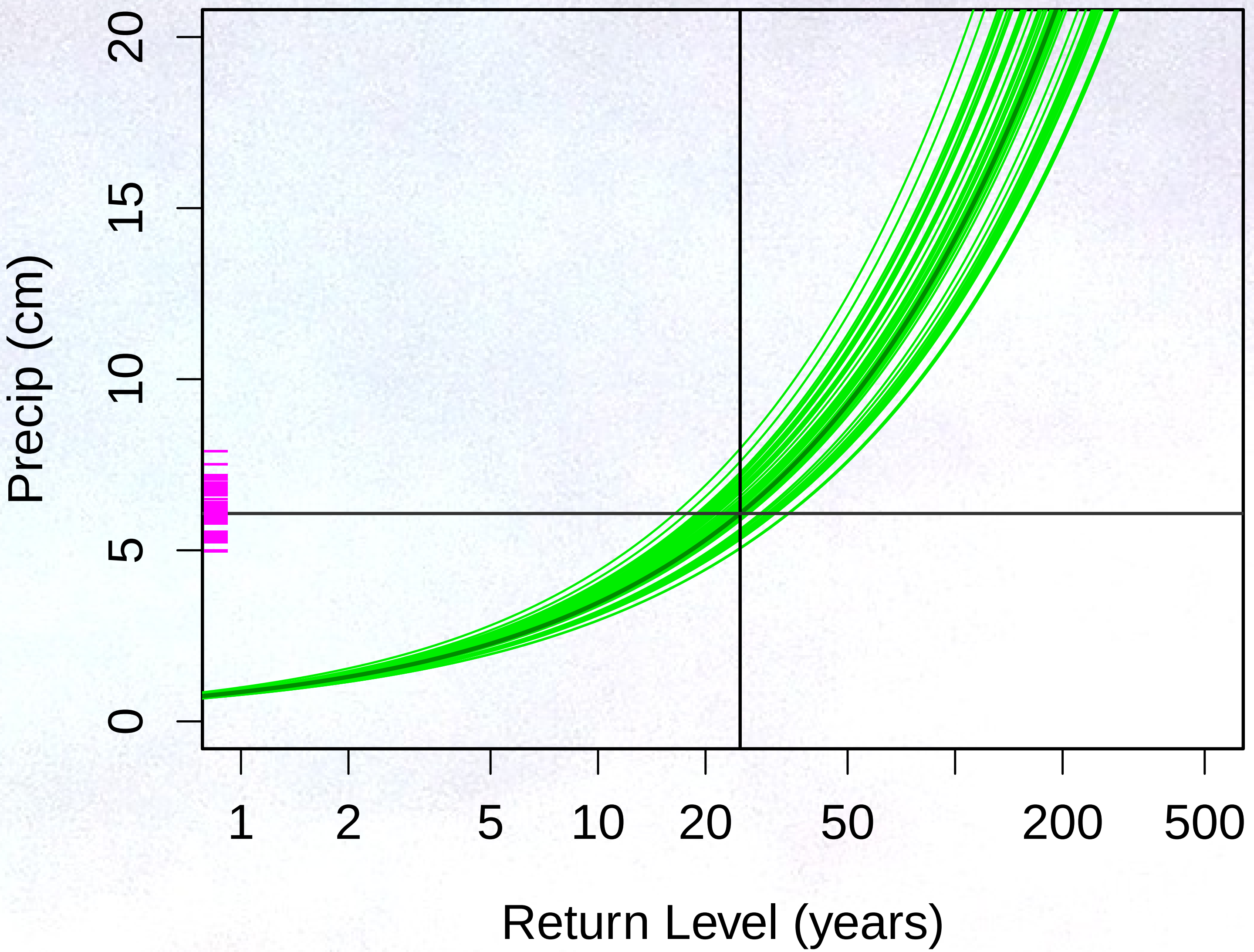
smoothed coefficients



Uncertainty

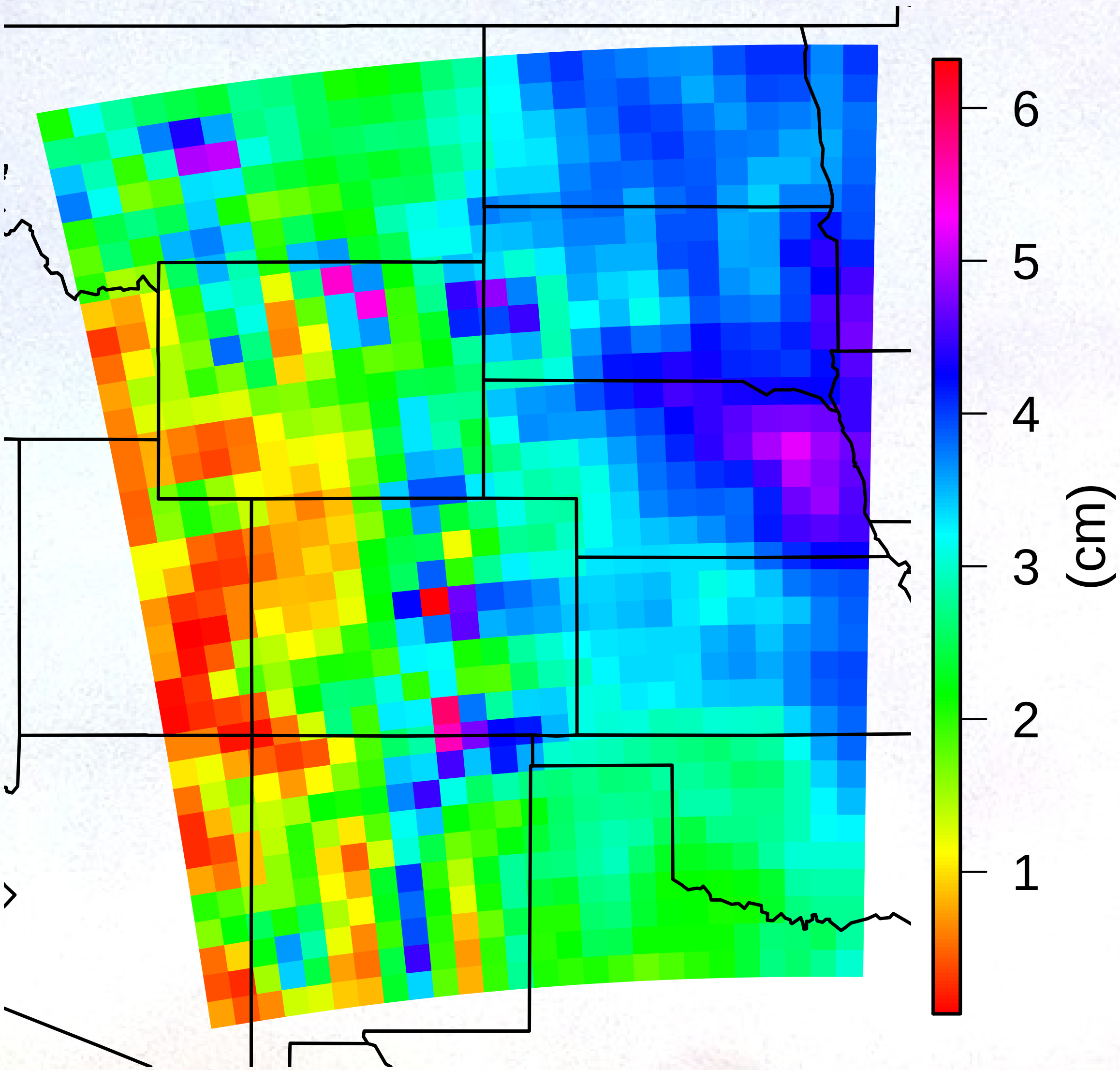


Boulder grid box 25 year return



25 year return surface

"posterior mode" for MM5I model.



PART 4:

Data analysis on *Yellowstone*

If I have to wait too long for my answer I forget my question. – Rich Loft

The Yellowstone supercomputer.

$\approx$72K cores = 4536 (nodes) $\times$ 16 (cores)
and each core with 2Gb memory
16 Pb parallel file system

- Core-hours are available to the NSF research community.
- Simple application process for graduate student allocations.
- *Supports R in both interactive and batch mode.*

Cheyene will be running January 2017 and will have 3 times capacity of *Yellowstone*.

Using the `Rmpi` package.

In R ...

```
library(Rmpi)
# Spawn 4 workers
mpi.spawn.Rworkers(nworkers=4)
 # Broadcast an R function to all workers

mpi.bcast.Robj2worker(doStats)
 # apply this function to 100 tasks (each worker will get about 25)

output <- mpi.iapplyLB(1:100, doStats)
```

`output` *is a list (100 components) with the result for each case.*

Are many R workers processes feasible?

- Time to initiate 100 - 1000 workers nearly constant at 3 seconds

- Workers lose little time reading common data files.

- Median execution time of task per worker is nearly constant.

- Successfully used for fitting extreme value distributions, spatial fields, covariance models.

Summary

- Nonparametric methods are available for estimating the tail behavior of climate distributions.

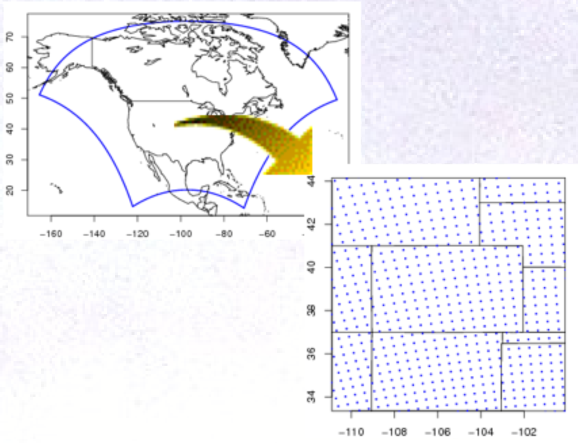
- Borrowing strength from spatial neighbrs and dimension reduction help to make them work

- Methods can be easily migrated to large computing systems.

Thank you

Regional simulations for N. America

North American Regional Climate Change and Assessment Program (NARCCAP)



4GCMS × 6RCMs:

12 runs – balanced half fraction design

Global observations × 6RCMs

X High resolution global atmosphere

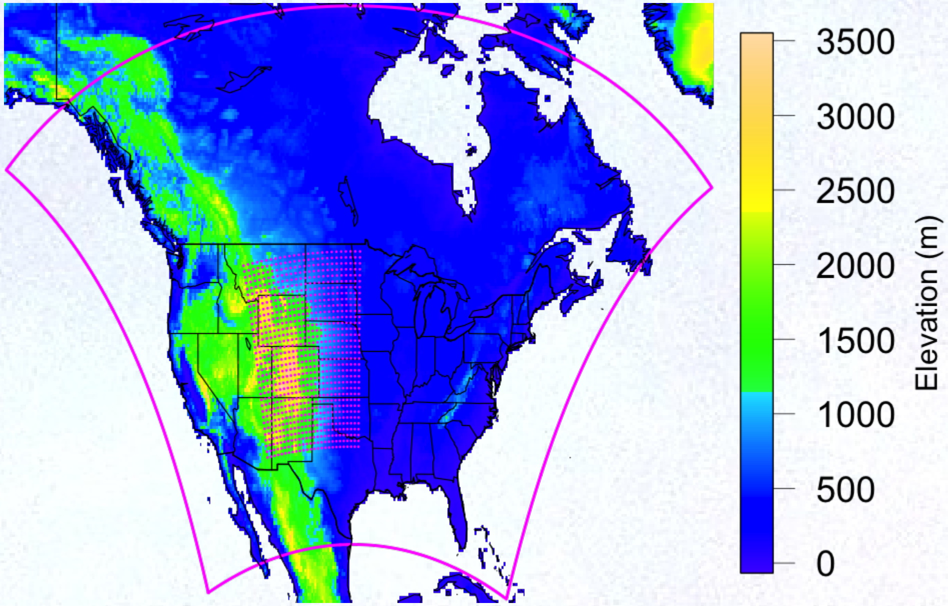
GLOBAL MODEL	**REGIONAL MODELS**					
	MM5I	WRF	HADRM	REGCM	RSM	CRCM
GFDL			●	●	○	
HADCM3	●		●		●	
CCSM	●	●				●
CGCM3		●		●		●
Reanalysis	■	■	●	■	■	●

NCAR grid over land is ≈ 8-9K grid points.

Study region

NARCCAP domain and Rocky Mountain MM5I grid cells.

(About 800 grid points in subregion.)



How do extremes of daily summer rainfall vary over space and and over climate models?