Climate Extremes, Computing Extremes

Douglas Nychka, National Center for Atmospheric Research

National Science Foundation

PIMS Environmental Big Data, Vancouver, May 2015

Summary

- Regional Climate models
- Precipitation extremes
- Adding a spatial element
- High performance computing

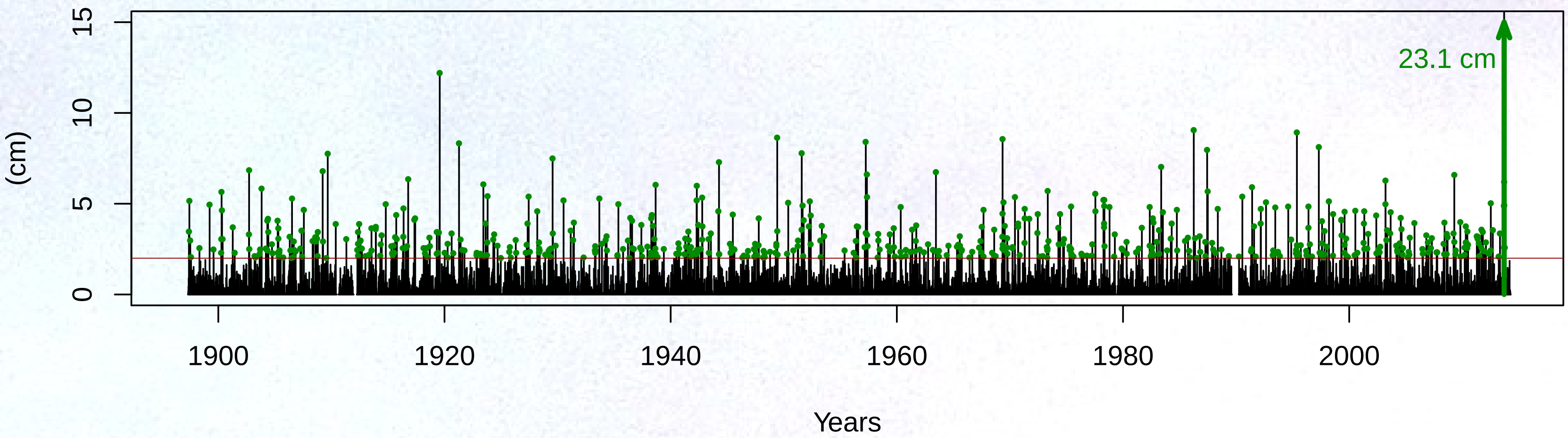
Challenges: NonGaussian distributions, functional data

How do we reduce dimensions? How do we borrow strength?

See Climate Extremes chapter (Zwiers et al) in *Climate Science for Serving Society: Research, Modeling and Prediction Priorities*

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.

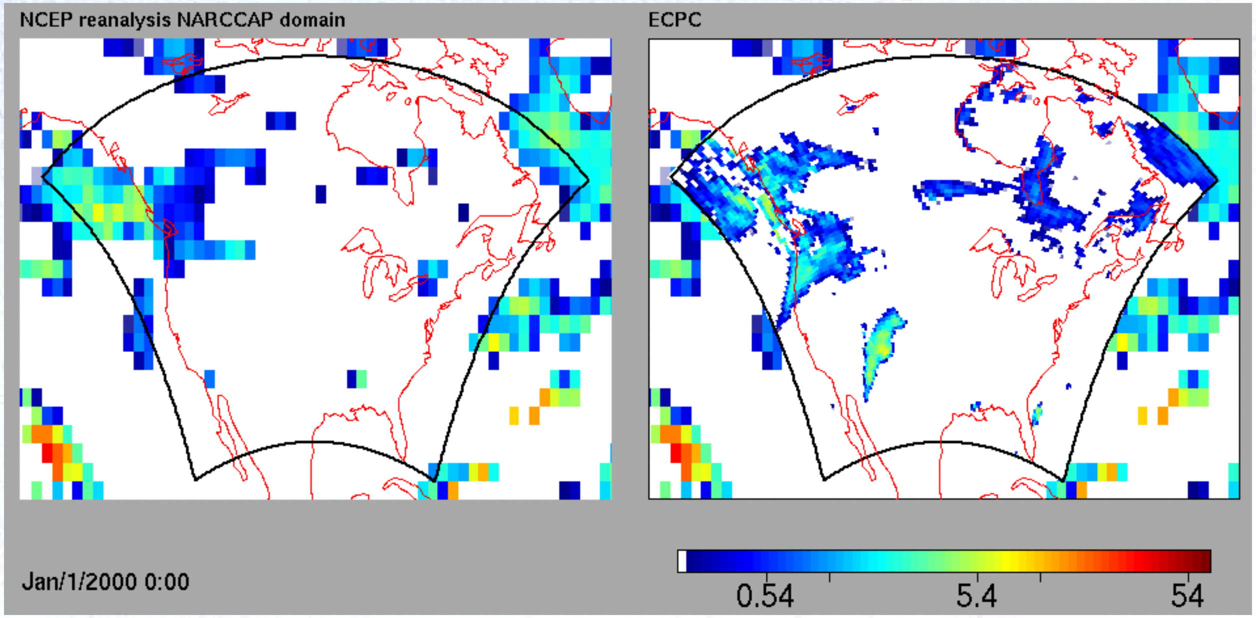
How does this vary over space?

PART 1:
Climate change and regional climate models

An approach to Regional Climate

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

Regional model simulates higher resolution weather based on the global model for boundary values and fluxes.

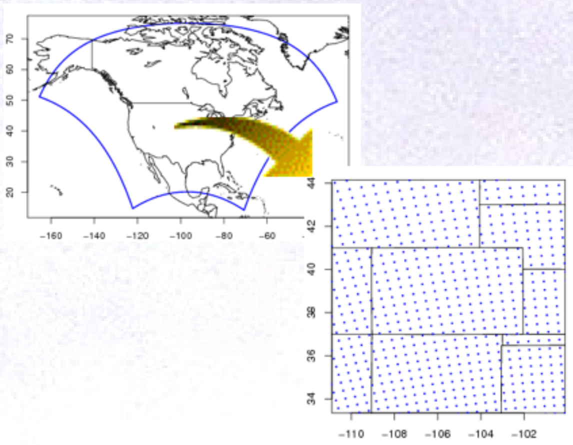


A snapshot from the 3-dimensional RSM3 model (right) forced by global observations (left)

- Consider different combinations of global and regional models to characterize model uncertainty.

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

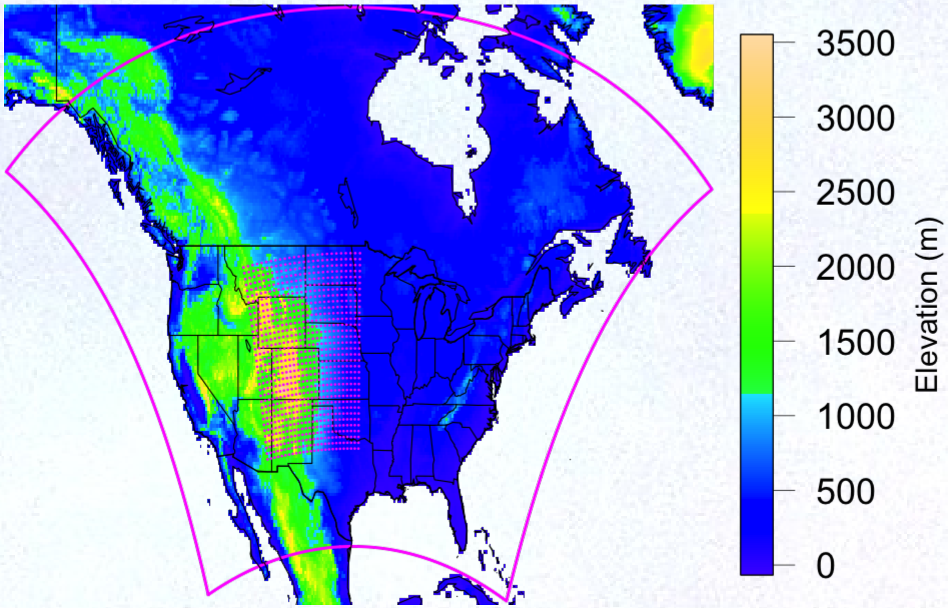
GLOBAL MODEL	REGIONAL MODELS					
	MM5I	WRF	HADRM	REGCM	RSM	CRCM
GFDL			●	●	○	
HADCM3	●		●		●	
CCSM	●	●				●
CGCM3		●		●		●
Reanalysis	■	■	●	■	■	●

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.)



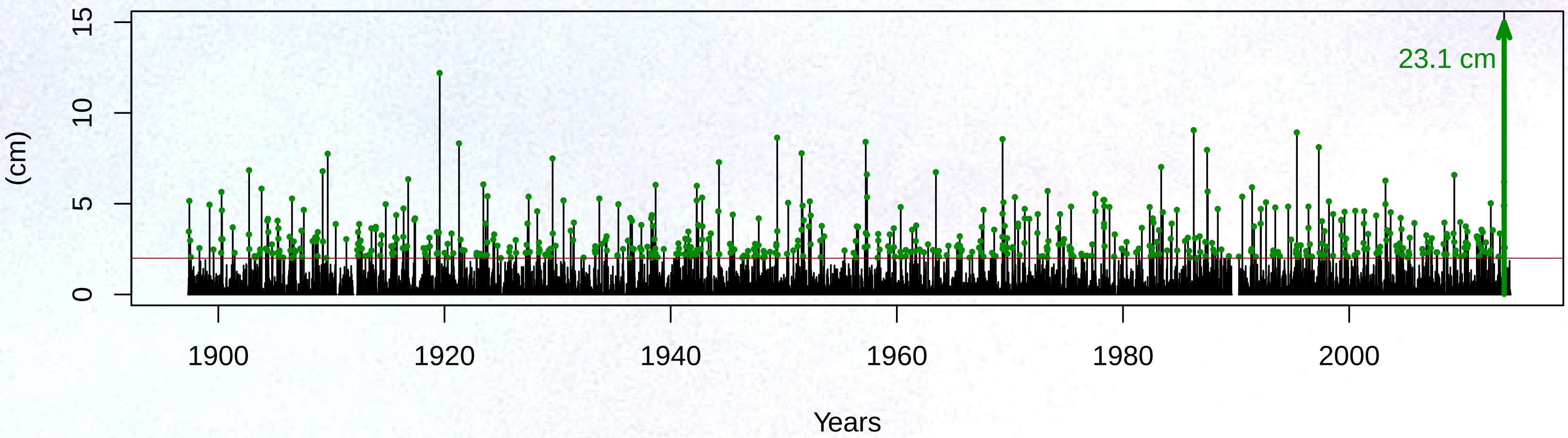
How do extremes of daily summer rainfall vary over space and and over climate models?

PART 2:
Estimates of climate extremes leading to spatial fields

- Three parameters of Generalized Pareto
- Nonparametric density estimates

Precipitation extremes for Boulder

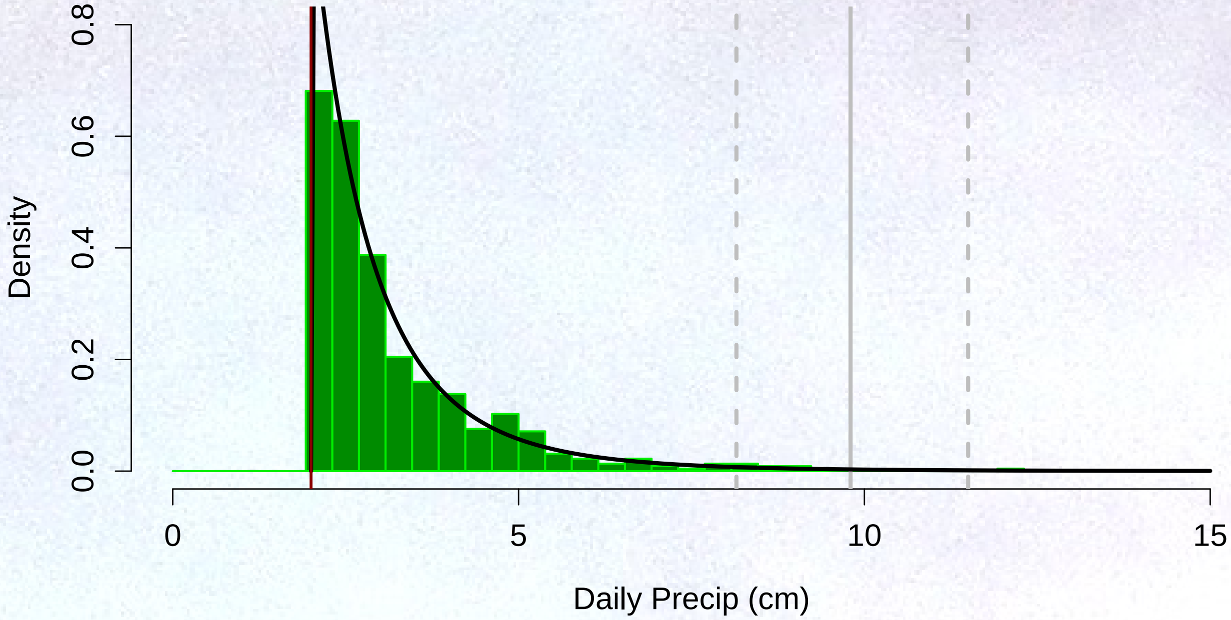
Daily precipitation amounts for Boulder



25 year daily return level:

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

Generalized Pareto Fit:



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

Generalized Pareto: depends on three parameters:

$$P(Z > z + \mu | Z > \mu) = \left((1 + \xi \frac{z}{\sigma} \right)_{+}^{-\frac{1}{\xi}}$$

- (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.

Functional data and space.

How do to manage the estimated distributions at many locations?

- Borrow strength from neighboring locations

- Reduce dimensions to the three Generalized Pareto parameters.

$\boldsymbol{u}$ a location in the region:

- scale($\boldsymbol{u}$)
- shape($\boldsymbol{u}$)
- prob exceedence ($\boldsymbol{u}$)

Beyond the Pareto

Probability density function:

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

Estimate g as a flexible spline function
and in the log scale of precipitation. i.e. x = log(precip)

- Constrain the spline function to extrapolate as a linear function – this implies polynomial tail behavior for the density in the untransformed scale.

- *logspline* – Kooperberg R package, Stone et al (1997)

- Chong Gu spline density estimate

- Adapt *gam, mgcv* – S. Woods R packages

Approximate, but fast, log densities

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

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

- Normalize estimate of g to integrate to one.

With lots of knots this is also a spatial process model.

log Penalized likelihood,

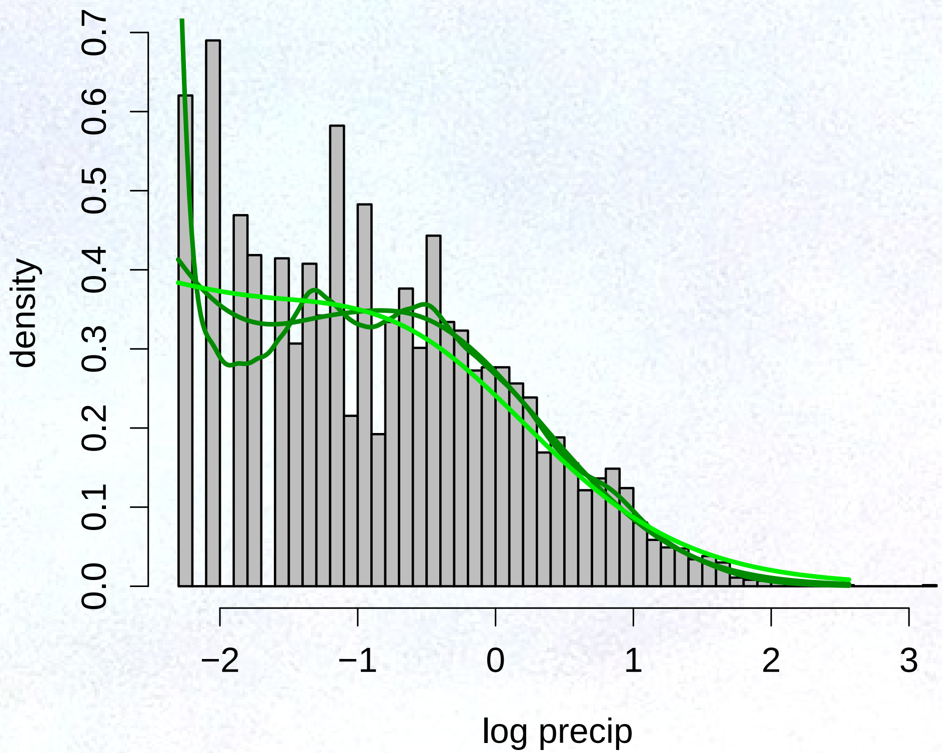
$$min_g \left(\sum_{j=1}^{N} -e^{(g_j)} + \boldsymbol{y}_j g_j - log(\boldsymbol{y}_j!) \right) - \lambda \left(\int_{[x_1, x_N]} (g''(x))^2 dx \right)$$

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

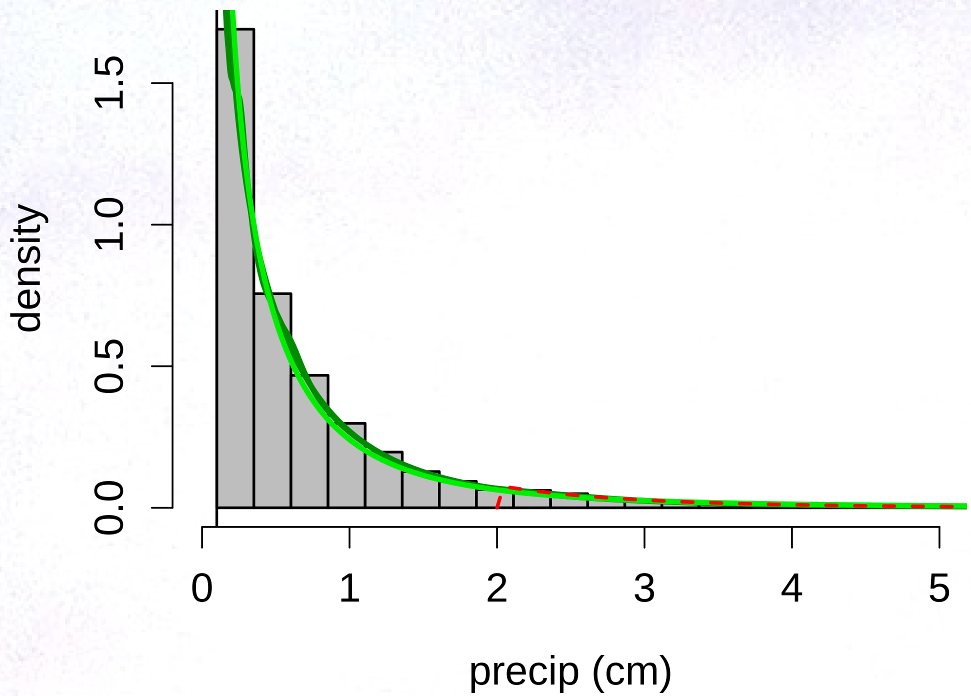
Fit to Boulder data

Three different smoothing parameters:

Log scale

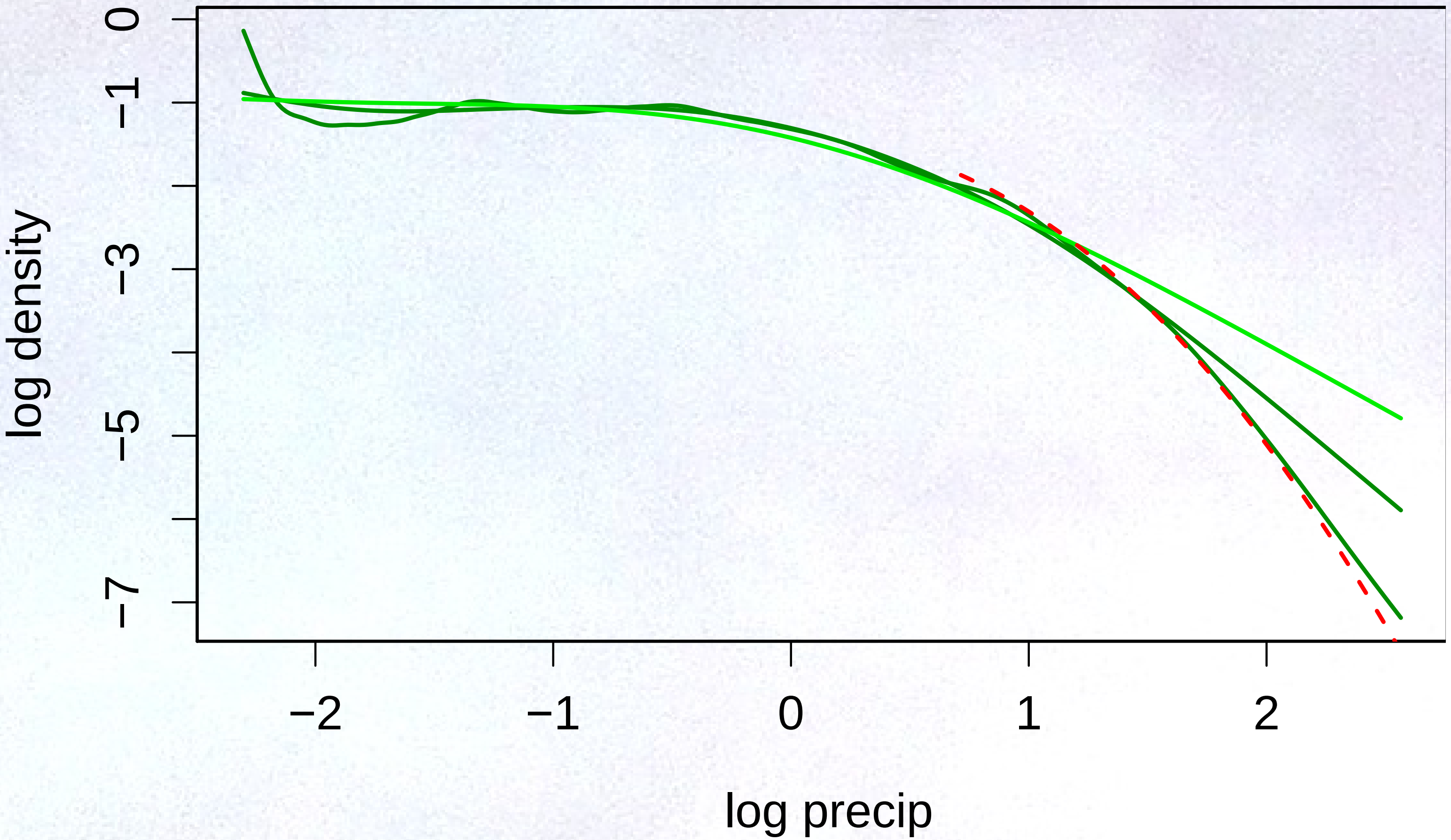


Raw scale:



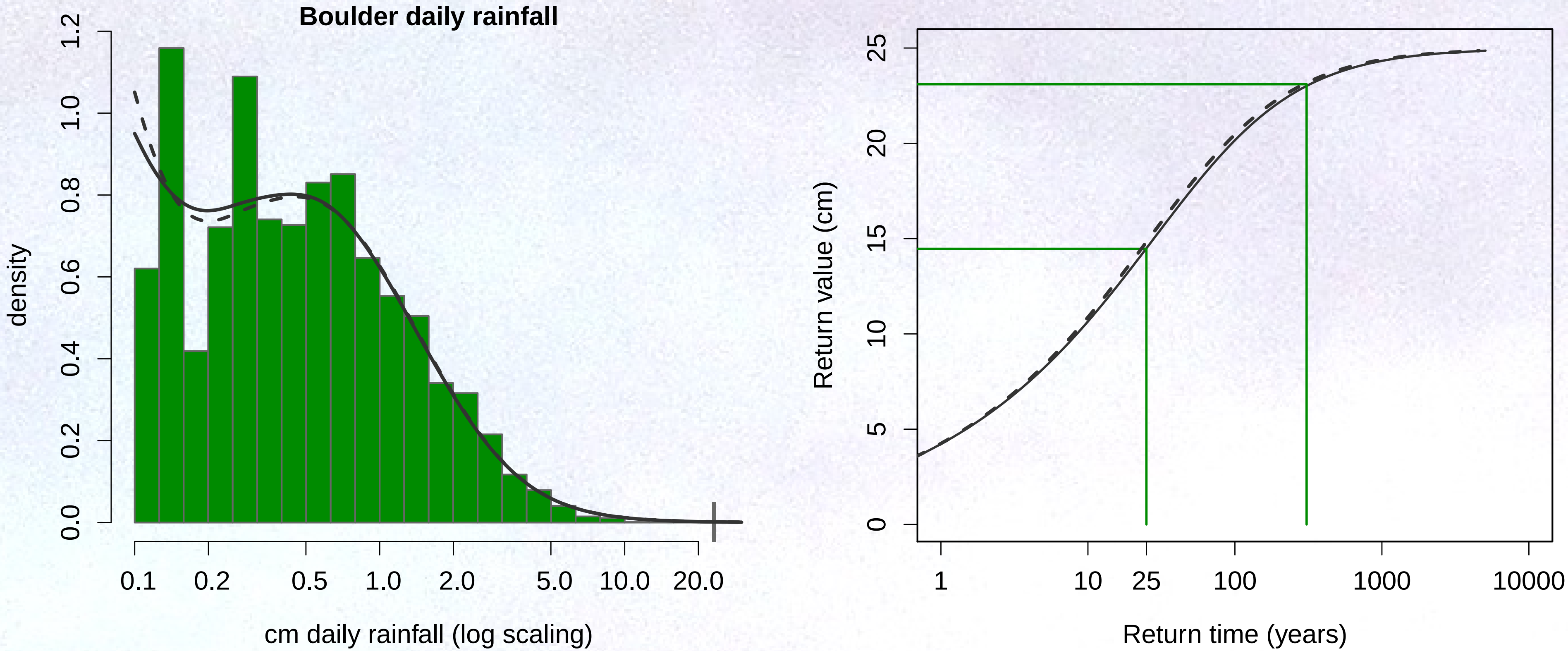
Cross validation choice for λ is effected by discretization at small precipitation amounts.

log densities



log spline rough , log spline smooth, Generalized Pareto

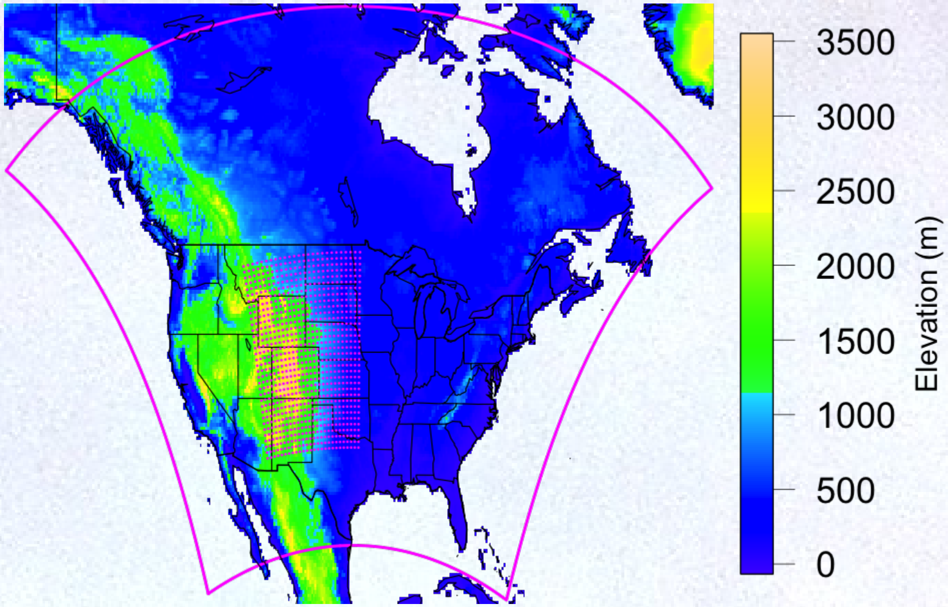
Return Levels



25 year return = quantile for $p = 1 - 1/(365 * 25)$

PART 3: Back to NARCCAP

Back to NARCCAP

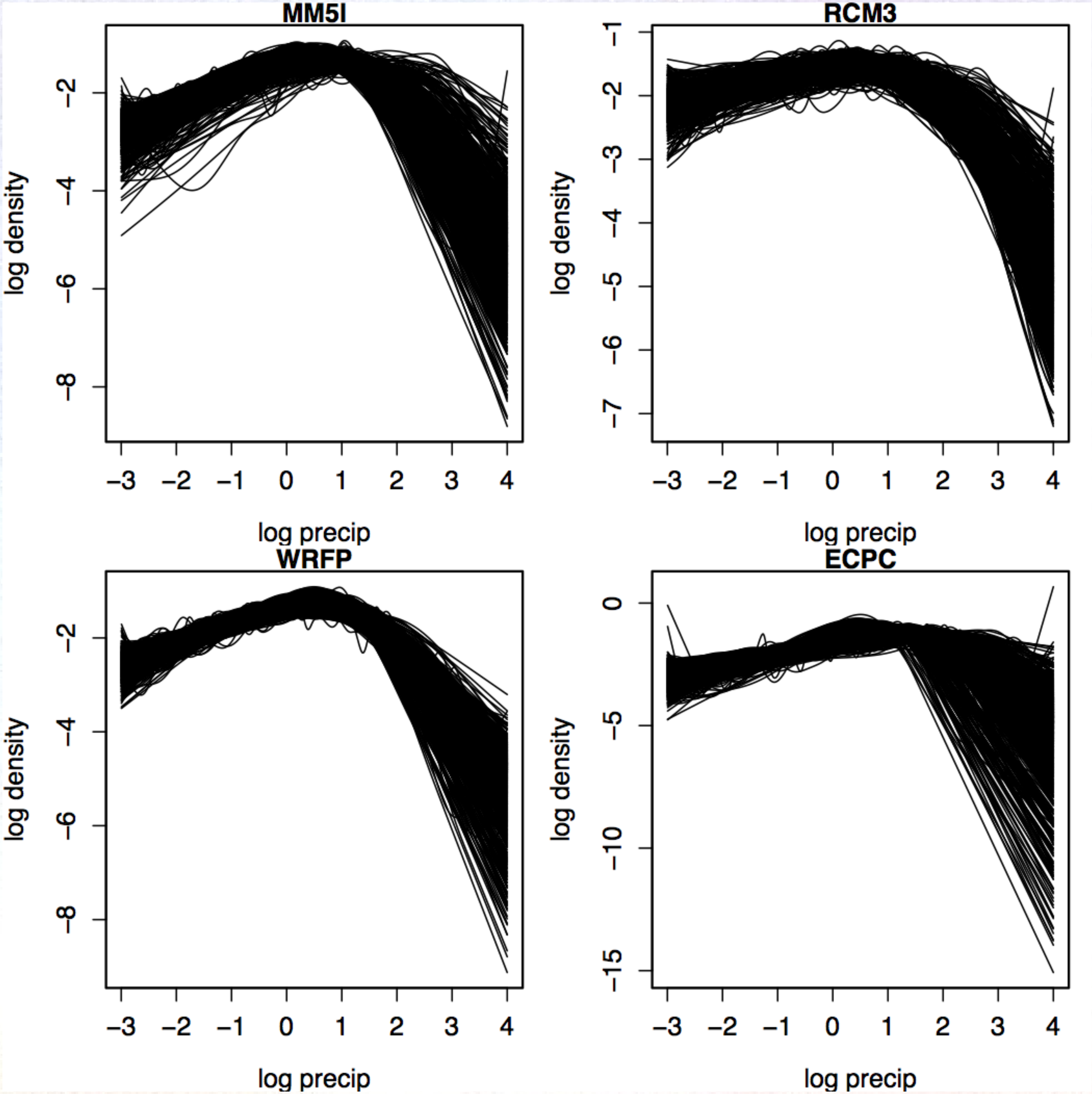


• Four regional models (MM5I, RCM3, WRFP, ECPC) that are driven by observed atmosphere at the boundaries of the NARCCAP domain.

• 20 years of daily downscaled weather about 800 grid points for each model.

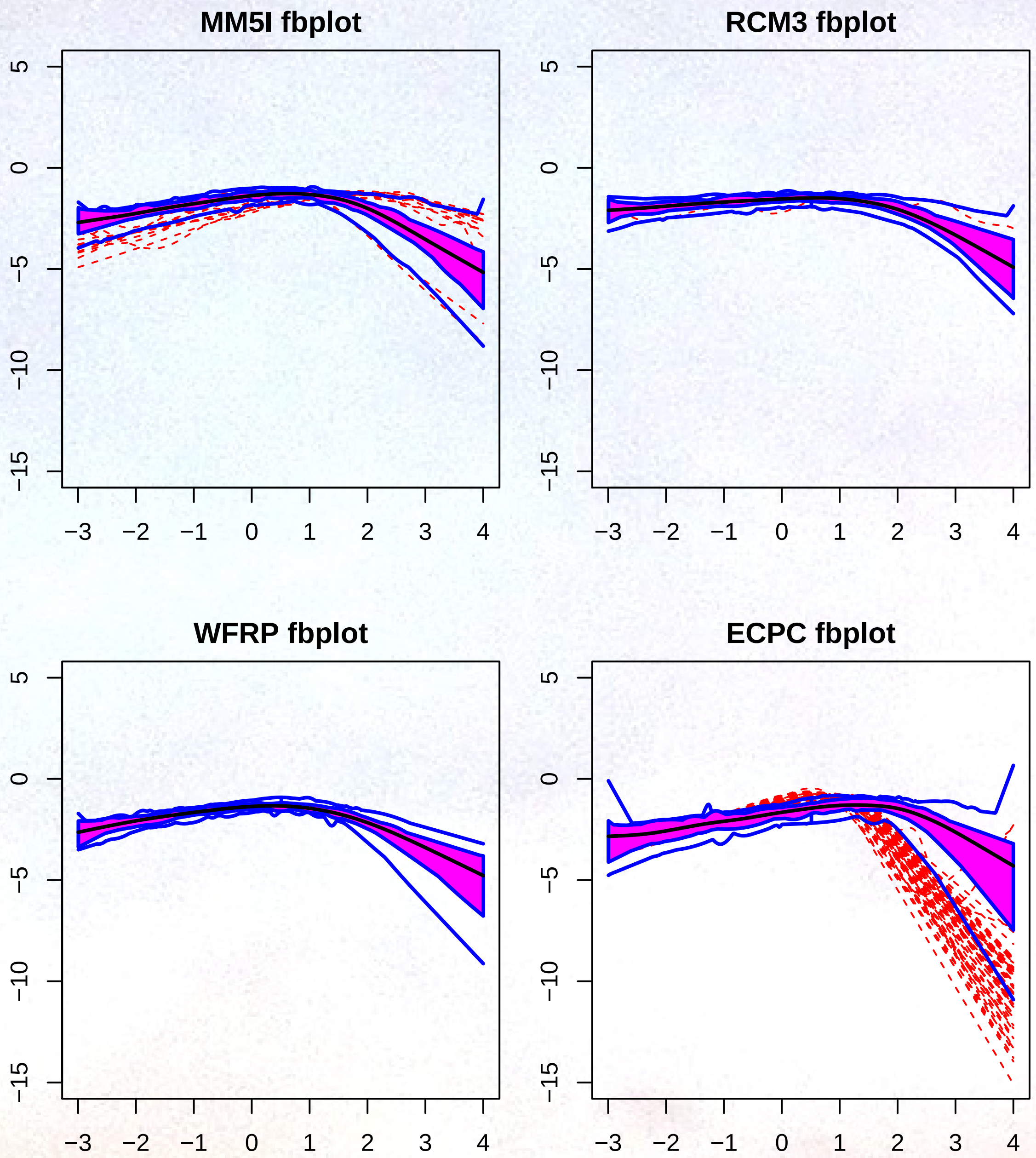
How do extremes of daily summer rainfall vary over space and and over climate models?

Fitted log spline densities



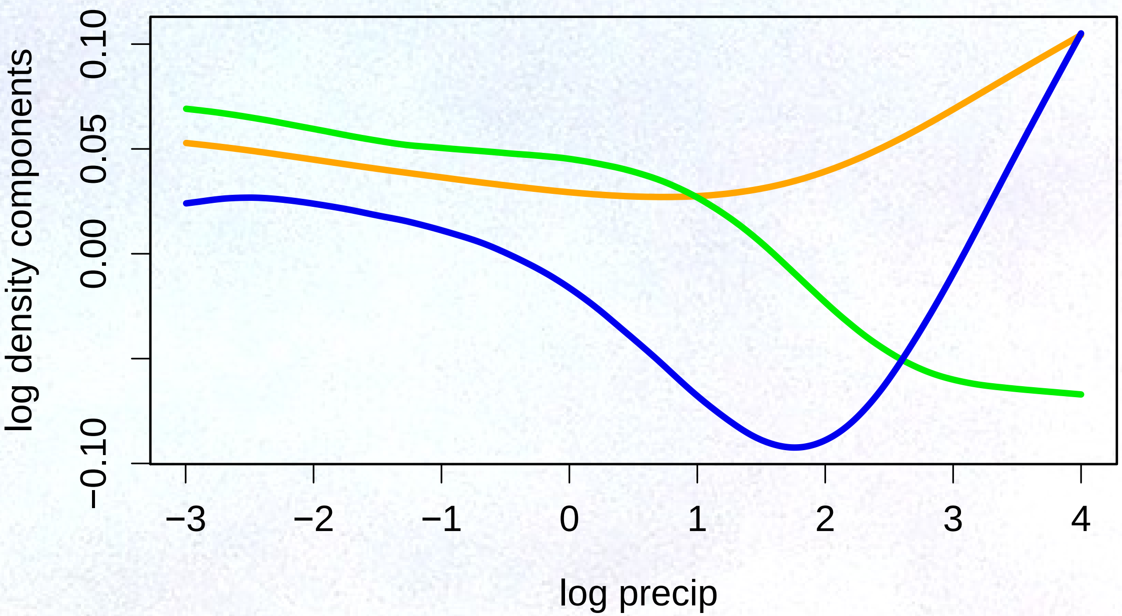
Functional boxplots

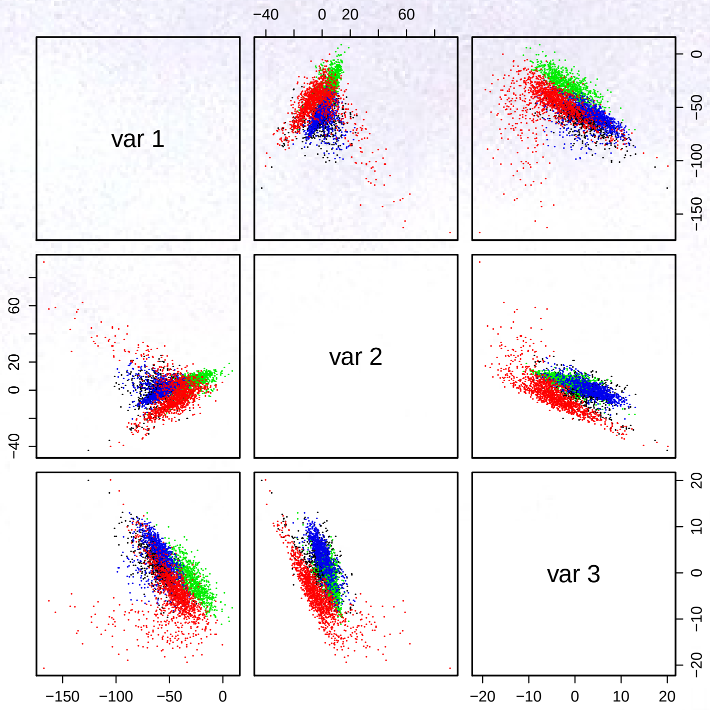
See Yang and Genton (2011)



Principle components

First three principle components of log densities





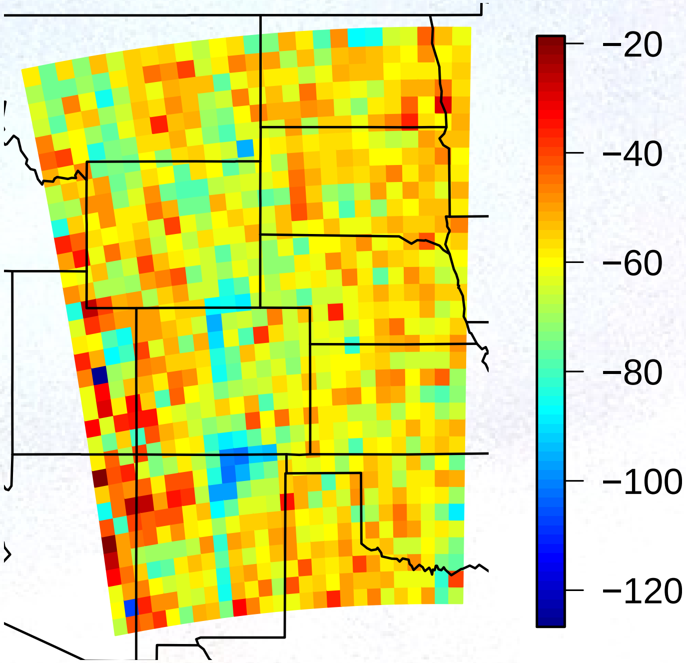
Use these as basis functions to refit models using standard GLM maximum likelihood

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

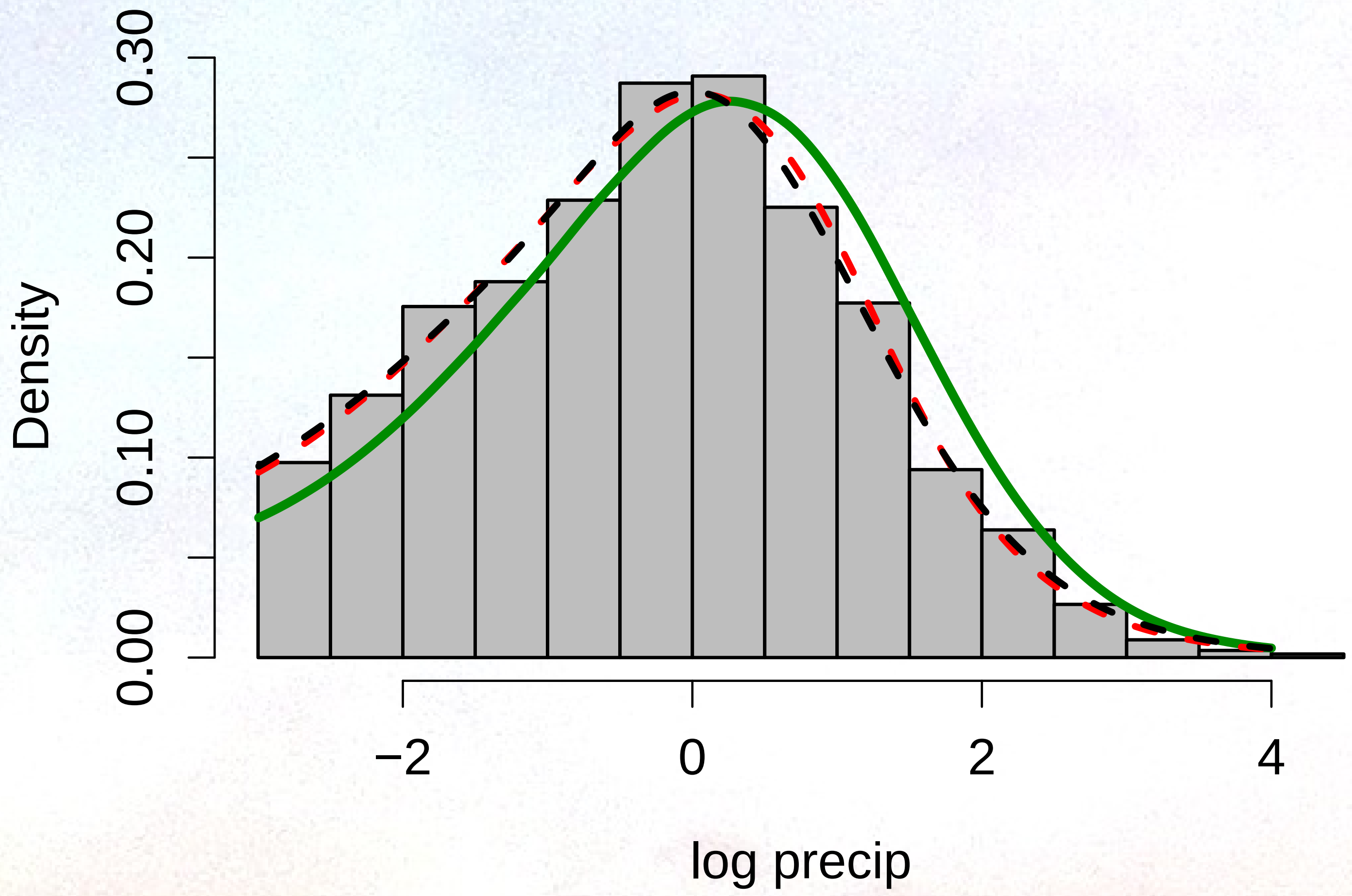


- Transform each climate models coefficients to be uncorrelated.

- Smooth transformed coefficients using spatial statistics. (Approximate thin plate spline.)

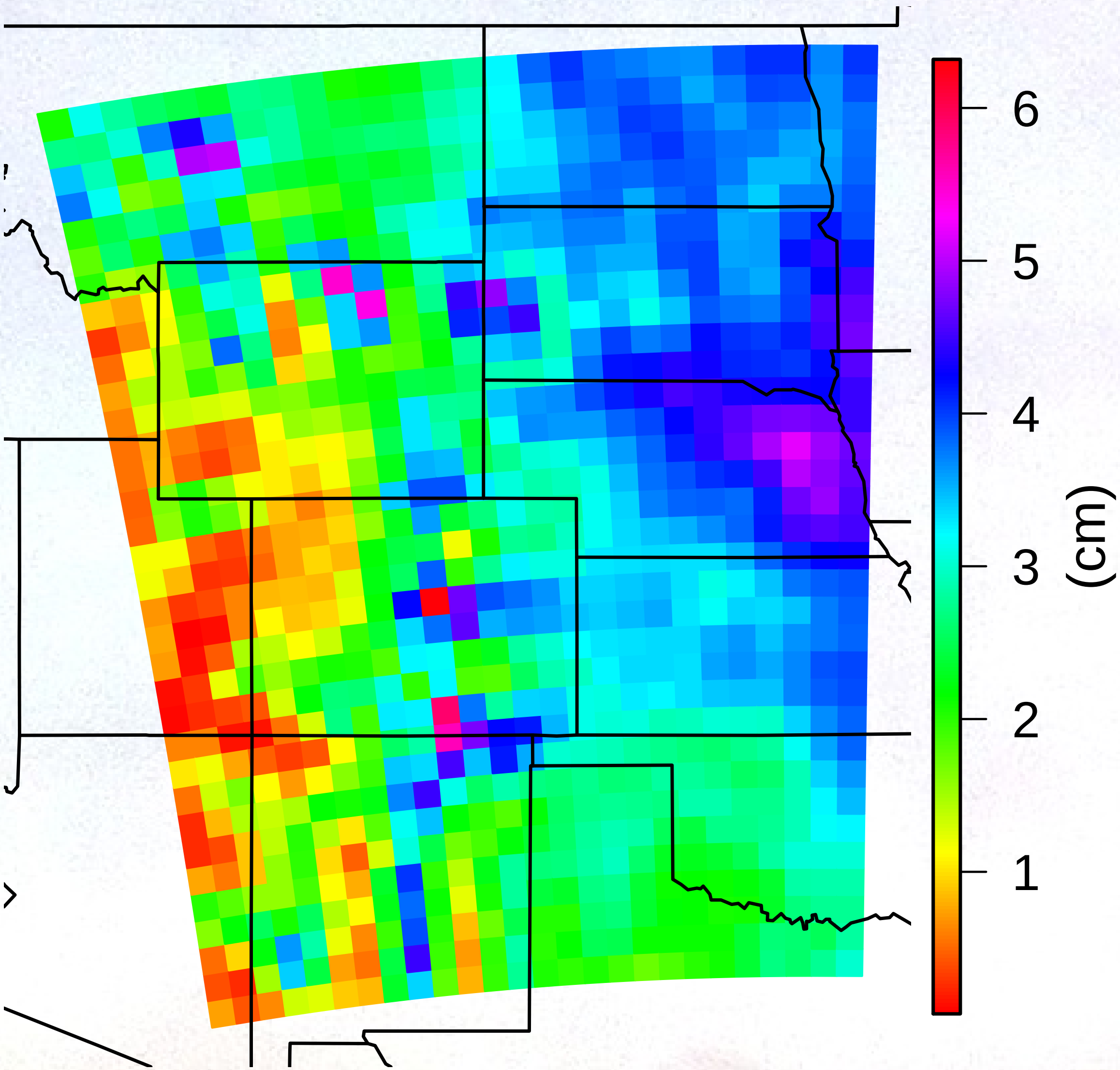
Reconstucting the Boulder grid box

log spline , GLM with 3 basis functions,
smoothed coefficients

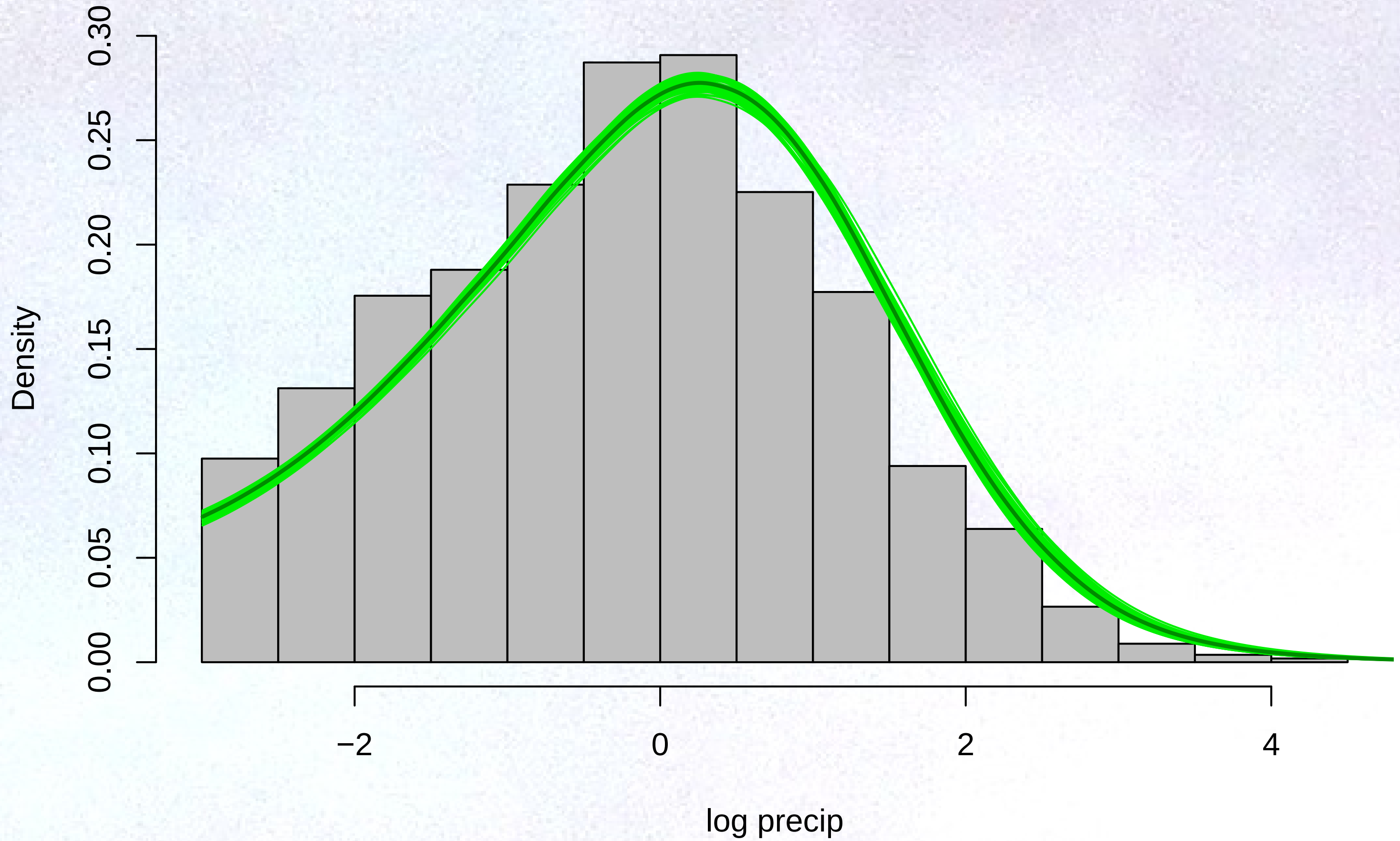


25 year return surface

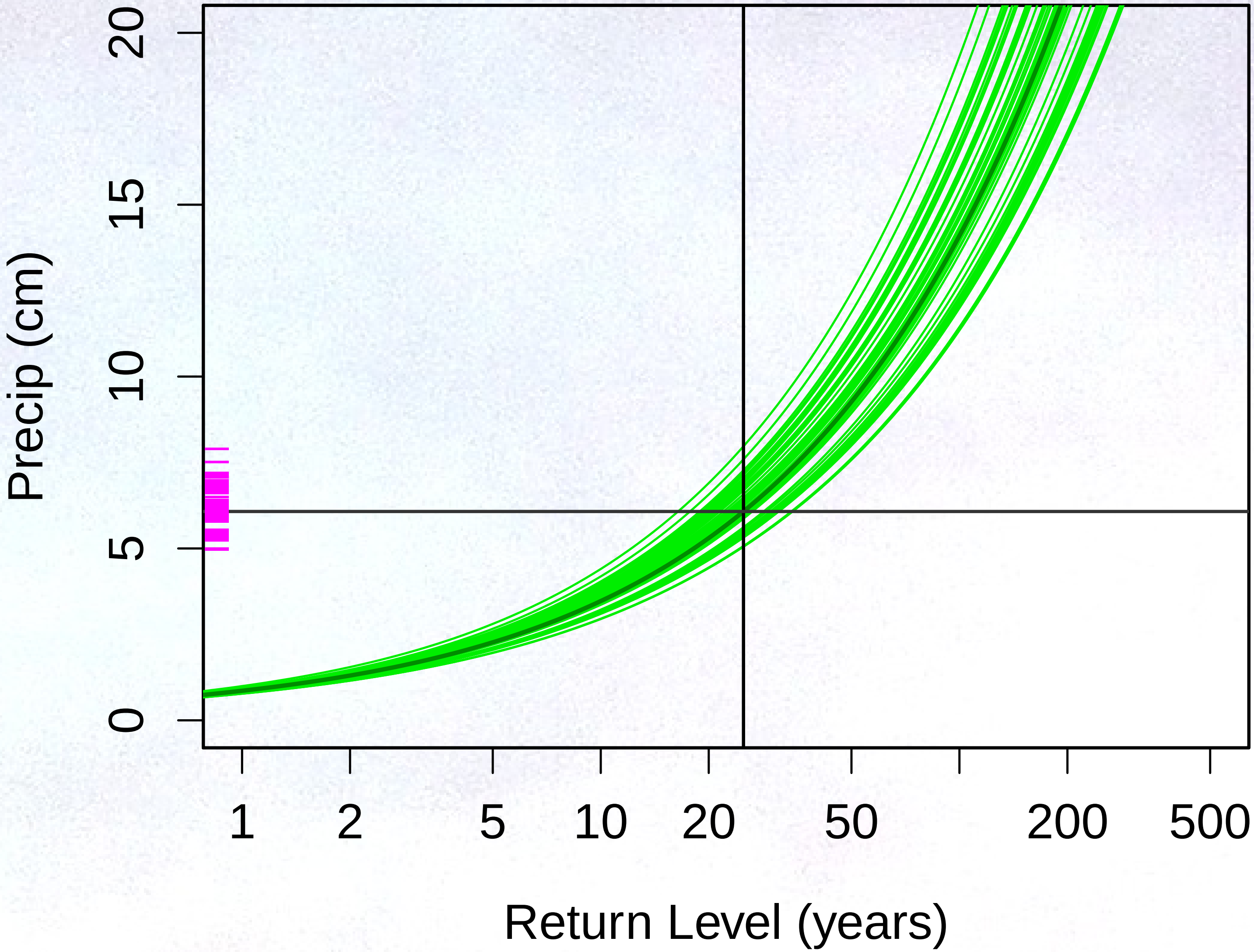
"posterior mode" for MM5I model.



Uncertainty



Boulder grid box 25 year return



Summary

- Statistical methods for estimating and quantifying uncertainty in the tail behavior of climate distributions.

- These are different from traditional climate statistics and require borrowing strength and dimension reduction to make them work

PART 4:

Large spatial data sets

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

The Yellowstone supercomputer.

≈72K cores = 4536 (nodes) × 16 (cores)
and each core with 2Gb memory
16 Pb parallel file system

- Core-hours are available to the NSF geosciences community with a friendly application process for student allocations.
- *Supports R in both interactive and batch mode.*

The Master R session.

In R ...

```
library(Rmpi)
# Spawn 4 slaves
mpi.spawn.Rslaves(nslaves=4)
 # Broadcast the function to all slaves

mpi.bcast.Robj2slave(lambdaKrig)
 # apply this function to 100 tasks (each slave will get about 25)

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

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

Are many R slave processes are feasible?

- `Rmpi` used to initiative many parallel, slaved R sessions from within a master R session.
- Time to initiate 100 - 1000 slaves nearly constant at 3 seconds
- Slaves lose little time reading common data files.
- Median execution time of task per slave is nearly constant.

Is a standard spatial analysis possible?

Embarrassing parallel steps:

Parameter estimation: Searching parameter space to maximize a likelihood or minimize cross validation mean square error.

Computing prediction error: Monte Carlo sampling from the error distribution (a.k.a. conditional simulation).

Iterate between spatial fitting and temporal fitting for space-time data

Thank you!