EXTREME PRECIPITATION TRENDS OVER THE CONTINENTAL UNITED STATES

Richard L. Smith, Amy M. Grady and Gabriele C. Hegerl

Climate Change and Extreme Value Theory
EURANDOM and KNMI, The Netherlands
May 11, 2009

Acknowledgements to NCAR, NCDC, NSF, NOAA, EPA, NISS

STATEMENT OF PROBLEM

During the past decade, there has been extensive research by climatologists documenting increases in the levels of extreme precipitation, both in observational and model-generated data. Groisman *et al.* (*Journal of Climate*, 2005) have a comprehensive review of this whole field.

With a few exceptions (papers by Katz, Zwiers and co-authors) this literature have not made use of the extreme value distributions and related constructs

There are however a few papers by statisticians that have explored the possibility of using more advanced extreme value methods (e.g. Cooley, Naveau and Nychka, *JASA* 2007; Sang and Gelfand, *Env. Ecol. Stat.* 2008)

In this paper, we explore systematically the development of extreme value and spatial models for this problem

CURRENT METHODOLOGY

(see e.g. Groisman *et al.* 2005)

Most common method is based on counting exceedances over a high threshold (e.g. 99.7% threshold)

- Express exceedance counts as anomalies from 30-year mean at each station
- Average over regions using "geometric weighting rule": first average within 1^{o} grid boxes, then average grid boxes within region
- Calculate standard error of this procedure using exponential spatial covariances with nugget (range of 30–500 km, nugget-sill ratio of 0–0.7).
- Increasing trends found in many parts of the world, nearly always stronger than trends in precipitation means, but spatially and temporally heterogeneous. Strongest increase in US extreme precipitations is post-1970, about 7% overall

DATA SOURCES

- NCDC Rain Gauge Data (Groisman 2000)
  - Daily precipitation from 5873 stations
  - Select 1970–1999 as period of study
  - 90% data coverage provision — 4939 stations meet that

- NCAR-CCSM climate model runs
  - 20×41 grid cells of side 1.4^{o}
  - 1970–1999 and 2070–2099 (A2 scenario)

- PRISM data
  - 1405×621 grid, side 4km
  - Elevations
  - Mean annual precipitation 1970–1997

EXTREME VALUE DISTRIBUTIONS

Suppose $X_1, X_2, ...,$ are independent random variables with the same probability distribution, and let $M_n = \max(X_1, ..., X_n)$. Under certain circumstances, it can be shown that there exist *normalizing constants* $a_n > 0, b_n$ such that

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq x\right\} = F(a_n x + b_n)^n \to H(x).$$

The *Three Types Theorem* (Fisher-Tippett, Gnedenko) asserts that if nondegenerate H exists, it must be one of three types:

$$H(x) = \exp(-e^{-x}), \text{ all } x \quad \text{(Gumbel)}$$

$$H(x) = \begin{cases} 0 & x < 0 \\ \exp(-x^{-\alpha}) & x > 0 \end{cases} \quad \text{(Fréchet)}$$

$$H(x) = \begin{cases} \exp(-|x|^{\alpha}) & x < 0 \\ 1 & x > 0 \end{cases} \quad \text{(Weibull)}$$

In Fréchet and Weibull, $\alpha > 0$.

The three types may be combined into a single *generalized extreme value* (GEV) distribution:

$$H(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\psi}\right)_+^{-1/\xi}\right\},$$

$(y_+ = \max(y, 0))$

where μ is a location parameter, $\psi > 0$ is a scale parameter and ξ is a shape parameter. $\xi \to 0$ corresponds to the Gumbel distribution, $\xi > 0$ to the Fréchet distribution with $\alpha = 1/\xi$, $\xi < 0$ to the Weibull distribution with $\alpha = -1/\xi$.

$\xi > 0$: "long-tailed" case, $1 - F(x) \propto x^{-1/\xi}$,

$\xi = 0$: "exponential tail"

$\xi < 0$: "short-tailed" case, finite endpoint at $\mu - \xi/\psi$

EXCEEDANCES OVER THRESHOLDS

Consider the distribution of X conditionally on exceeding some high threshold u:

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)}.$$

As $u \to \omega_F = \sup\{x : \ F(x) < 1\}$, often find a limit

$$F_u(y) \approx G(y; \sigma_u, \xi)$$

where G is *generalized Pareto distribution* (GPD)

$$G(y; \sigma, \xi) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)_+^{-1/\xi}.$$

The Generalized Pareto Distribution

$$G(y;\sigma,\xi) = 1 - \left(1 + \xi\frac{y}{\sigma}\right)_+^{-1/\xi}.$$

$\xi > 0$: long-tailed (equivalent to usual Pareto distribution), tail like $x^{-1/\xi}$,

$\xi = 0$: take limit as $\xi \to 0$ to get

$$G(y;\sigma,0) = 1 - \exp\left(-\frac{y}{\sigma}\right),$$

i.e. exponential distribution with mean σ,

$\xi < 0$: finite upper endpoint at $-\sigma/\xi$.

The *Poisson-GPD model* combines the GPD for the excesses over the threshold with a Poisson distribtion for the number of exceedances. Usually the mean of the Poisson distribution is taken to be λ per unit time.

POINT PROCESS APPROACH

Homogeneous case:

Exceedance $y > u$ at time t has probability

$$\frac{1}{\psi}\left(1+\xi\frac{y-\mu}{\psi}\right)_+^{-1/\xi-1}\exp\left\{-\left(1+\xi\frac{u-\mu}{\psi}\right)_+^{-1/\xi}\right\}dydt$$

- $\mu,\ \psi,\ \xi$ are GEV parameters for annual maxima
- N-year return value — the level y_N that is exceeded in any one year with probability $\frac{1}{N}$.

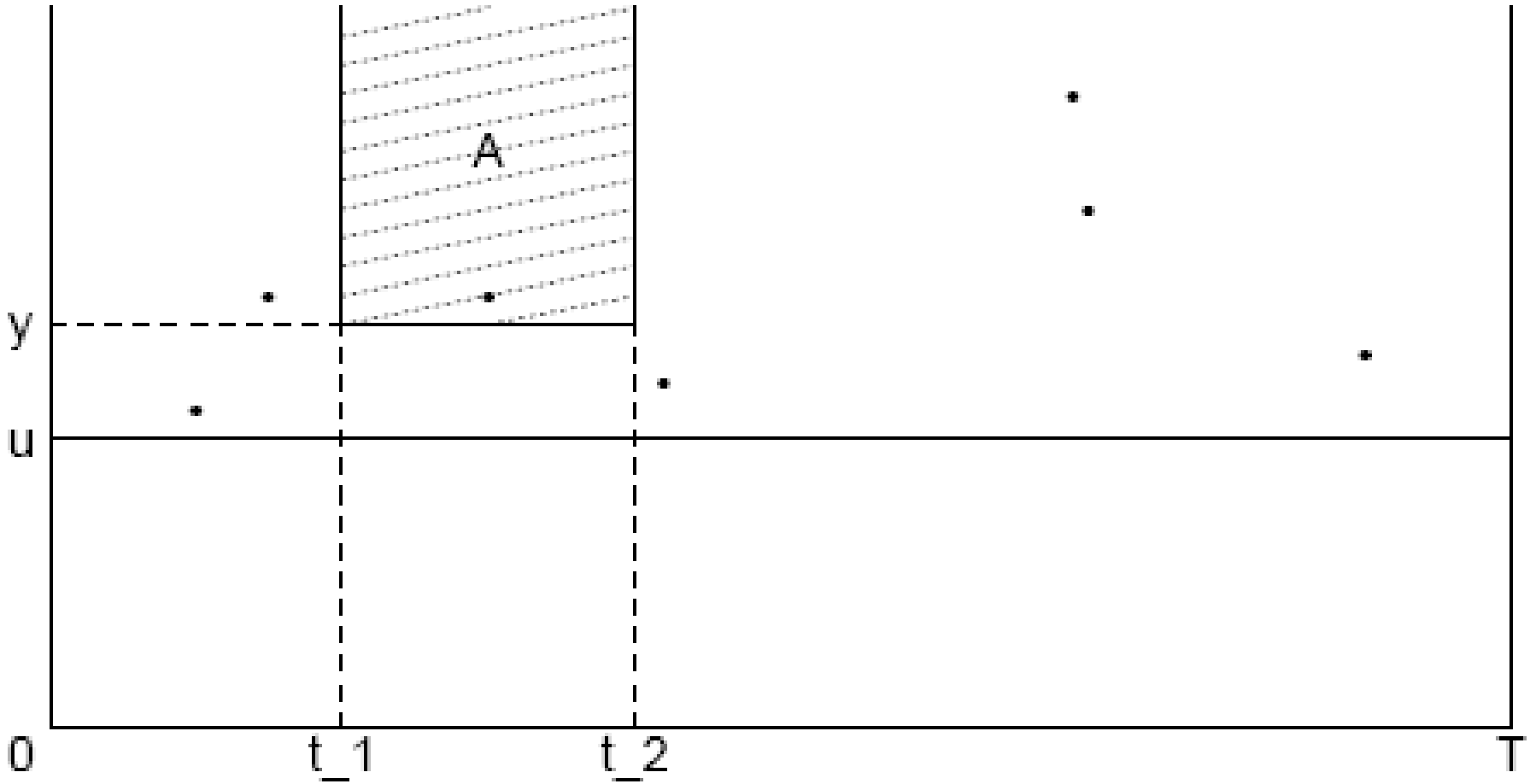


Illustration of point process model.

Inhomogeneous case:

- Time-dependent threshold u_t and parameters $\mu_t,\ \psi_t,\ \xi_t$
- Exceedance $y > u_t$ at time t has probability

$$\frac{1}{\psi_t}\left(1+\xi_t\frac{y-\mu_t}{\psi_t}\right)_+^{-1/\xi_t-1}\exp\left\{-\left(1+\xi_t\frac{u_t-\mu_t}{\psi_t}\right)_+^{-1/\xi_t}\right\}dydt$$

- Estimation by maximum likelihood

Seasonal models without trends

General structure:

$$\begin{aligned}
\mu_t &= \theta_{1,1} + \sum_{k=1}^{K_1} \left(\theta_{1,2k} \cos \frac{2\pi kt}{365.25} + \theta_{1,2k+1} \sin \frac{2\pi kt}{365.25} \right), \\
\log \psi_t &= \theta_{2,1} + \sum_{k=1}^{K_2} \left(\theta_{2,2k} \cos \frac{2\pi kt}{365.25} + \theta_{2,2k+1} \sin \frac{2\pi kt}{365.25} \right), \\
\xi_t &= \theta_{3,1} + \sum_{k=1}^{K_3} \left(\theta_{3,2k} \cos \frac{2\pi kt}{365.25} + \theta_{3,2k+1} \sin \frac{2\pi kt}{365.25} \right).
\end{aligned}$$

Call this the (K_1, K_2, K_3) model.

Note: This is all for one station. The θ parameters will differ at each station.

Model selection

Use a sequence of likelihood ratio tests

- For each (K_1, K_2, K_3), construct LRT against some (K_1', K_2', K_3'), $K_1' \geq K_1, K_2' \geq K_2, K_3' \geq K_3$ (not all equal) using standard χ^2 distribution theory

- Look at proportion of rejected tests over all stations. If too high, set $K_j = K_j'$ $(j = 1, 2, 3)$ and repeat procedure

- By trial and error, we select $K_1 = 4, K_2 = 2, K_3 = 1$ (17 model parameters for each station)

Models with trend

Add to the above:

- Overall linear trend $\theta_{j,2K+2}t$ added to any of μ_t $(j = 1)$, $\log \psi_t$ $(j = 1)$, ξ_t $(j = 1)$. Define K_j^* to be 1 if this term is included, o.w. 0.

- Interaction terms of form

$$t \cos \frac{2\pi kt}{365.25}, \quad t \sin \frac{2\pi kt}{365.25}, \quad k = 1, ..., K_j^{**}.$$

Typical model denoted

$$(K_1, K_2, K_3) \times (K_1^*, K_2^*, K_3^*) \times (K_1^{**}, K_2^{**}, K_3^{**})$$

Eventually use $(4, 2, 1) \times (1, 1, 0) \times (2, 2, 0)$ model (27 parameters for each station)

Details

- Selection of time-varying threshold — based on the 95th percentile of a 7-day window around the date of interest
- Declustering by r-runs method (Smith and Weissman 1994) — use $r = 1$ for main model runs
- Computation via *sampling the likelihood*: evaluate contributions to likelihood for all observations above threshold, but sample only 5% or 10% of those below, then renormalize to provide accurate approximation to full likelihood

Details (continued)

- Covariances of parameters at different sites:

  $\widehat{\theta}_s$ is MLE at site s, solves $\nabla \ell_s(\widehat{\theta}_s) = 0$

  For two sites s, s',

  $$\mathrm{Cov}\left(\widehat{\theta}_s, \widehat{\theta}_{s'}\right) \approx \left(\nabla^2 \ell_s(\widehat{\theta}_s)\right)^{-1} \mathrm{Cov}\left(\nabla \ell_s(\theta_s), \nabla \ell_{s'}(\theta_{s'})\right) \left(\nabla^2 \ell_{s'}(\widehat{\theta}_{s'})\right)^{-1}$$

  Estimate covariances on RHS empirically, using a subset of days (*same* subset for all stations)

  Also employed when $s = s'$.

  Open question: Should we "regularize" this covariance matrix?

Details (continued)

- Calculating the N-year return value

  For one year ($t = 1, ..., T$), find $y_{\theta,N}$ numerically to solve

$$\sum_{1}^{T} \left(1 + \xi_t \frac{y_{\theta,N} - \mu_t}{\psi_t}\right)^{-1/\xi_t} = -\log\left(1 - \frac{1}{N}\right).$$

  - Also calculate $\frac{\partial y_{\theta,N}}{\partial \theta_j}$ by numerical implementation of inverse function formula
  - Covariances between return level estimates at different sites by

$$\text{Cov}\left\{y_{\widehat{\theta}_s,N}, y_{\widehat{\theta}_{s'},N}\right\} \approx \left(\frac{dy_{\widehat{\theta}_s,N}}{d\theta_s}\right)^T \text{Cov}\left(\widehat{\theta}_s, \widehat{\theta}_{s'}\right) \left(\frac{dy_{\widehat{\theta}_{s'},N}}{d\theta_{s'}}\right).$$

  - Also apply to ratios of return level estimates, such as

$$\frac{25-\text{year return level at } s \text{ in } 1999}{25-\text{year return level at } s \text{ in } 1970}$$

SPATIAL SMOOTHING

Let Z_s be field of interest, indexed by s (typically the logarithm of the 25-year RV at site s, or a log of ratio of RVs. Taking logs improves fit of spatial model, to follow.)

Don't observe Z_s — estimate $\hat{Z}_s$. Assume

$$\begin{aligned} \hat{Z} \mid Z &\sim N[Z, W] \\ Z &\sim N[X\beta, V(\phi)] \\ \hat{Z} &\sim N[X\beta, V(\phi) + W]. \end{aligned}$$

for known W; X are covariates, β are unknown regression parameters and ϕ are parameters of spatial covariance matrix V.

- ϕ by REML
- β given ϕ by GLS
- Predict Z at observed and unobserved sites by kriging

Details

- Covariates
  - Always include intercept
  - Linear and quadratic terms in elevation and log of mean annual precipitation
  - Contrast "climate space" approach of Cooley *et al.* (2007)

Details (continued)

- Spatial covariances
  - Matérn
  - Exponential with nugget
  - Intrinsically stationary model

$$\text{Var}(Z_s - Z_{s'}) \;=\; \phi_1 d_{s,s'}^{\phi_2} + \phi_3$$

  - Matérn with nugget

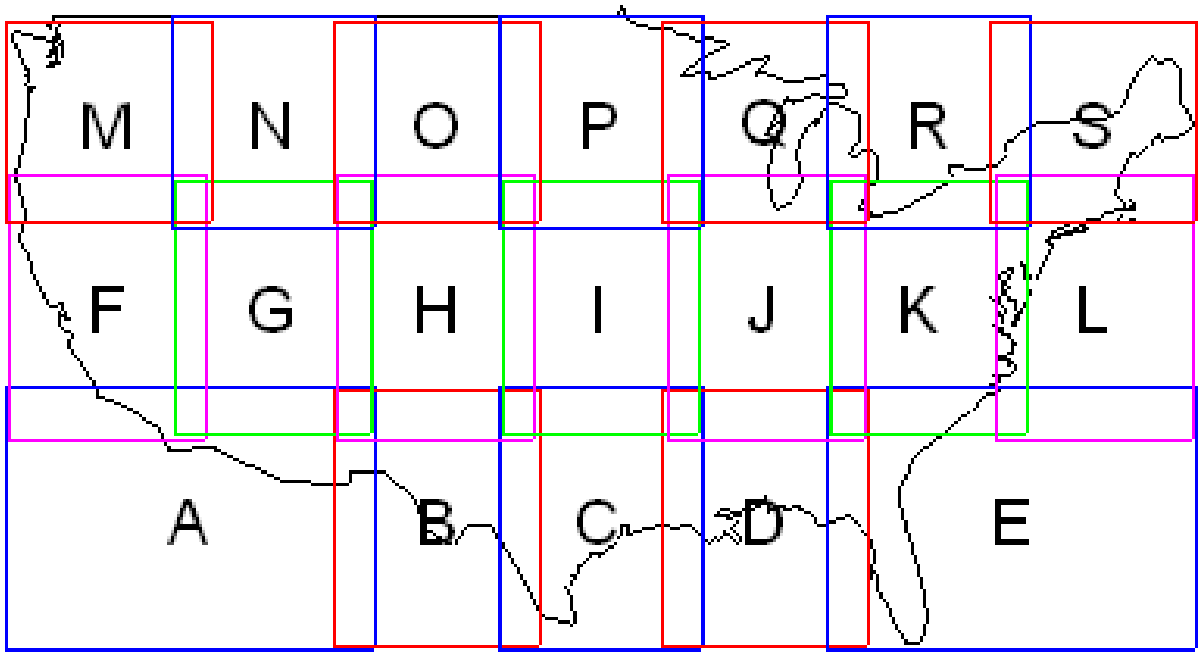
The last-named contains all the previous three as limiting cases and appears to be the best overall, though is often slow to converge (e.g. sometime the range parameter tends to ∞, which is almost equivalent to the intrinsically stationary model)

Details (continued)

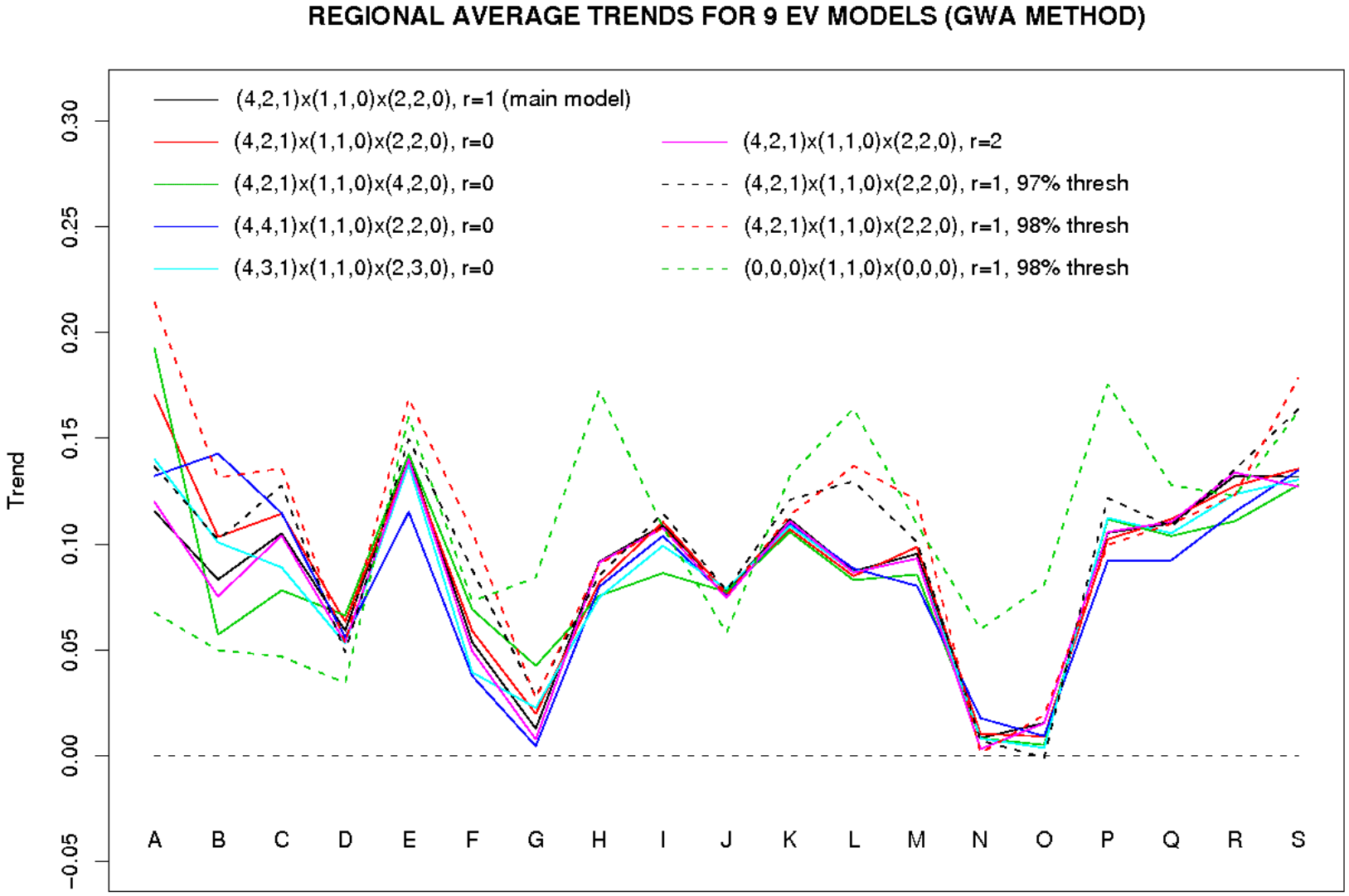
- Spatial heterogeneity

  Divide US into 19 overlapping regions, most $10^o \times 10^o$

  - Kriging within each region
  - Linear smoothing across region boundaries
  - Same for MSPEs
  - Also calculate regional averages, including MSPE



Continental USA divided into 19 regions



Trends across 19 regions (measured as change in log RV25) for 8 different seasonal models and one non-seasonal model with simple linear trends. Regional averaged trends by geometric weighted average approach.

Summary of models shown on previous slide:

1: Preferred covariates model ($r = 0$ for declustering, uses 95% threshold calculated from 7-day window)

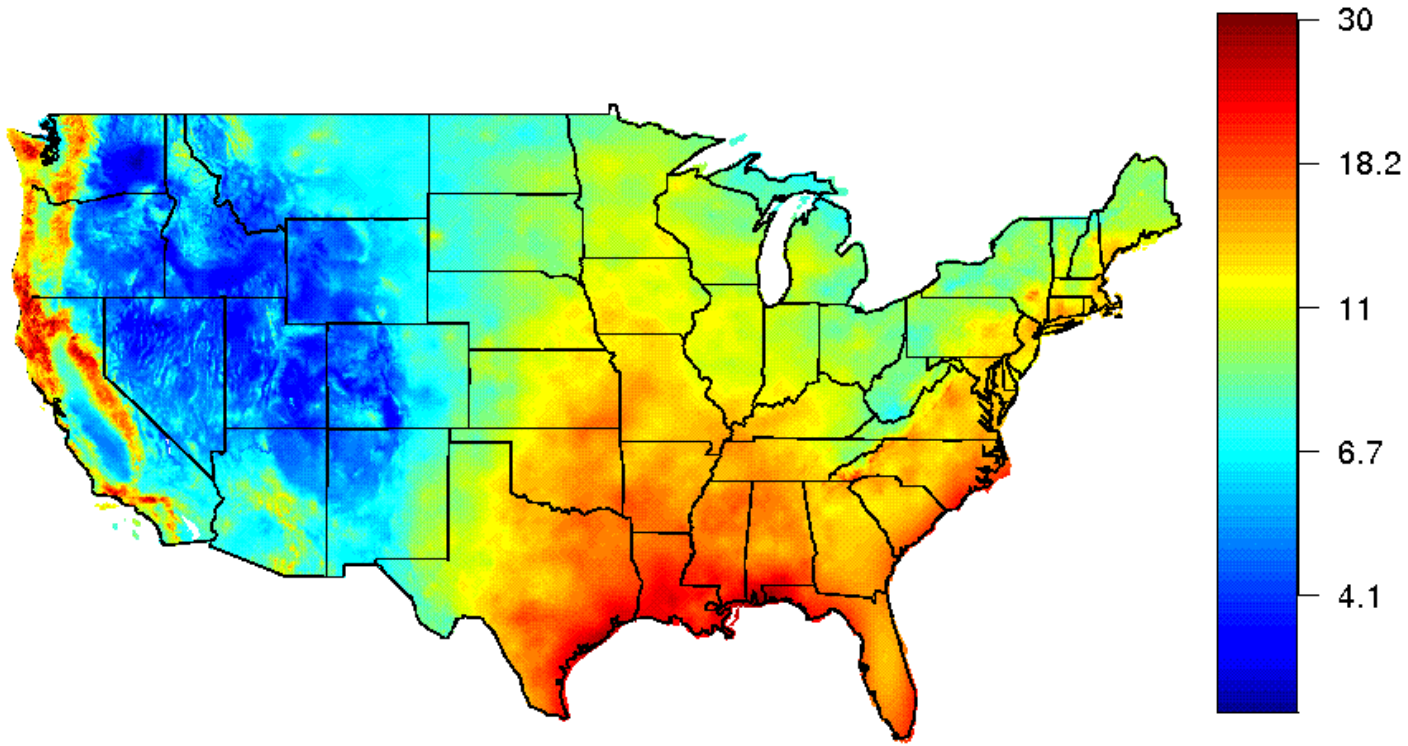
2–4: Three variants where we add covariates to μ_t and/or $\log \psi_t$

5: Replace $r = 0$ by $r = 1$ (subsequent results are based on this model)

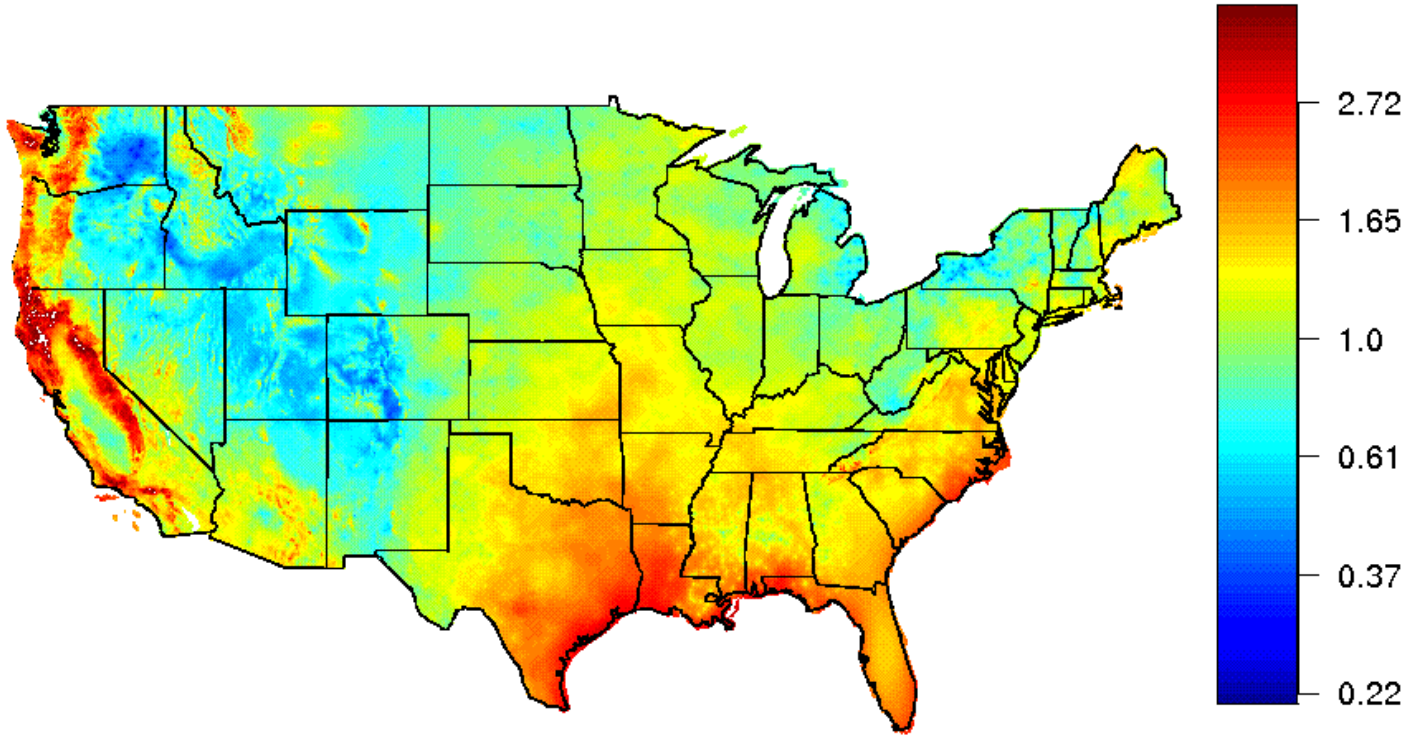
6: Replace $r = 0$ by $r = 2$

7: 97% threshold calculated from 14-day window

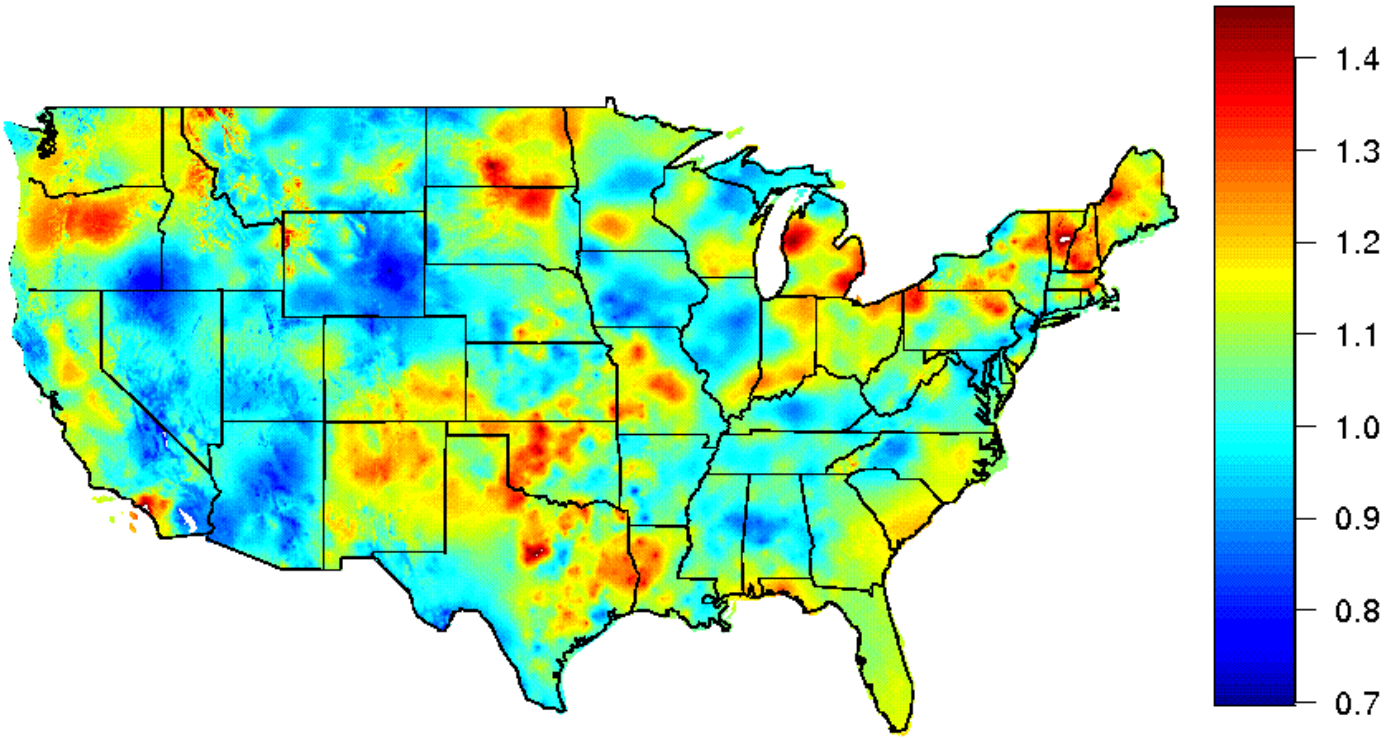
8: 98% threshold calculated from 28-day window



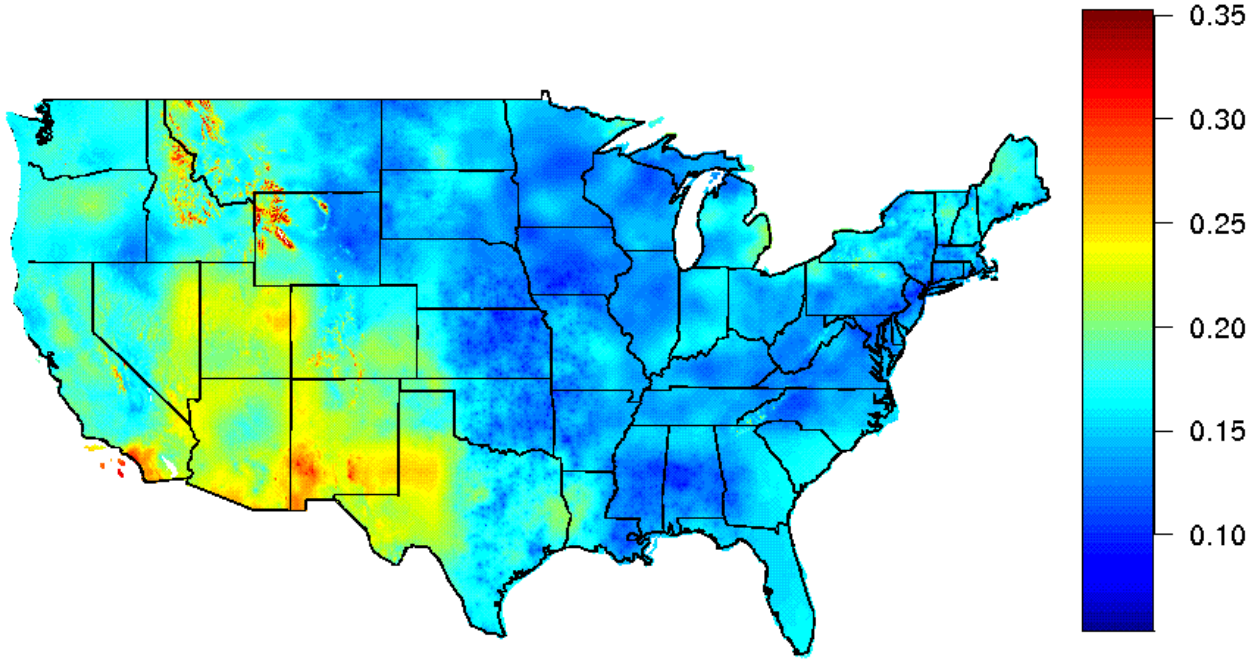
Map of 25-year return values (cm.) for the years 1970–1999



Root mean square prediction errors for map of 25-year return values for 1970–1999



Ratios of return values in 1999 to those in 1970



Root mean square prediction errors for map of ratios of 25-year return values in 1999 to those in 1970

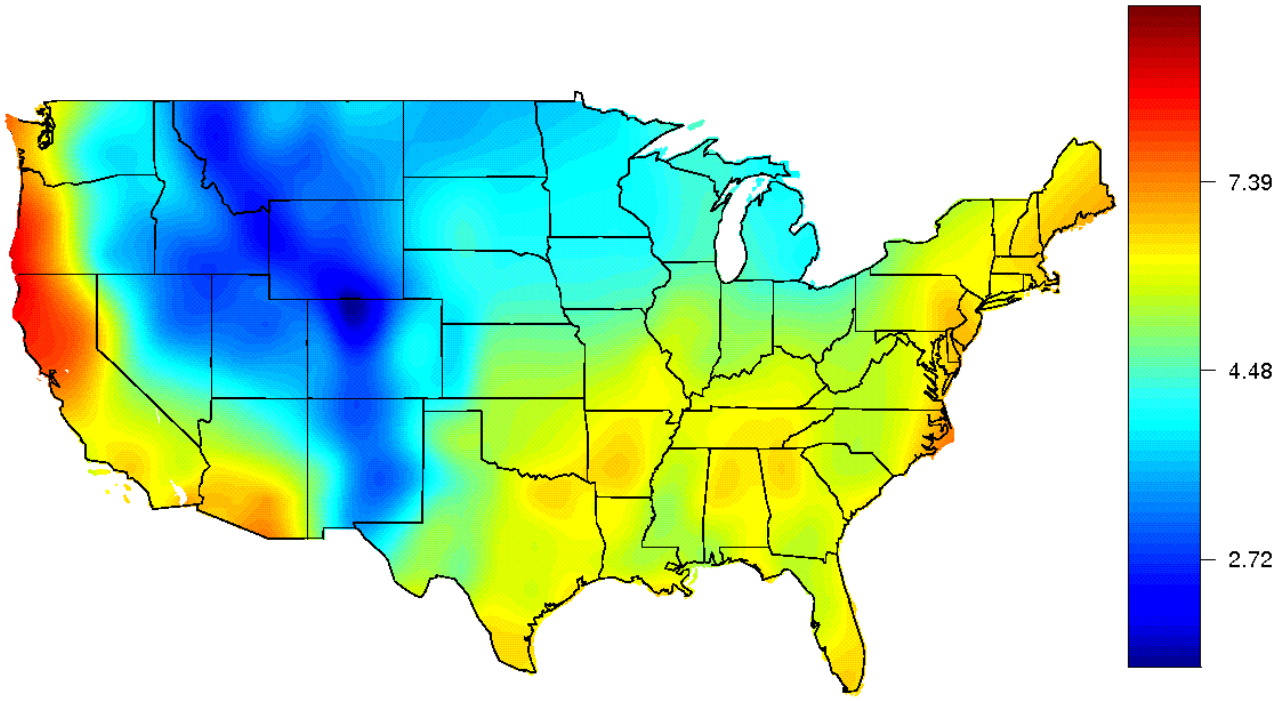
	Δ_1	S_1	Δ_2	S_2		Δ_1	S_1	Δ_2	S_2
A	−0.01	.03	0.05**	.05	K	0.08***	.01	0.09**	.03
B	0.07**	.03	0.08***	.04	L	0.07***	.02	0.07*	.04
C	0.11***	.01	0.10	.03	M	0.07***	.02	0.10**	.03
D	0.05***	.01	0.06	.05	N	0.02	.03	0.01	.03
E	0.13***	.02	0.14*	.05	O	0.01	.02	0.02	.03
F	0.00	.02	0.05*	.04	P	0.07***	.01	0.11***	.03
G	−0.01	.02	0.01	.03	Q	0.07***	.01	0.11***	.03
H	0.08***	.01	0.10***	.03	R	0.15***	.02	0.13***	.03
I	0.07***	.01	0.12***	.03	S	0.14***	.02	0.12*	.06
J	0.05***	.01	0.08**	.03					

Δ_1: Mean change in log 25-year return value (1970 to 1999) by kriging

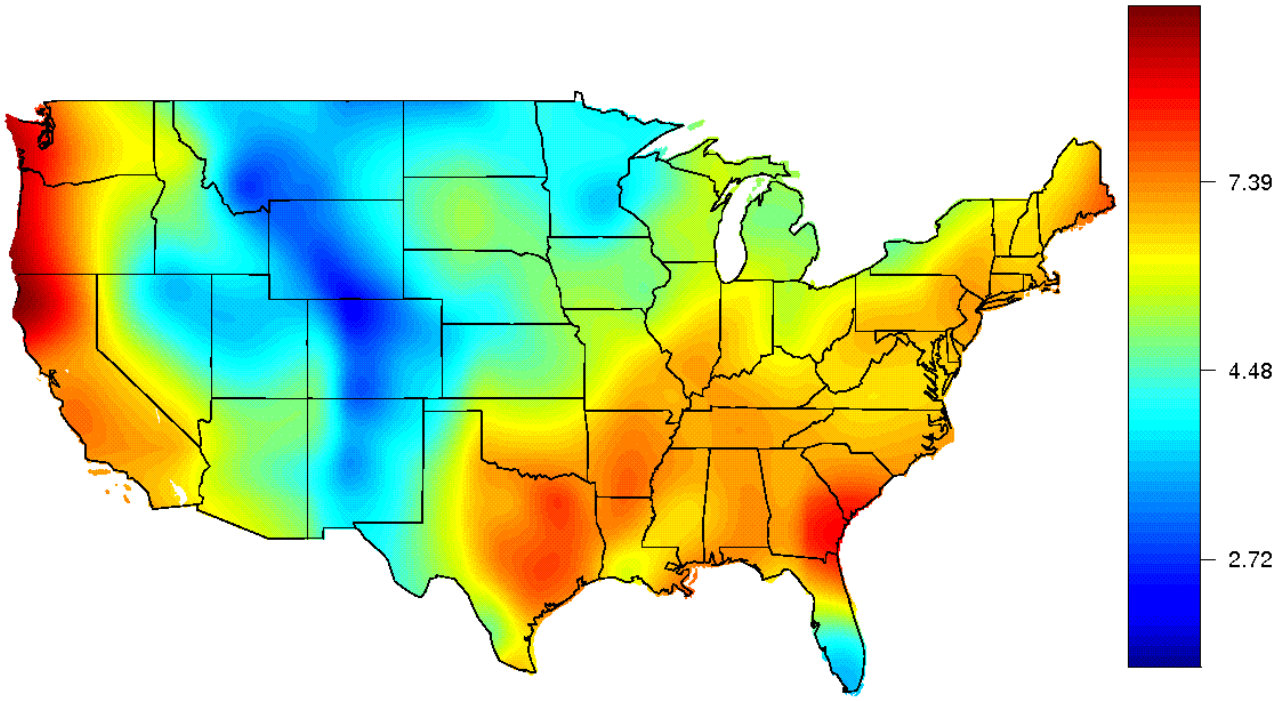
S_1: Corresponding standard error (or RMSPE)

Δ_2, S_2: same but using geometrically weighted average (GWA)

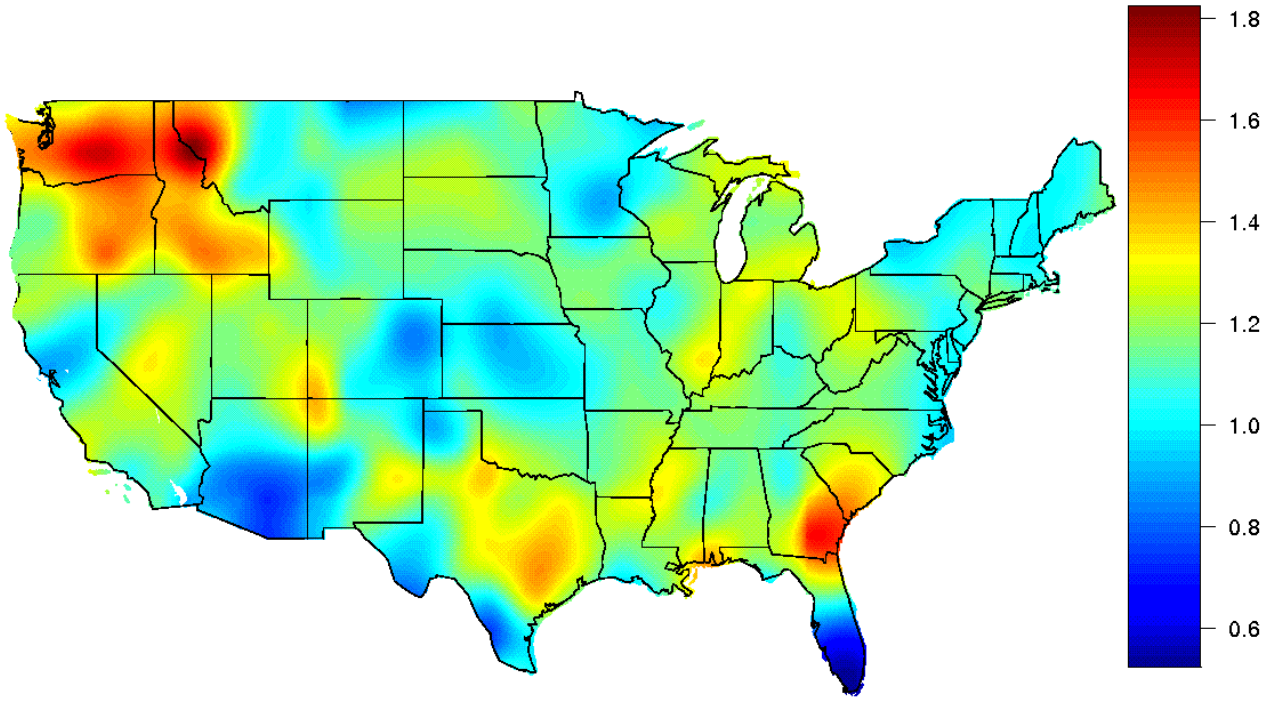
Stars indicate significance at 5%*, 1%**, 0.1%***.



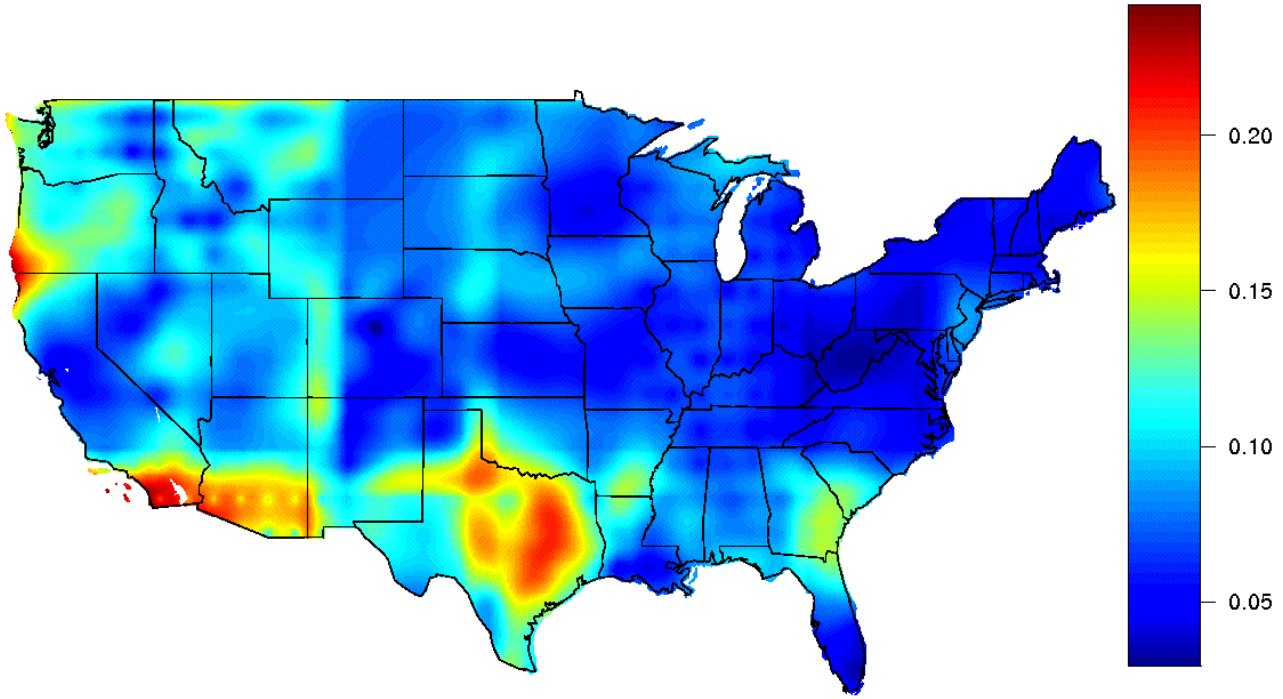
Return value map for CCSM data (cm.): 1970–1999



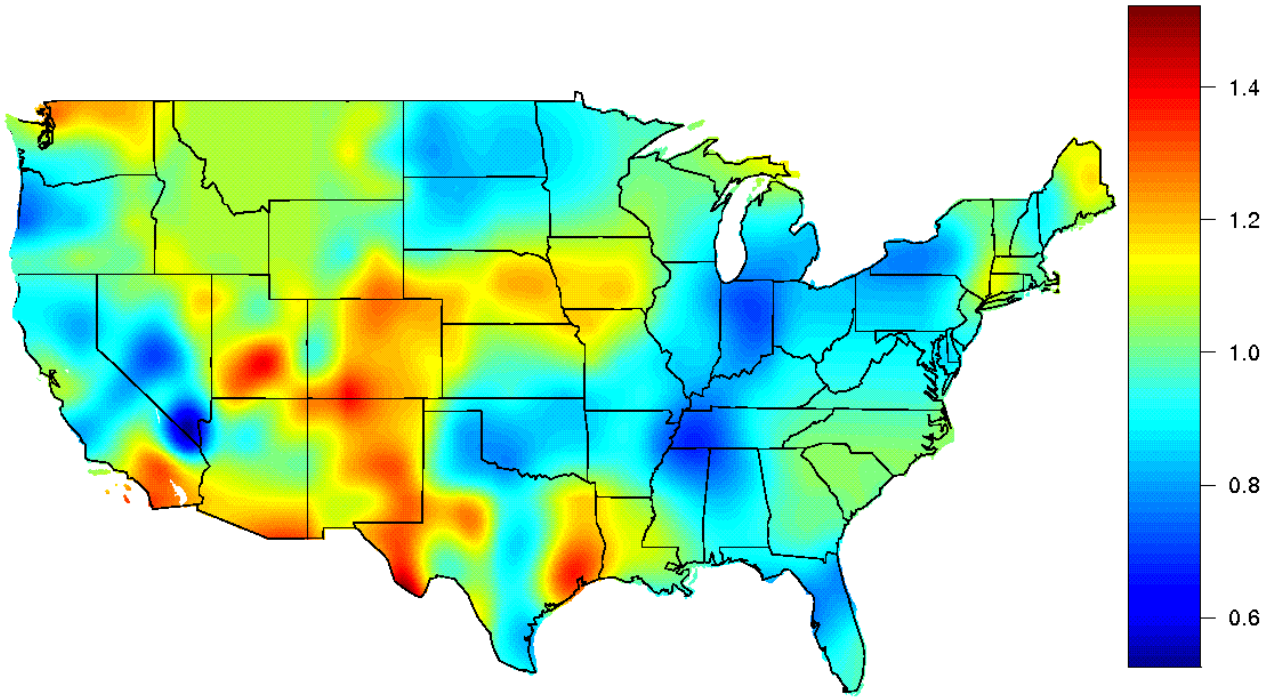
Return value map for CCSM data (cm.): 2070–2099



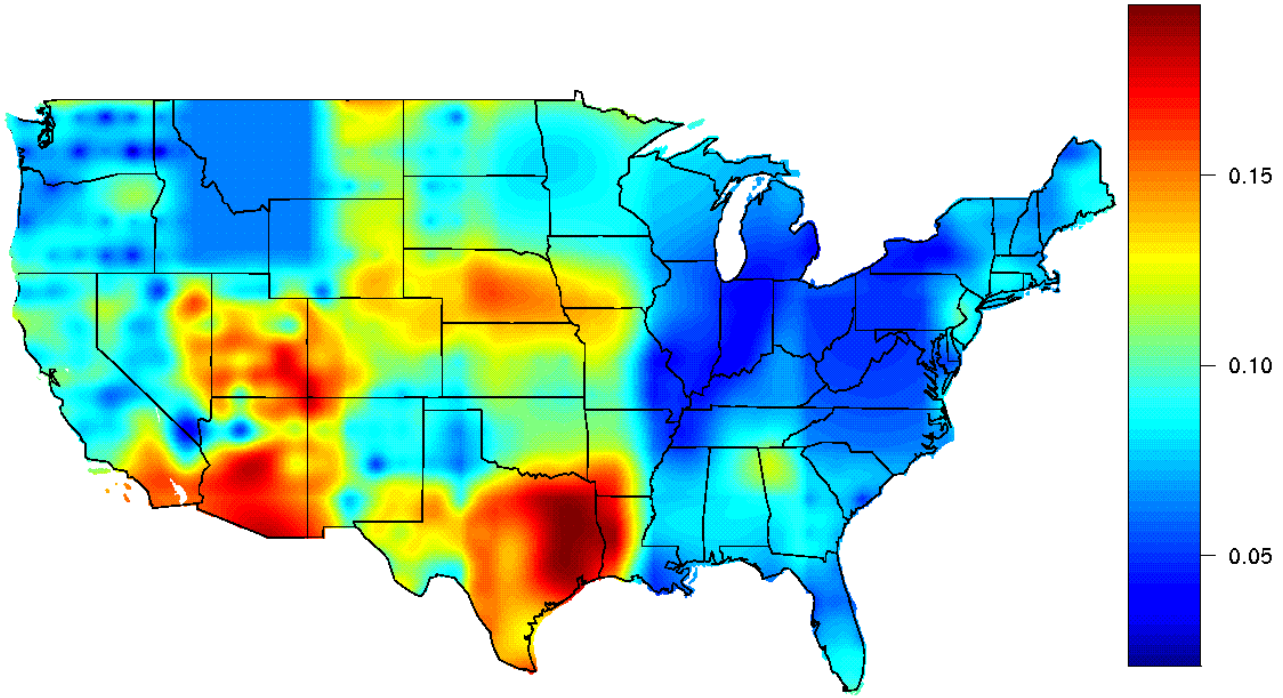
Estimated ratios of 25-year return values for 2070–2099 to those of 1970–1999, based on CCSM data, A2 scenario



RMSPE for map in previous slide



Extreme value model with trend: ratio of 25-year return value in 1999 to 25-year return value in 1970, based on CCSM data



RMSPE for map in previous slide

	Δ_3	S_3	Δ_4	S_4		Δ_3	S_3	Δ_4	S_4
A	0.16^{**}	.07	0.24^{**}	.10	K	-0.08^{***}	.02	-0.11^{*}	.05
B	0.14^{***}	.04	0.12^{***}	.06	L	-0.04	.04	-0.03	.06
C	0.02	.05	-0.14	.11	M	0.01	.03	0.00	.08
D	-0.06	.04	-0.15	.10	N	0.06^{**}	.02	0.05	.06
E	-0.07^{*}	.03	-0.09	.08	O	-0.03	.04	-0.06	.07
F	-0.07^{*}	.04	-0.03	.05	P	-0.01	.04	-0.07	.07
G	0.03	.03	0.08^{*}	.04	Q	-0.04	.04	-0.03	.07
H	0.11^{***}	.03	0.08	.06	R	-0.17^{***}	.03	-0.06	.08
I	-0.02	.04	-0.05	.07	S	0.00	.04	0.02	.05
J	-0.15^{***}	.03	-0.16^{**}	.06					

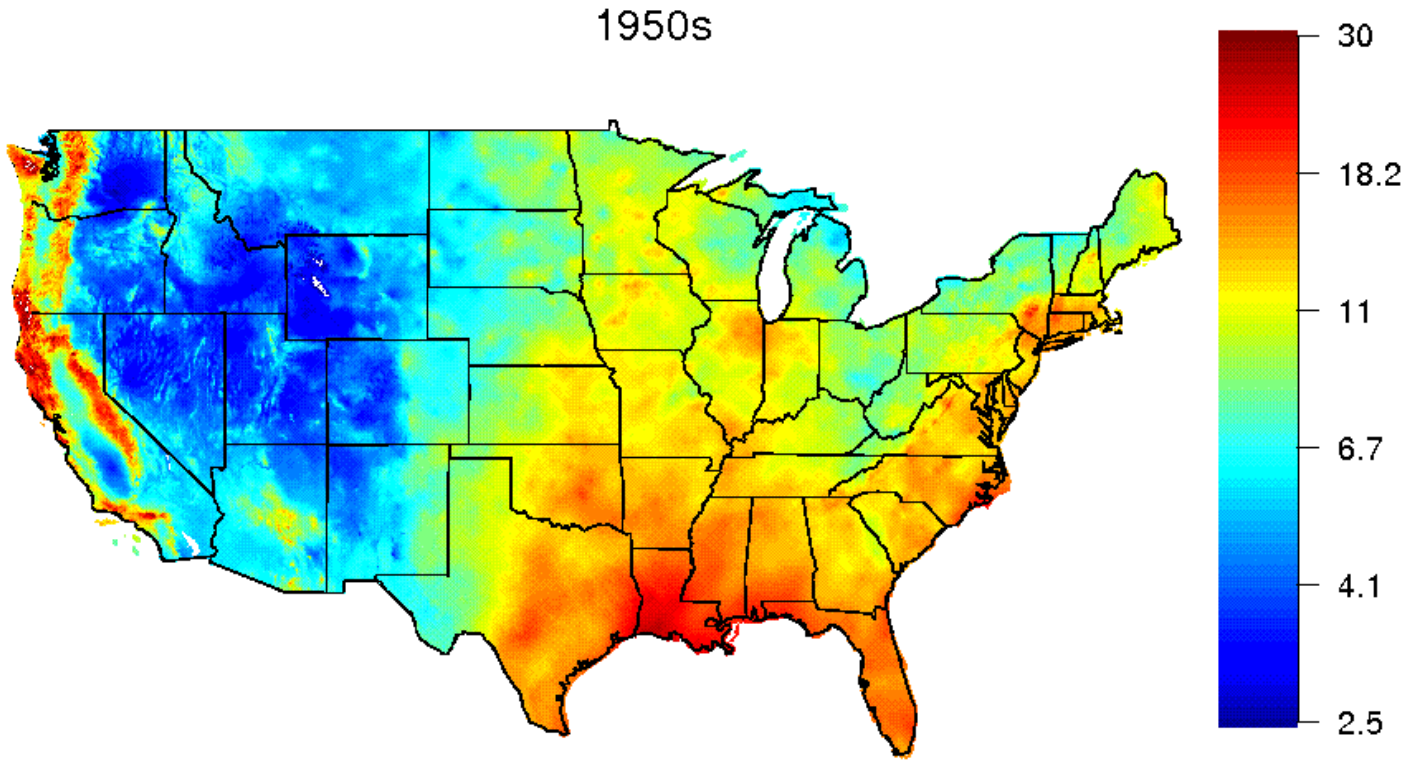
Δ_3: Mean change in log 25-year return value (1970 to 1999) for CCSM, by kriging

SE_3: Corresponding standard error (or RMSPE)

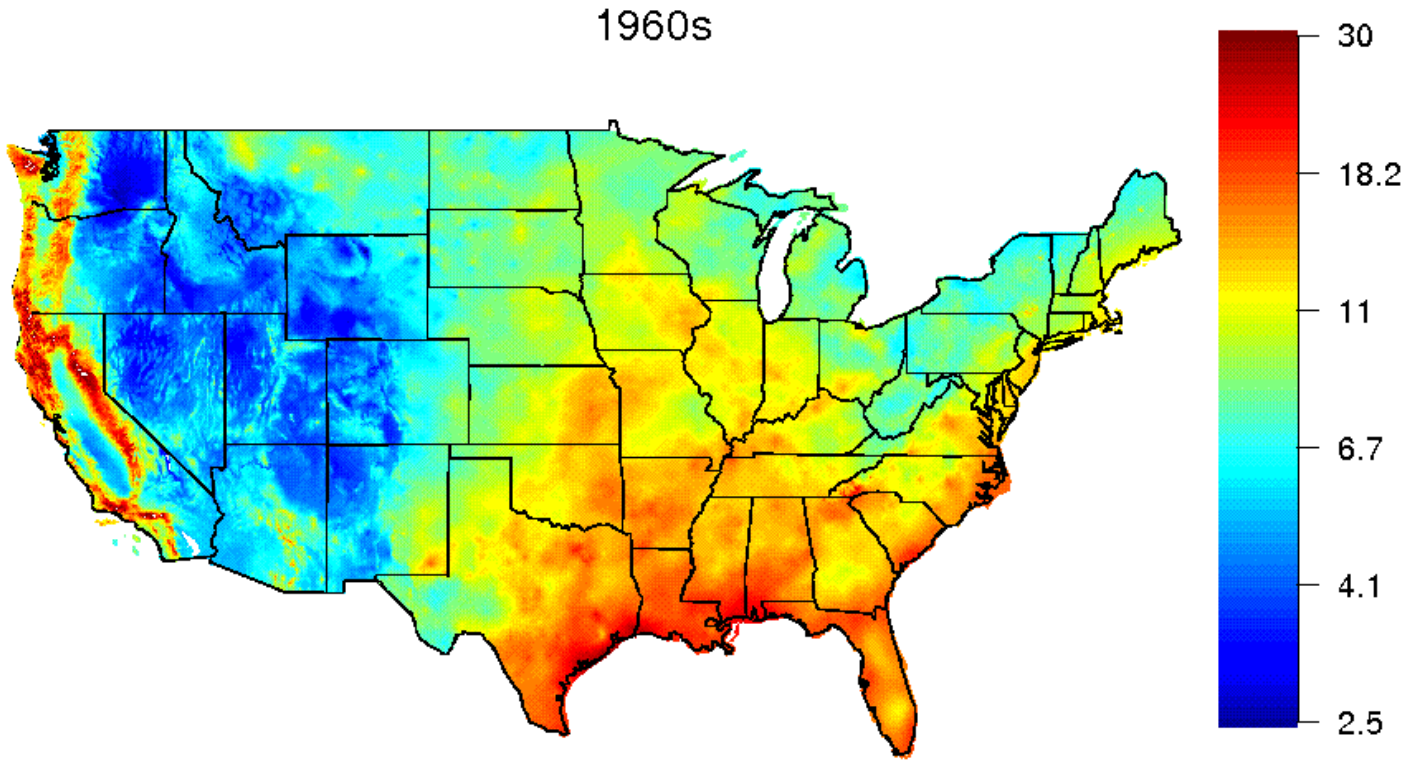
Δ_4, SE_4: Results using GWA

Stars indicate significance at 5%*, 1%**, 0.1%***.

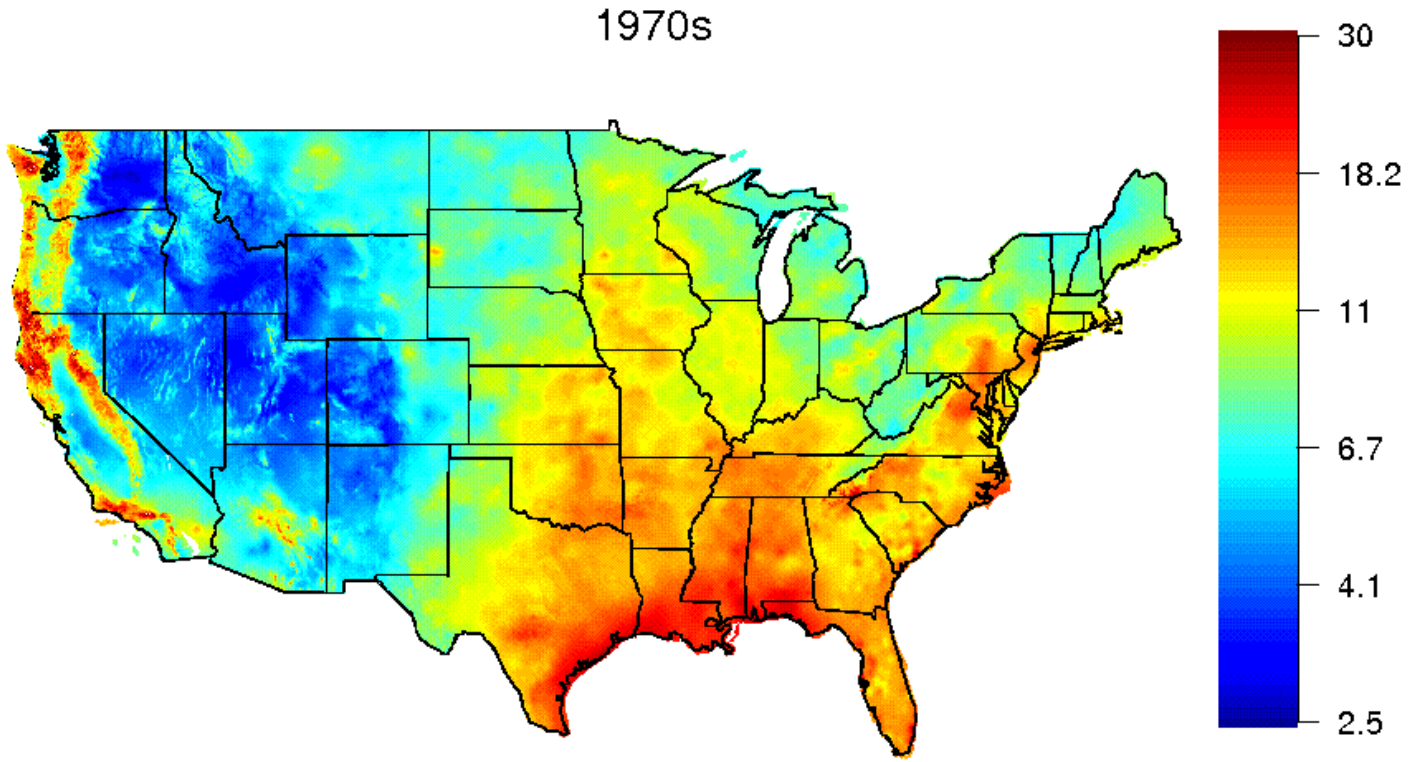
RETURN VALUE MAPS FOR INDIVIDUAL DECADES



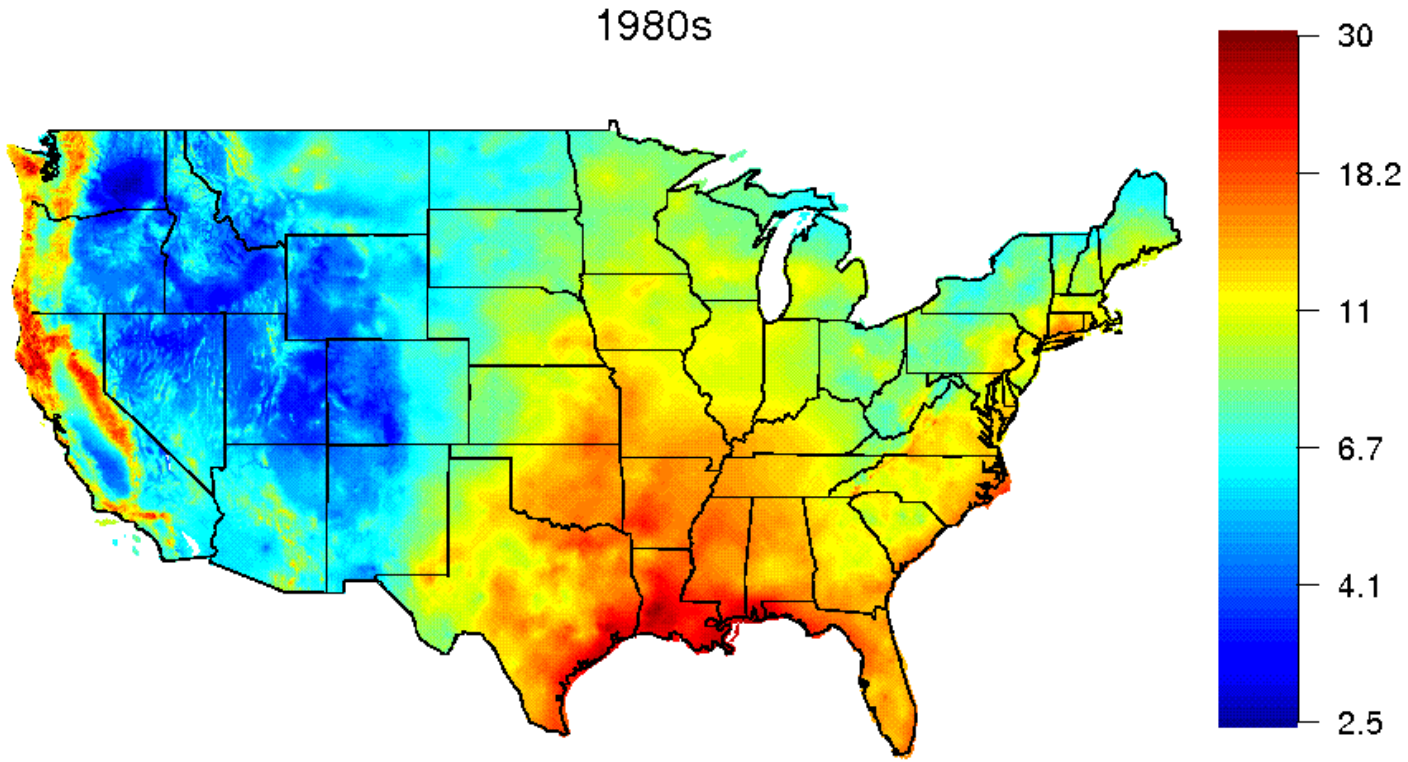
Map of 25-year return values (cm.) for the years 1950–1959



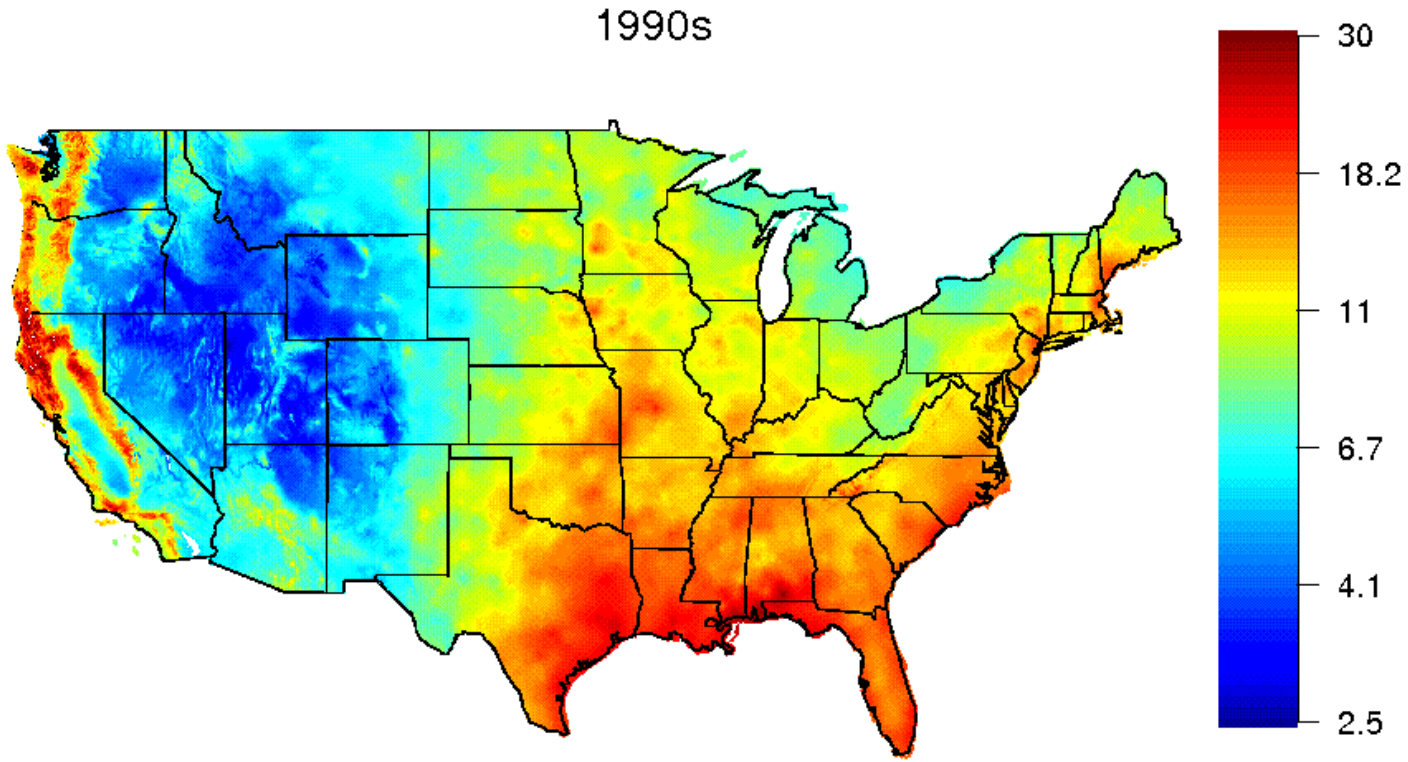
Map of 25-year return values (cm.) for the years 1960–1969



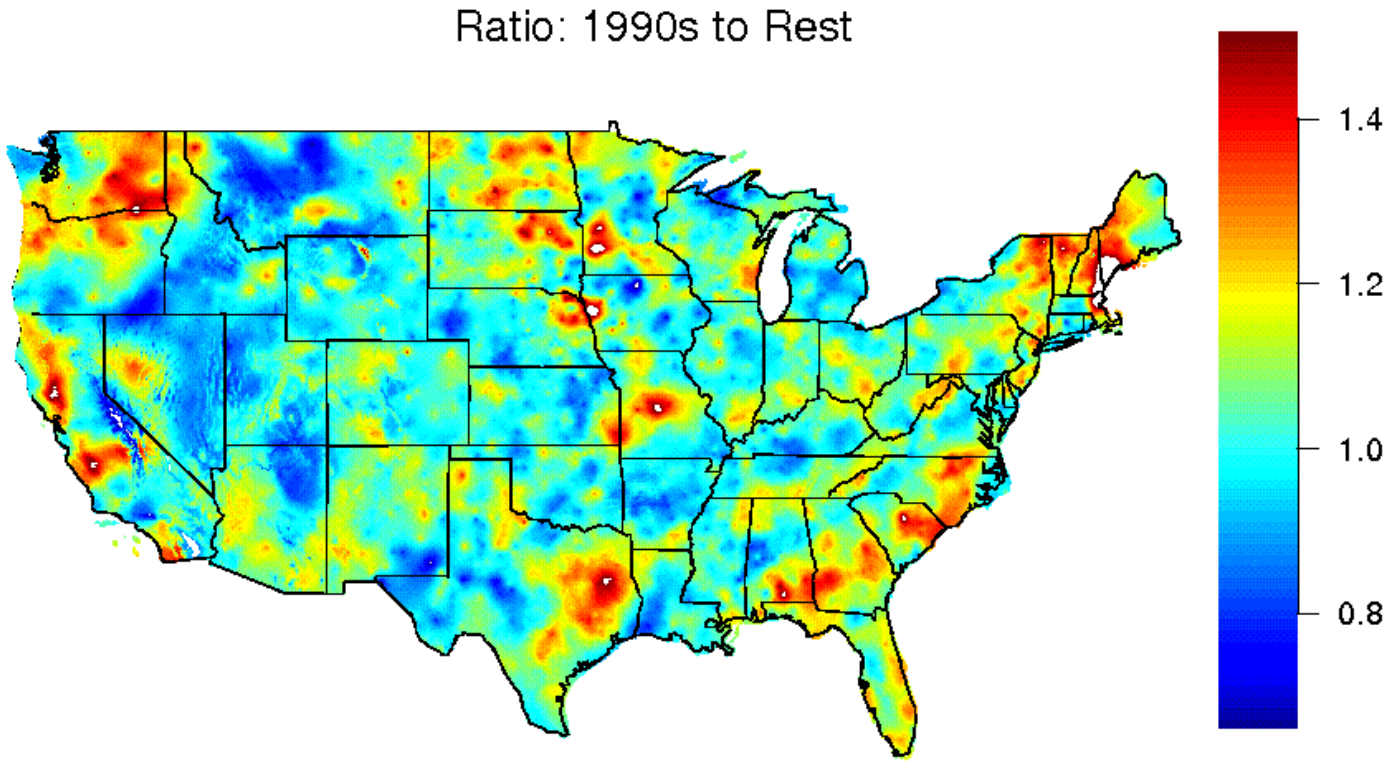
Map of 25-year return values (cm.) for the years 1970–1979



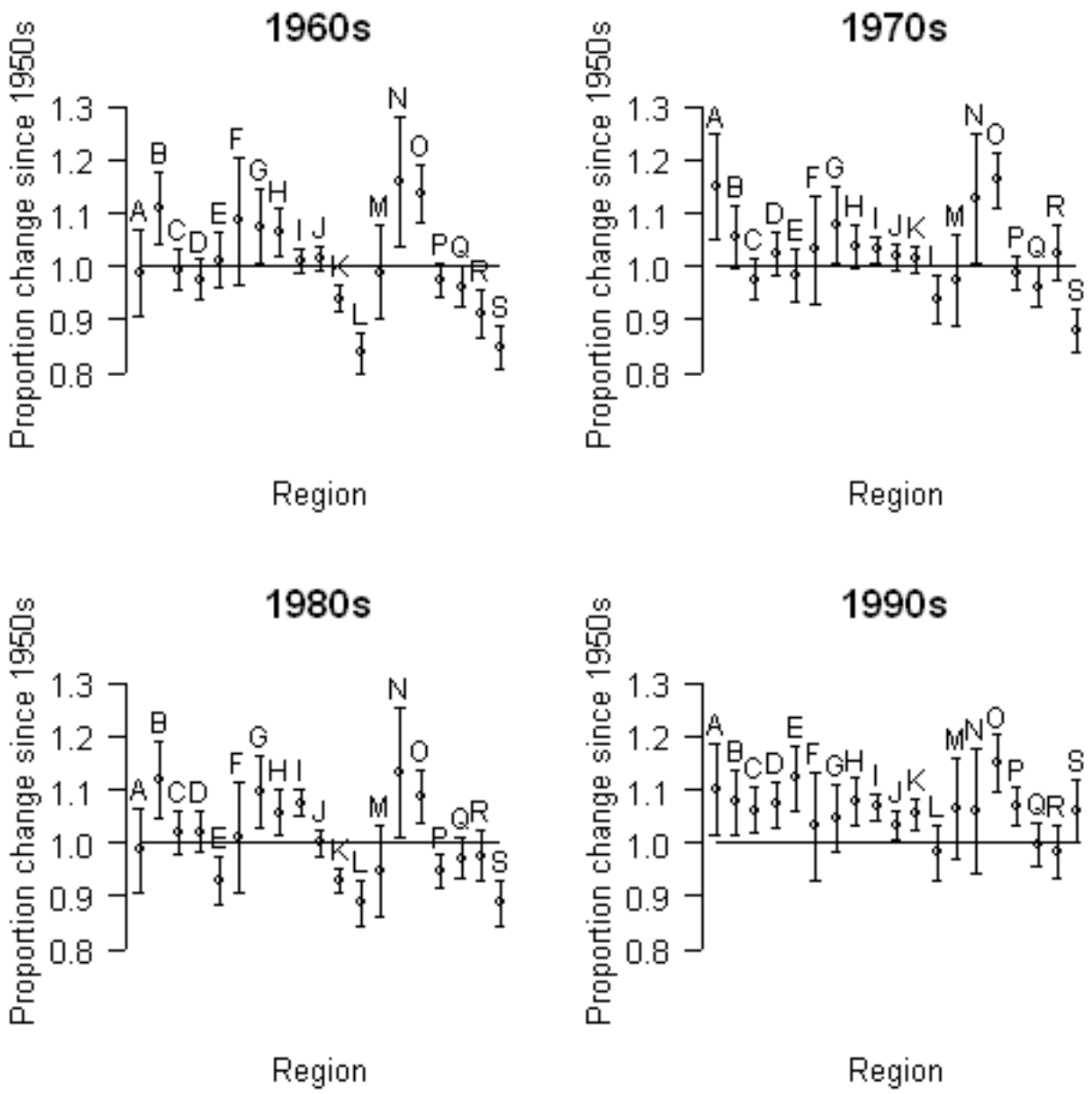
Map of 25-year return values (cm.) for the years 1980–1989



Map of 25-year return values (cm.) for the years 1990–1999



Estimated ratios of 25-year return values for 1990s compared with average at each location over 1950–1989



Regional changes in log RV25 for each decade compared with 1950s

CONCLUSIONS

1. Focus on N-year return values — strong historical tradition for this measure of extremes (we took $N = 25$ here)
2. Seasonal variation of extreme value parameters is a critical feature of this analysis
3. Overall significant increase over 1970–1999 except for parts of western states — average increase across continental US is 7%
4. Kriging better than GWA
5. *But...* based on CCSM data there is a completely different spatial pattern and no overall increase
6. Projections to 2070–2099 show further strong increases but note caveat based on point 5
7. Decadal variations since 1950s show strongest increases during 1990s.