EXTREME PRECIPITATION TRENDS OVER THE CONTINENTAL UNITED STATES

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

15th ‘Aha Huliko’a Hawaiian Winter Workshop
Honolulu
January 24, 2007

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

BACKGROUND

During the past decade, many papers have appeared in the climatology literature documenting increases in the frequency of extreme meteorological events. Much of this literature has been concerned with precipitation extremes.

Also, it is widely believed that such increases are directly linked to global warming (specifically, that increases in greenhouse gases are responsible for increases in frequency of extreme events, as well as increases in overall average temperatures).

Major feature of current reports on climate science, e.g. IPCC, CCSP

In this talk, I explore systematically the development of extreme value theory and spatial statistics 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

Criticisms of current methodology

- Exceedance-counting approach too limited to measure many quantities of practical interest, e.g. return levels
- Ignores seasonality
- "Geometric weighting" approach for calculating regional average trends is theoretically inferior to choosing optimal weights based on the covariance function (a.k.a. kriging, optimal interpolation,...), though it's not clear how much this matters
- They don't attempt to construct spatial maps of the trend, though they could, and should if they want to learn about small-scale variation
- Ignores other covariates that could be highly relevant, e.g. El Niño, other circulation indices such as NAO, AO, AMO, PDO, etc. (but I'm going to ignore those as well)

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

Some thoughts on Bayesianism, frequentism and all that

- I am not a frequentist
  - We have a dataset of $\sim$10,000 observations at each of 5,000 stations. In what sense is this one replication of a hypothetically infinite sequence of experiments?

- But I'm not sure I meet Jay Kadane's definition of a Bayesian either
  - My personal probabilities about climate change are certainly different from Dick Lindzen's. They are also very different from Phil Jones's.
  - Yet I believe it's possible to make probabilistic statements about scientific issues without depending excessively on personal beliefs

- Most of my statements about standard errors, significance levels, etc. can be reinterpreted as statements about the posterior distribution (under a flat prior), and that is probably the best interpretation

- There is a wide body of theoretical and empirical literature that suggests that in large samples, the role of the prior is unimportant

- To most people, $n = 50,000,000$ looks like a large sample

- But there is another point of view by which the sample is of size 1

- What is the correct sample size here?

Statements of uncertainty are starting to appear in the press.

“In fresh drafts of a summary of its next report, the group, the Intergovernmental Panel on Climate Change, has said that it is more than 90 percent likely that global warming since 1950 has been driven mainly by the buildup of carbon dioxide and other heat-trapping greenhouse gases, and that more warming and rising sea levels are on the way.

In its last report, published in 2001, the panel concluded that there was a 66 to 90 percent chance that human activities were driving the most recent warming.”

(Andrew Revkin in the *New York Times*, January 20 2007)

What does this mean?

BASICS OF EXTREME VALUE THEORY

We start with the *extreme value limit laws* (Fisher and Tippett 1928; Gnedenko 1943)

Let $X_1, X_2, ...,$ be independent identically distributed (IID) random variables with distribution function F.

Let $M_n = \max(X_1, ..., X_n)$. Then

$$\Pr\{M_n \le x\} = F^n(x) \to 0$$

for any x such that $F(x) < 1$.

To obtain interesting results *renormalize*: Find $a_n > 0$, b_n,

$$\begin{aligned} \Pr\left\{\frac{M_n - b_n}{a_n} \le x\right\} &= F^n(a_n x + b_n) \\ &\to G(x) \end{aligned}$$

where G is a nondegenerate limiting distribution function.

The Three Extreme Value Types

Type I (Gumbel)

$$\Lambda(x) = \exp(-e^{-x}),\ -\infty < x < \infty.$$

Type II (Fréchet)

$$\Phi_\alpha(x) = \begin{cases} 0, & \text{if } x \le 0, \\ \exp(-x^{-\alpha}), & \text{if } x \ge 0\ (\alpha > 0). \end{cases}$$

Type III (Weibull)

$$\Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^\alpha\}, & \text{if } x \le 0\ (\alpha > 0), \\ 1, & \text{if } x \ge 0. \end{cases}$$

Generalized EV Distribution:

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

($x_+ = \max(x, 0)$) where $-\infty < \mu < \infty$, $0 < \psi < \infty$, $-\infty < \xi < \infty$.
The limit $\xi \to 0$ corresponds to the Gumbel case.

Exceedances Over Thresholds

Exceedances over a high threshold u.

$$\begin{aligned} F_u(y) &= \Pr\{X \le u + y \mid X > u\} \\ &= \frac{F(u+y) - F(u)}{1 - F(u)}. \quad (y > 0) \end{aligned}$$

Look for scaling constants $\{c_u\}$ so that as $u \uparrow \omega_F = \sup\{x : F(x) < 1\}$,

$$F_u(zc_u) \to H(z)$$

where H is nondegenerate. In that case, H must be of form

$$H(z) = \begin{cases} 1 - \left(1 + \frac{\xi z}{\sigma}\right)_+^{-1/\xi}, & \text{if } \xi \neq 0, \\ 1 - e^{-z/\sigma}, & \text{if } \xi = 0, \end{cases}$$

where $\sigma > 0$ and $-\infty < \xi < \infty$.

This is the *Generalized Pareto Distribution* (Pickands 1975).

Statistical Approaches

Peaks Over Thresholds

Basic idea: fix a high threshold u say, and fit the Generalized Pareto distribution (GPD) to exceedances over the threshold.

May need separate analysis to model the probability of crossing the threshold as a function of covariates, e.g. logistic regression.

Extensions of the basic methodology:

- Selecting the threshold
- Incorporating covariates
- Dependence in the time series

Statistical Approaches, Continued

Point process approach

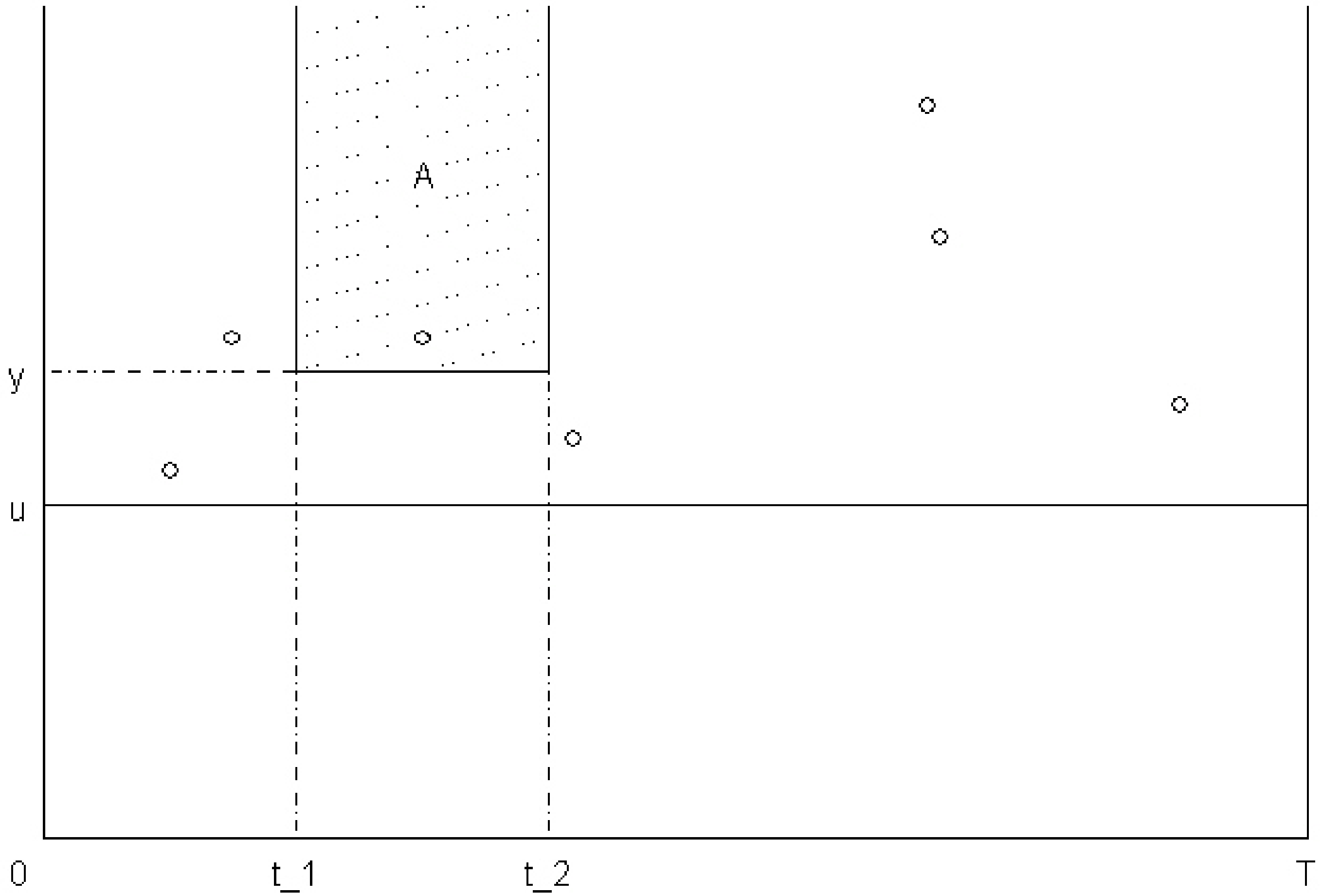
The expected number of exceedances in a box of the form of A is assumed to be

$$\Lambda(A) = (t_2 - t_1)\Psi(y; \mu, \psi, \xi)$$

where

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

In practice, allow parameters to depend on covariates.


A
y
u
0
t_1
t_2
T

The main technique used in this talk is a generalization of the point process approach in which the extreme value parameters $\mu,\ \psi,\ \xi$ are replaced by time-dependent parameters $\mu_t,\ \psi_t,\ \xi_t$. In this way, I am able to model both seasonal variation and trend.

This technique is initially applied one station at a time. Later, the results will be combined across stations to produce a spatial map.

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.

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)

The N-year return value is defined as the value that is exceeded in any given year with probability $\frac{1}{N}$. It is calculated by finding y_N numerically to solve

$$\sum_{t_1}^{t_2} \left(1 + \xi_t \frac{y_N - \mu_t}{\psi_t}\right)^{-1/\xi_t} \quad = \quad -\log\left(1 - \frac{1}{N}\right)$$

where $t = t_1, ..., t_2$ are the days within the year of interest.

Defined like this, I can treat Y_N itself as a time-dependent variable that possibly increases over the time period of the study, and compare the N-year return level for 1999 with that for 1970 (and later, that for 2070–2099 with 1970–1999 form a climate model).

For convenience, I fix $N = 25$ in subsequent discussion.

SPATIAL SMOOTHING

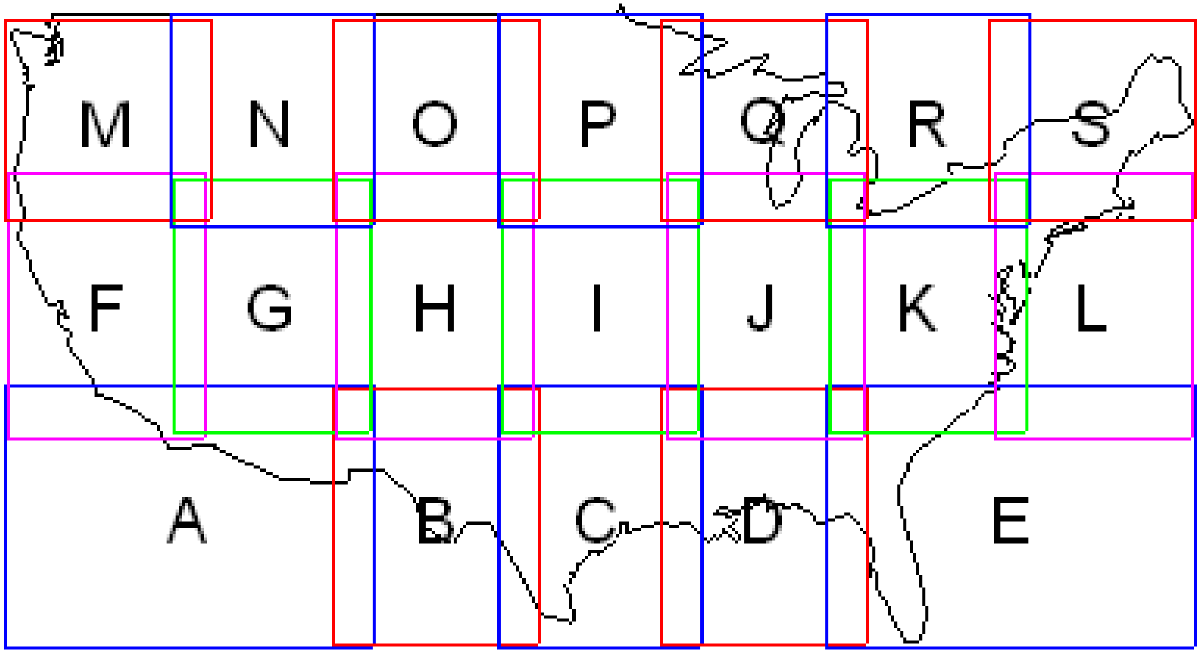
Let Z_s be field of interest, indexed by s (typically the logarithm of the 25-year RV at site s, or a log of ratio of RVs. Taking logs improves fit of spatial model, to follow.)

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

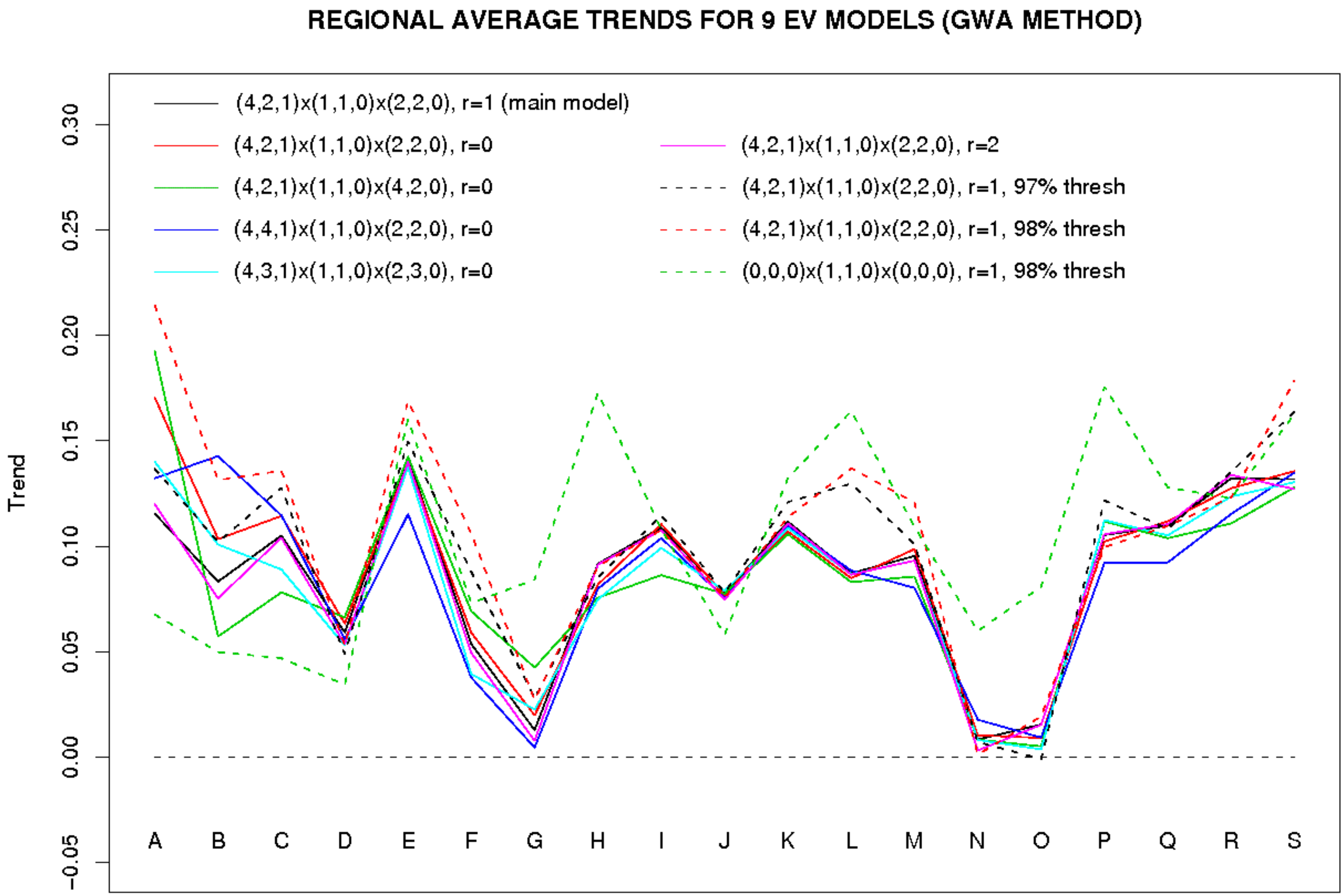
$$\begin{aligned} \hat{Z} \mid Z &\sim N[Z, W] \\ Z &\sim N[X\beta, V(\phi)] \\ \hat{Z} &\sim N[X\beta, V(\phi) + W]. \end{aligned}$$

for known W; X are covariates, β are unknown regression parameters and ϕ are parameters of spatial covariance matrix V. The covariates here include elevation and mean annual precipitation.

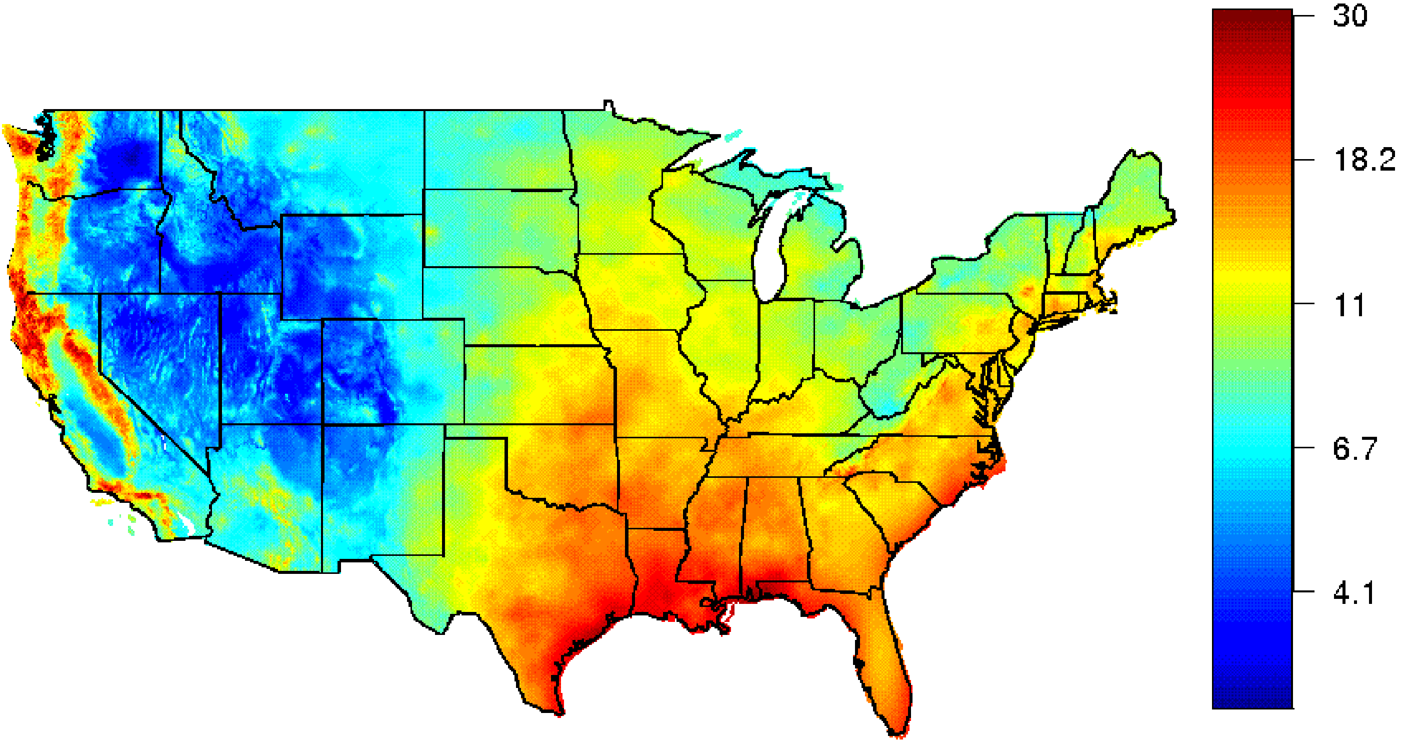
- ϕ by REML (Matérn covariances with nugget)
- β given ϕ by generalized least squares
- Predict Z at observed and unobserved sites by kriging



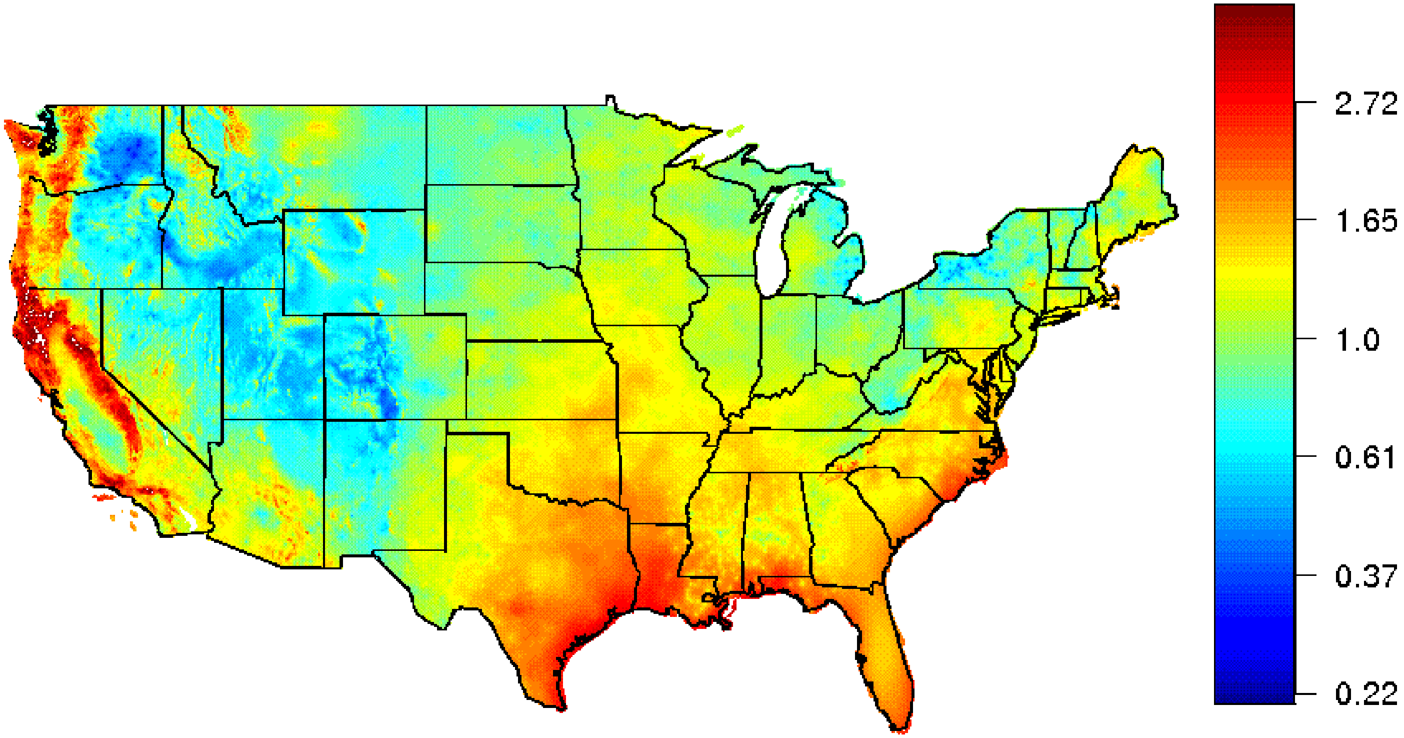
Continental USA divided into 19 regions



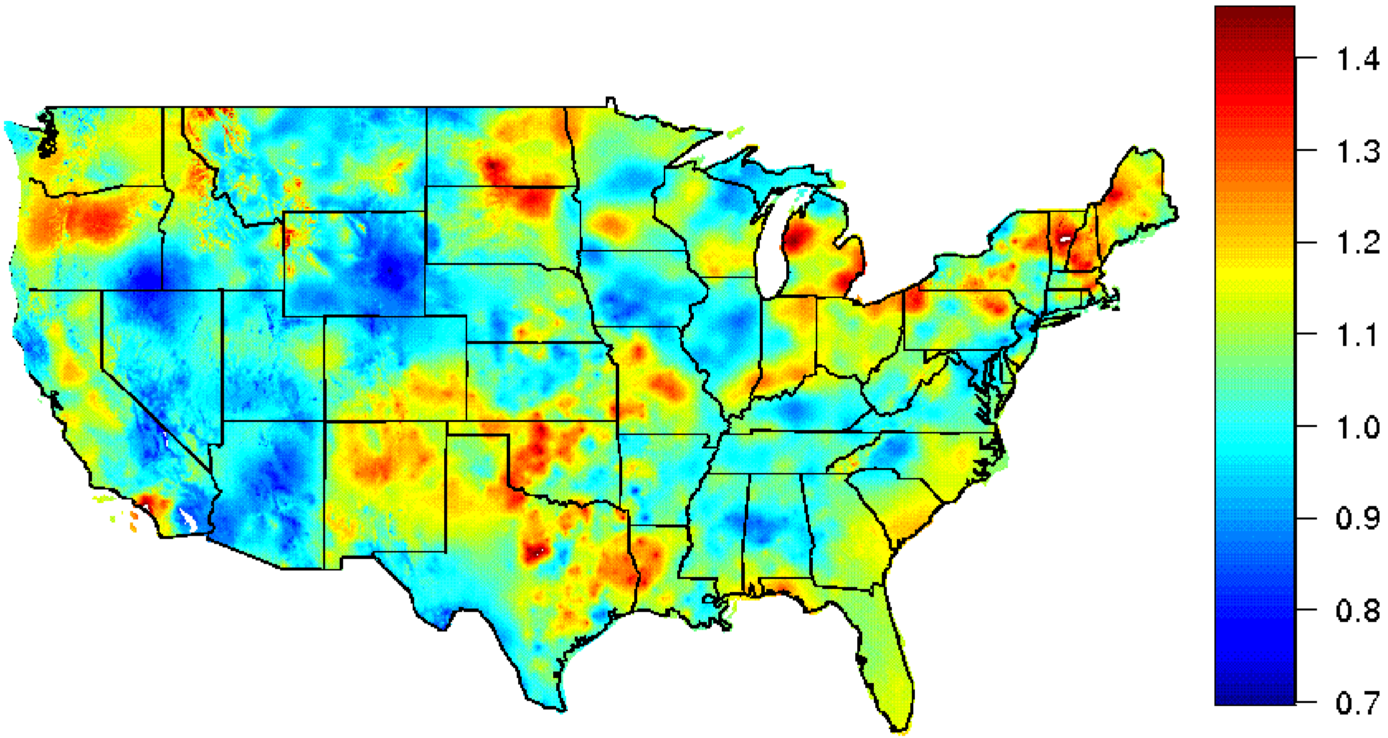
Trends across 19 regions (measured as change in log RV25) for 8 different seasonal models and one non-seasonal model with simple linear trends. Regional averaged trends by geometric weighted average approach.



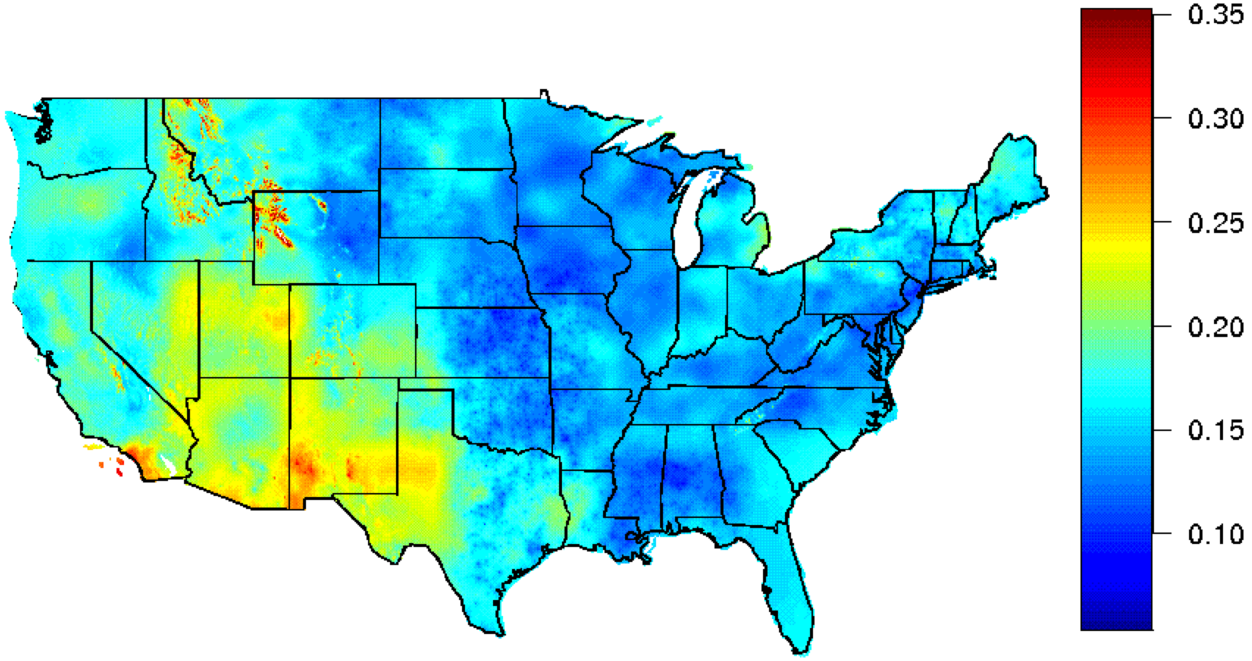
Map of 25-year return values (cm.) for the years 1970–1999



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

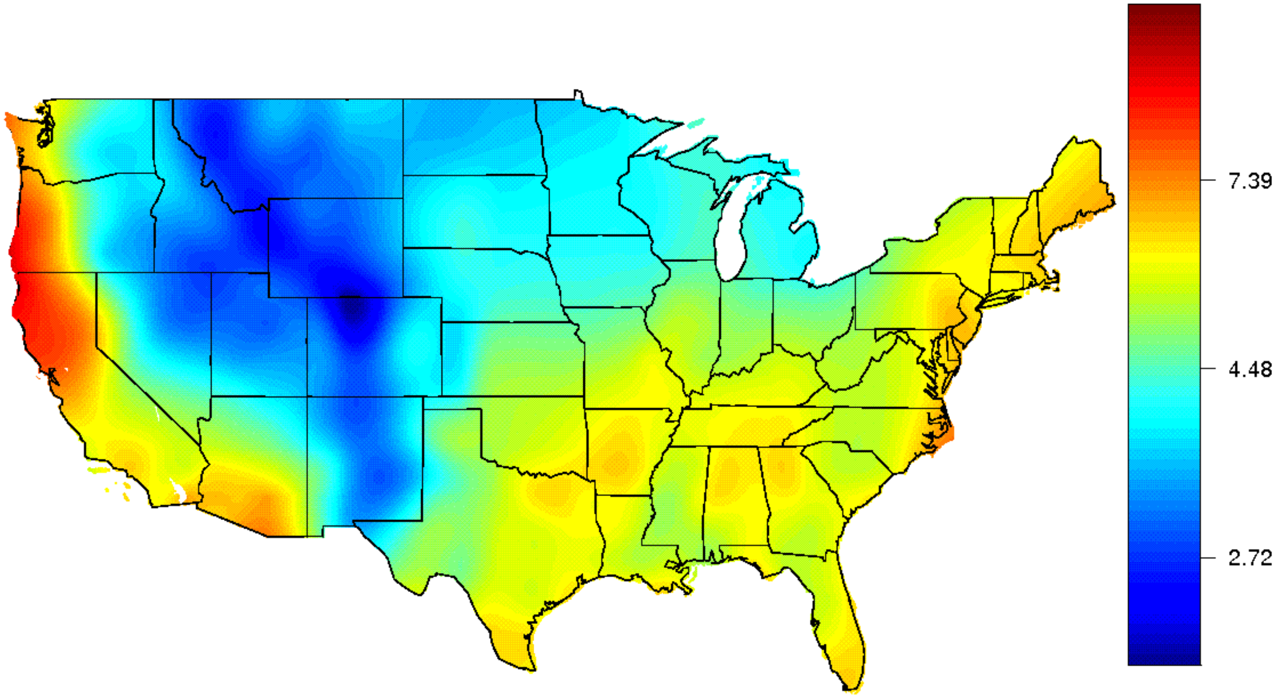
	Δ_1	S_1	Δ_2	S_2		Δ_1	S_1	Δ_2	S_2
A	-0.01	.03	0.05^{**}	.05	K	0.08^{***}	.01	0.09^{**}	.03
B	0.07^{**}	.03	0.08^{***}	.04	L	0.07^{***}	.02	0.07^{*}	.04
C	0.11^{***}	.01	0.10	.03	M	0.07^{***}	.02	0.10^{**}	.03
D	0.05^{***}	.01	0.06	.05	N	0.02	.03	0.01	.03
E	0.13^{***}	.02	0.14^{*}	.05	O	0.01	.02	0.02	.03
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

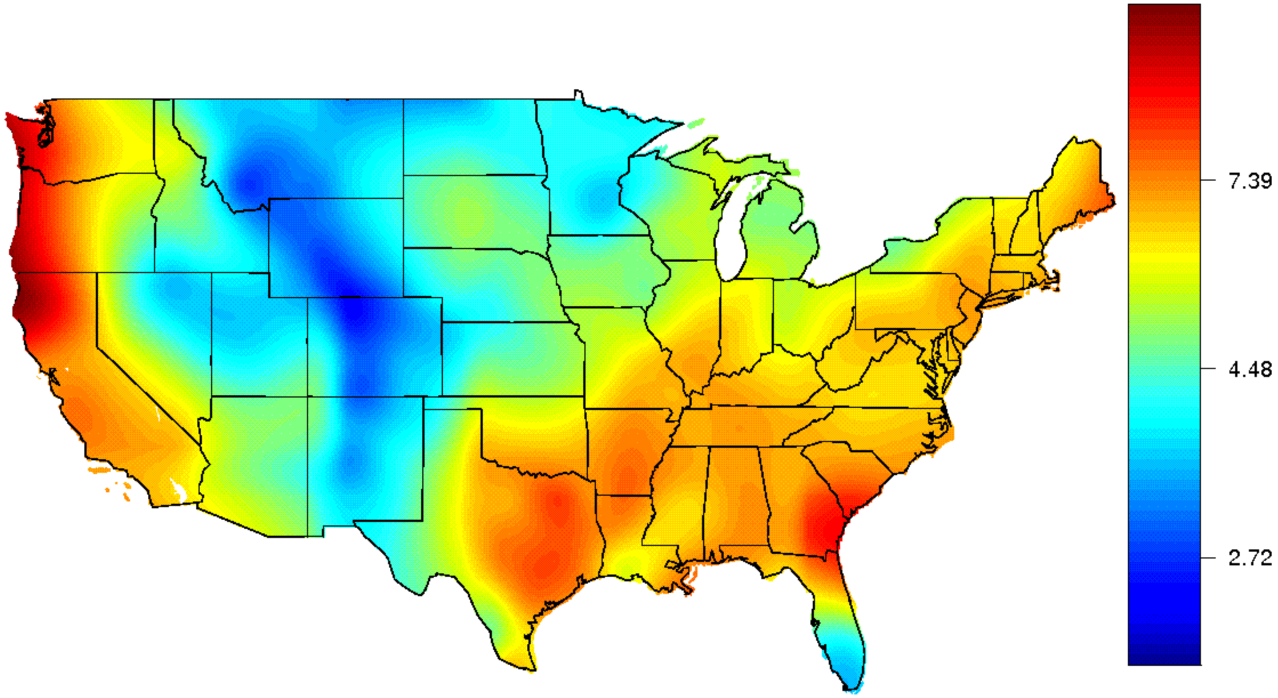
S_1: Corresponding standard error (or RMSPE)

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

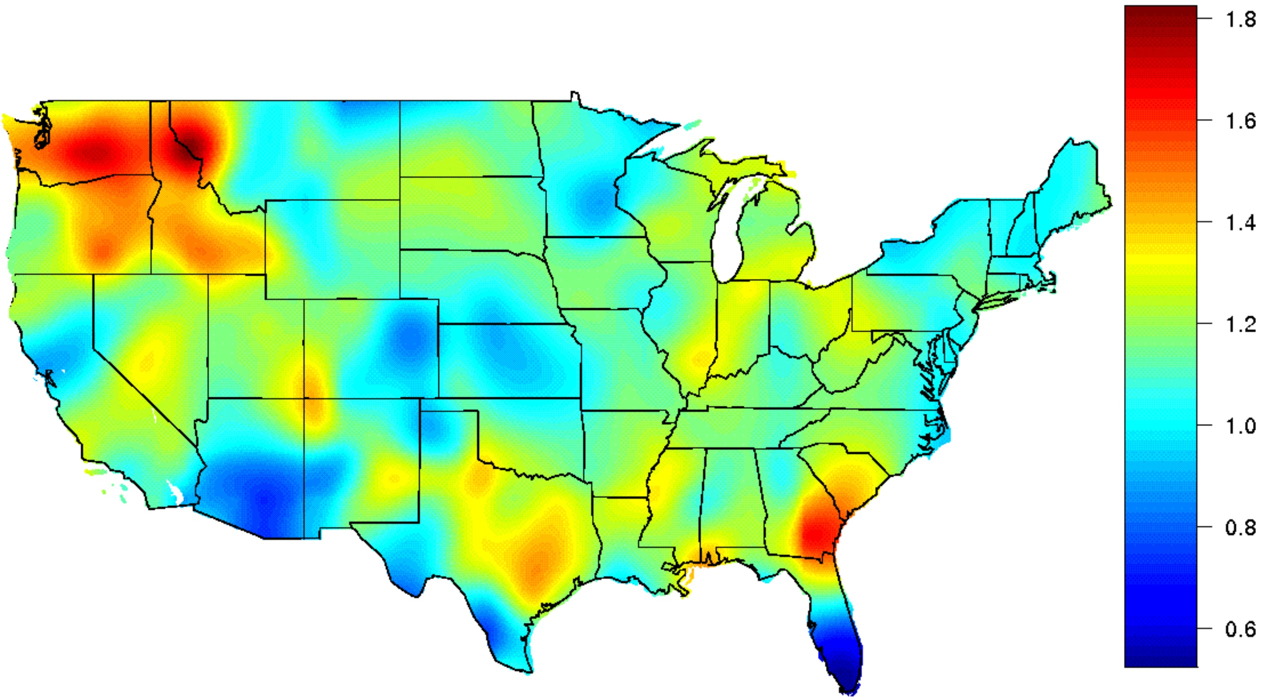
Stars indicate significance at 5%*, 1%**, 0.1%***.



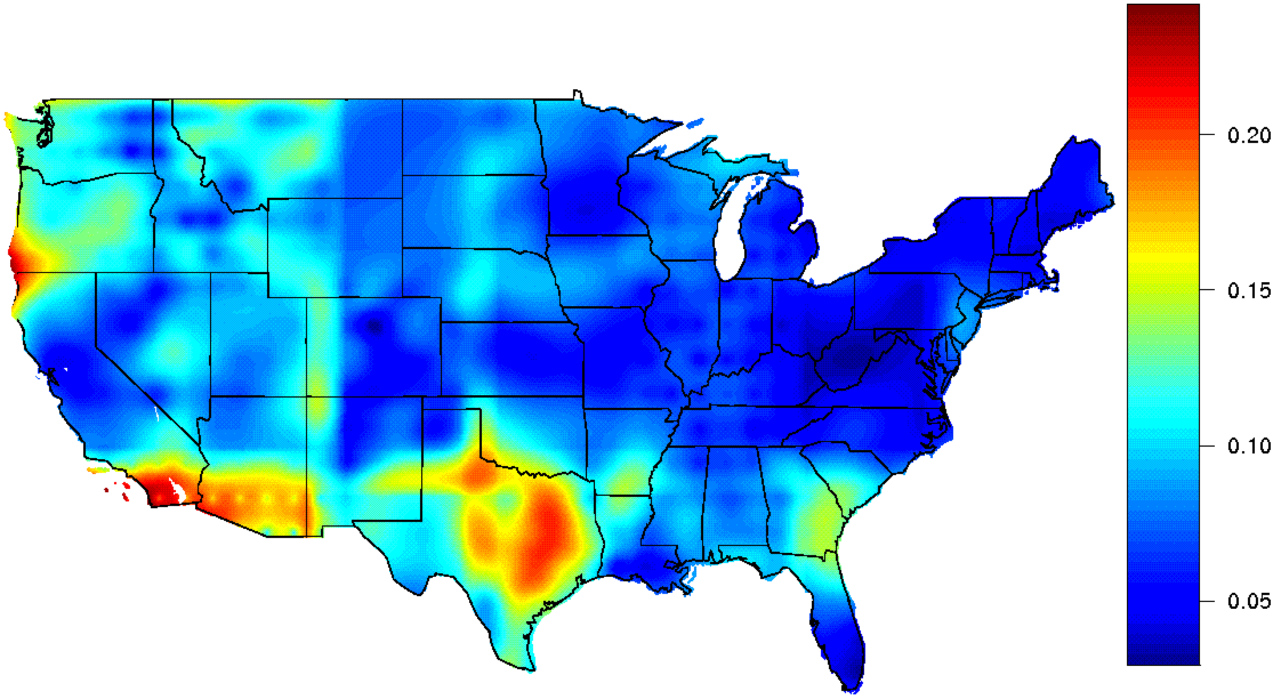
Return value map for CCSM data (cm.): 1970–1999



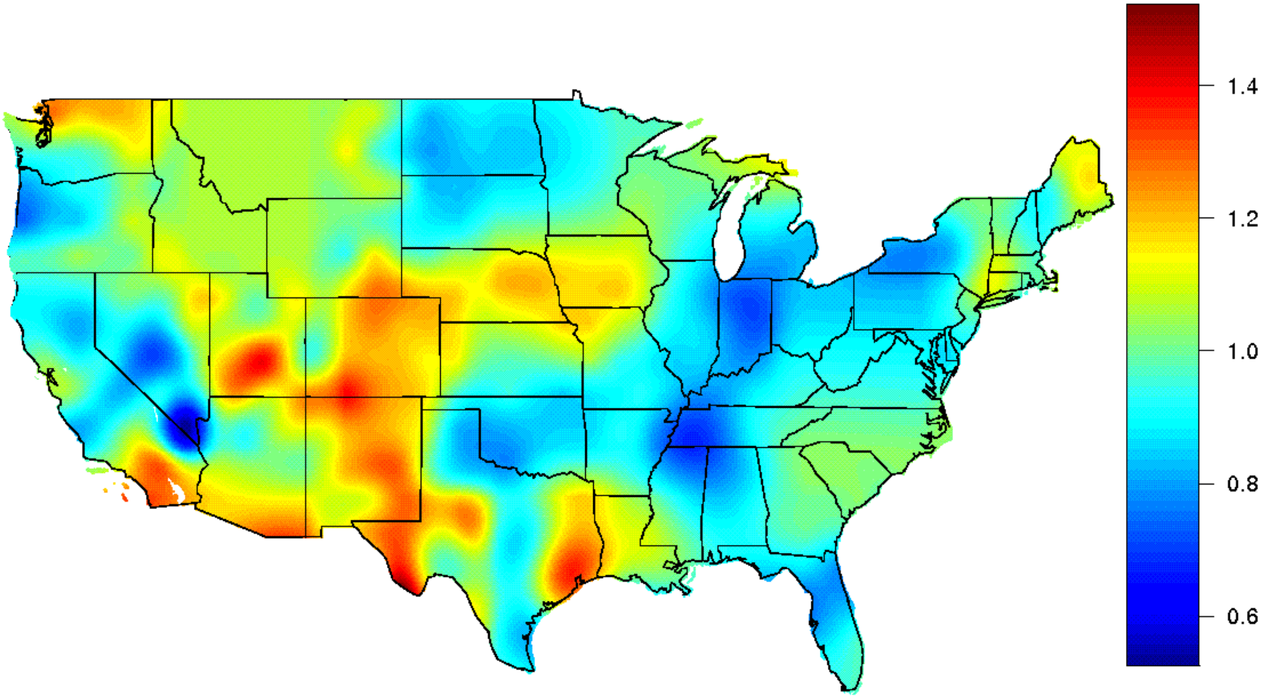
Return value map for CCSM data (cm.): 2070–2099



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

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

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%***.

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.