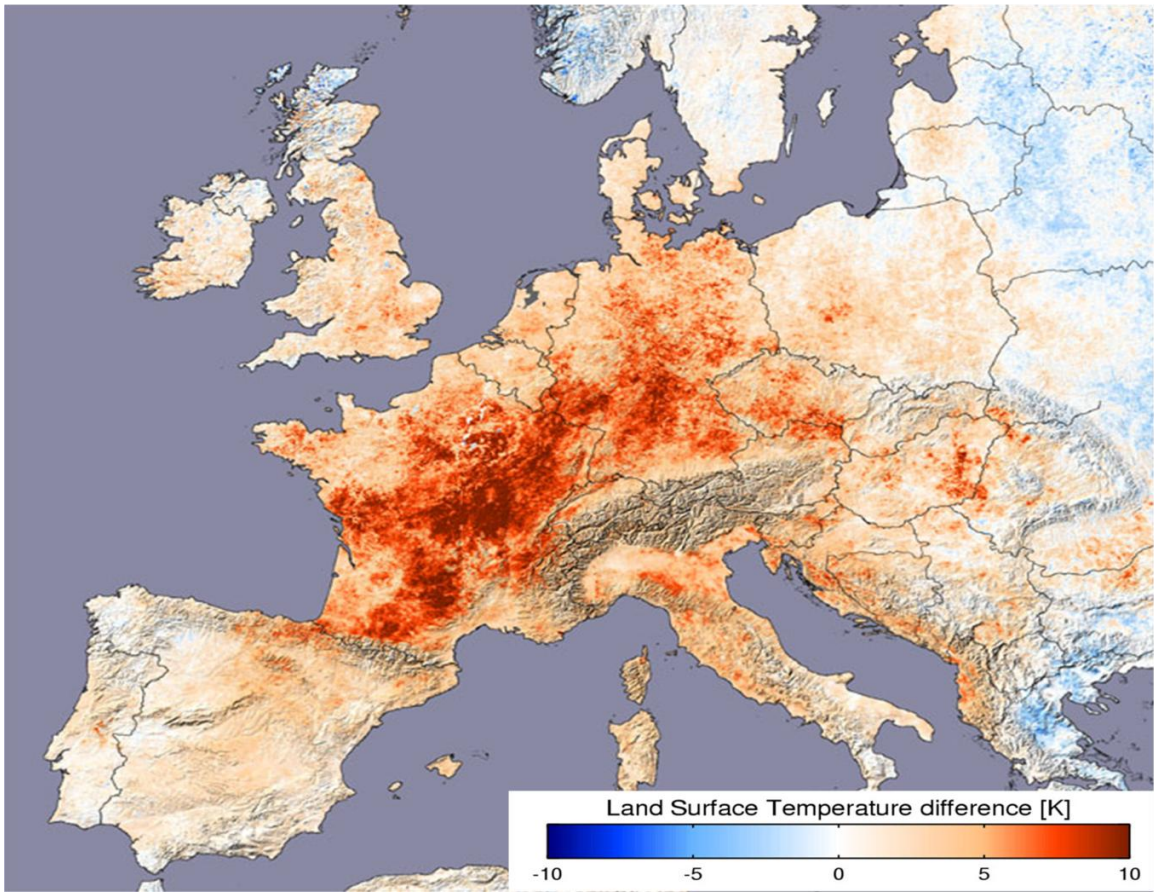
ATTRIBUTION OF EXTREME EVENTS USING OBSERVATIONAL DATA AND CLIMATE MODELS

Richard L. Smith
Department of Statistics and Operations Research
University of North Carolina, Chapel Hill
rls@email.unc.edu
and

Statistical and Applied Mathematical Sciences Institute

Data Hierarchices for Climate Modeling
Workshop at IPAM
May 28 2011



European temperatures in early August 2003, relative to 2001-2004 average

From NASA's MODIS - Moderate Resolution Imaging Spectrometer, courtesy of Reto Stöckli, ETHZ

(From a presentation by Myles Allen)

Human contribution to the European heatwave of 2003

Peter A. Stott[1], D. A. Stone[2,3] & M. R. Allen[2]

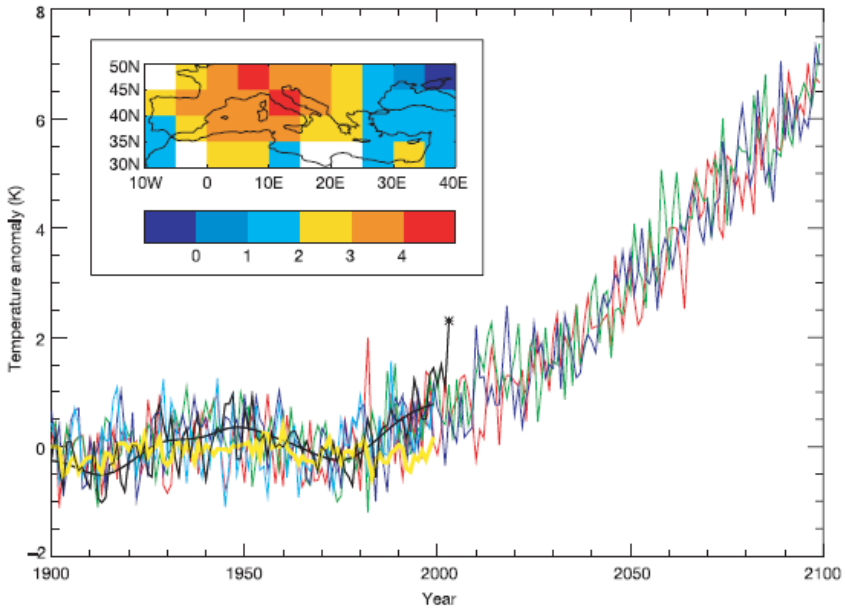


Figure 1 June–August temperature anomalies (relative to 1961–90 mean, in K) over the region shown in inset. Shown are observed temperatures (black line, with low-pass-filtered temperatures as heavy black line), modelled temperatures from four HadCM3 simulations including both anthropogenic and natural forcings to 2000 (red, green, blue and turquoise lines), and estimated HadCM3 response to purely natural natural forcings (yellow line). The observed 2003 temperature is shown as a star. Also shown (red, green and blue lines) are three simulations (initialized in 1989) including changes in greenhouse gas and sulphur emissions according to the SRES A2 scenario to 2100[22]. The inset shows observed summer 2003 temperature anomalies, in K.

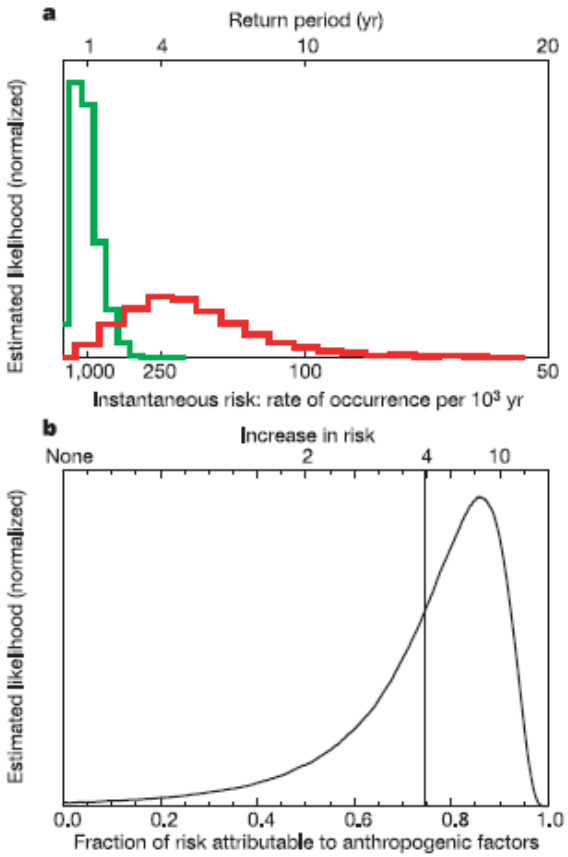


Figure 4 Change in risk of mean European summer temperatures exceeding the 1.6 K threshold. **a**, Histograms of instantaneous return periods under late-twentieth-century conditions in the absence of anthropogenic climate change (green line) and with anthropogenic climate change (red line). **b**, Fraction attributable risk (FAR). Also shown, as the vertical line, is the 'best estimate' FAR, the mean risk attributable to anthropogenic factors averaged over the distribution.

OUTLINE OF TALK

I. Extreme value theory

- Probability Models
- Estimation
- Profile Likelihoods

II. Trends in Extreme Rainfall Events

III. Analysis of 2003 European Heatwave

I. EXTREME VALUE THEORY

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

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

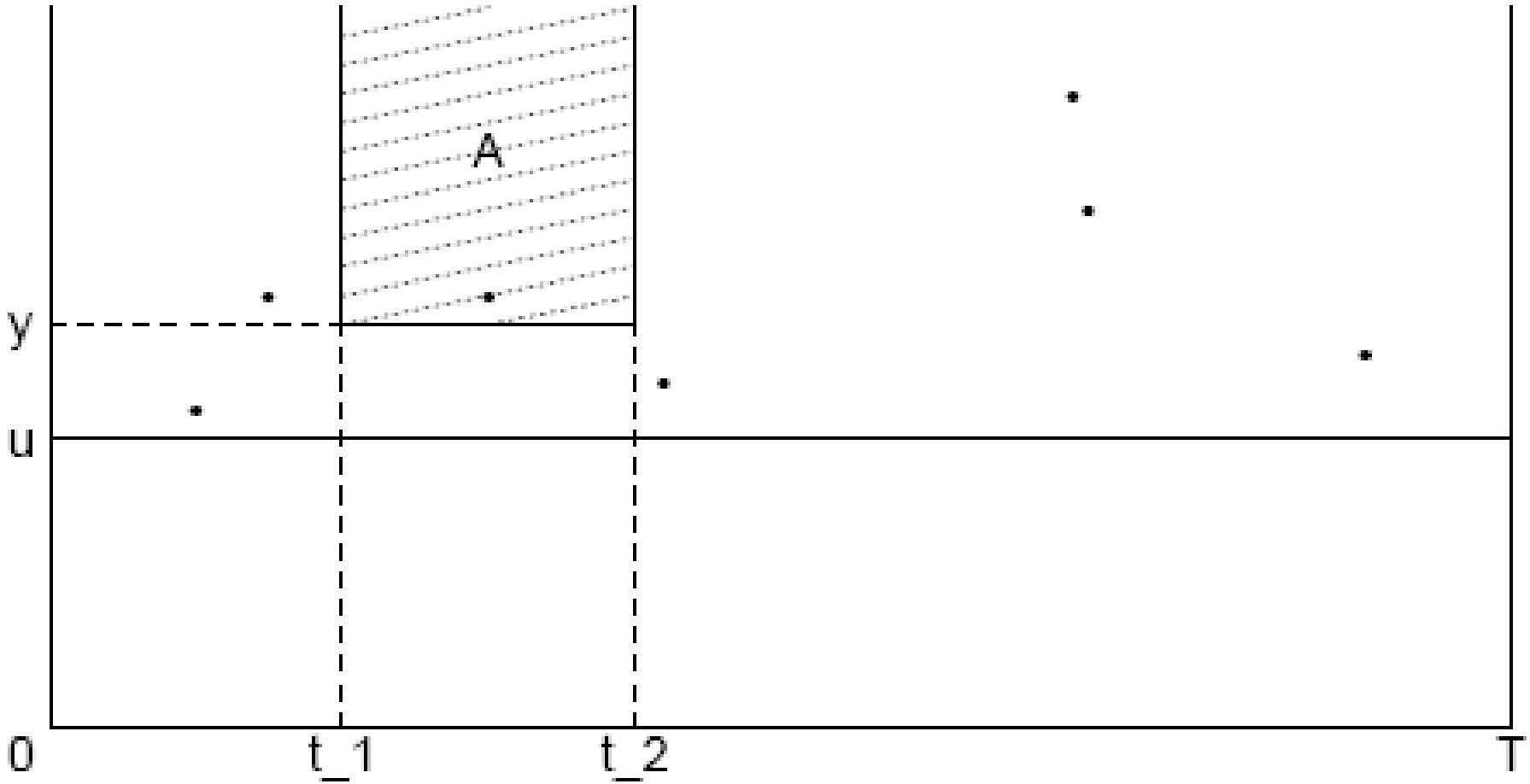


Illustration of point process model.

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

ESTIMATION

For the point process approach, likelihood

$$\prod_i \left\{ \frac{1}{\psi_{T_i}} \left(1 + \xi_{T_i} \frac{Y_i - \mu_{T_i}}{\psi_{T_i}} \right)^{-1/\xi_{T_i} - 1} \right\} \exp \left\{ - \int_0^T \left(1 + \xi_t \frac{u_t - \mu_t}{\psi_t} \right)^{-1/\xi_t} dt \right\}$$

In practice, integral replaced by Riemann sum: neg log likelihood is

$$\sum_i \left\{ \log \psi_{T_i} + \left(\frac{1}{\xi_{T_i}} + 1 \right) \log \left(1 + \xi_{T_i} \frac{Y_i - \mu_{T_i}}{\psi_{T_i}} \right) \right\}$$
$$+ \delta \sum_{t=1}^{T} \left(1 + \xi_t \frac{u_t - \mu_t}{\psi_t} \right)^{-1/\xi_t}$$

where δ is the interval between successive time points, e.g. for daily observations in a situation where annual maxima are of interest, $\delta = 1/365$.

II. TREND IN PRECIPITATION EXTREMES

(joint work with Amy Grady and Gabi Hegerl)

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. However, some papers by statisticians have explored 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 work, we explore systematically the development of extreme value and spatial models

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

Statistical Analysis

The basic method uses the point process approach to extremes with time-dependent parameters $\mu_t,\ \psi_t,\ \xi_t$.

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)

SPATIAL SMOOTHING

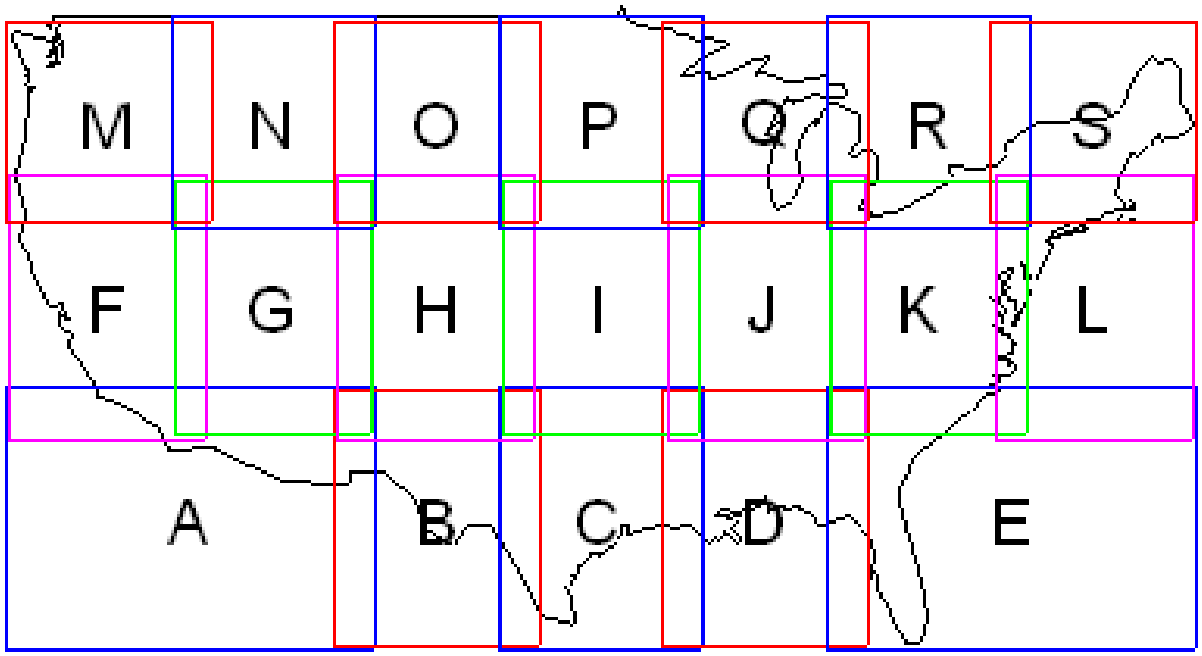
Let Z_s be field of interest, indexed by s (typically the logarithm of the 25-year RV at site s, or a log of ratio of RVs).

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

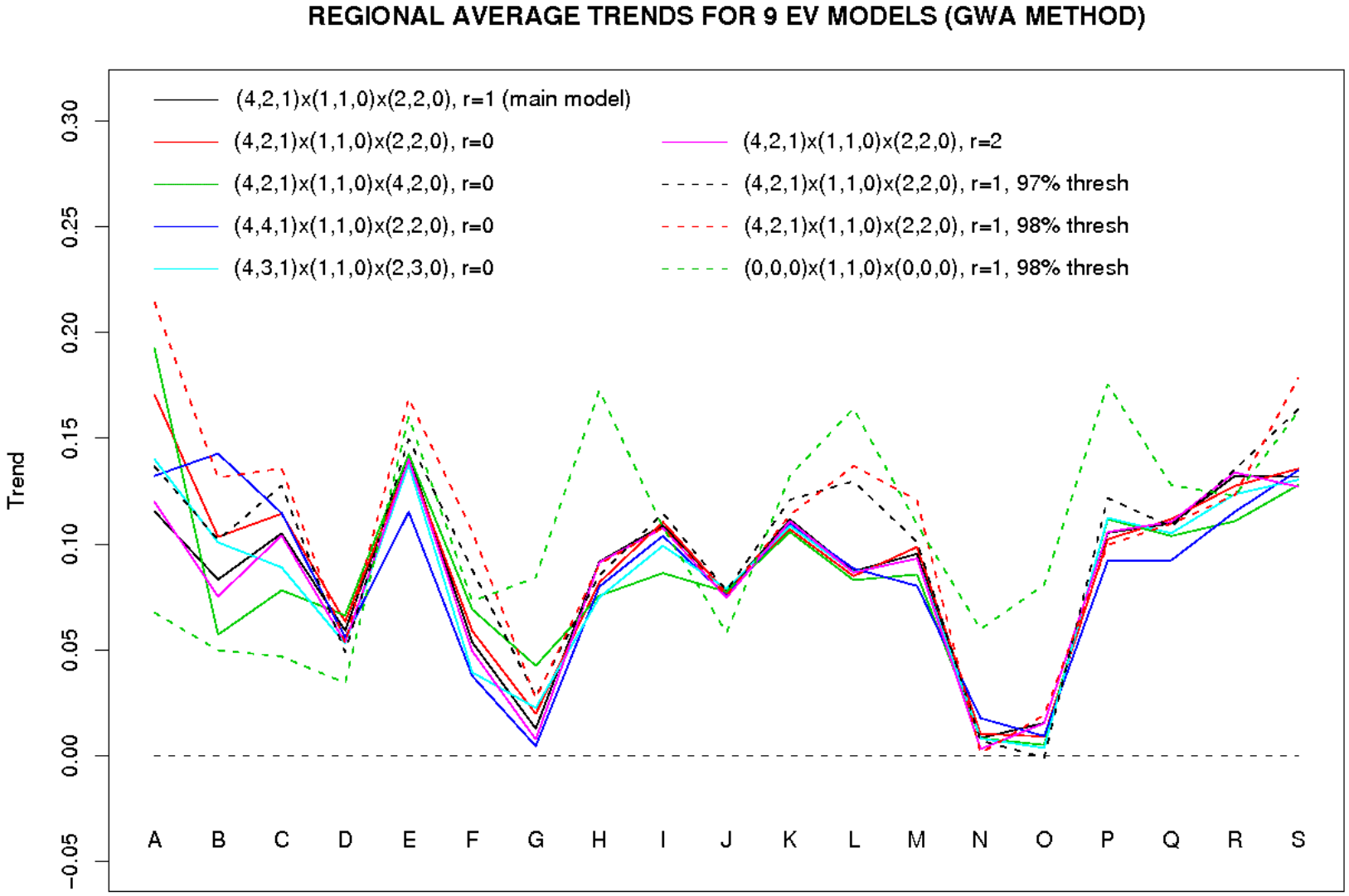
$$\begin{aligned} \hat{Z} \mid Z &\sim N[Z, W] \\ Z &\sim N[X\beta, V(\phi)] \\ \hat{Z} &\sim N[X\beta, V(\phi) + W]. \end{aligned}$$

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

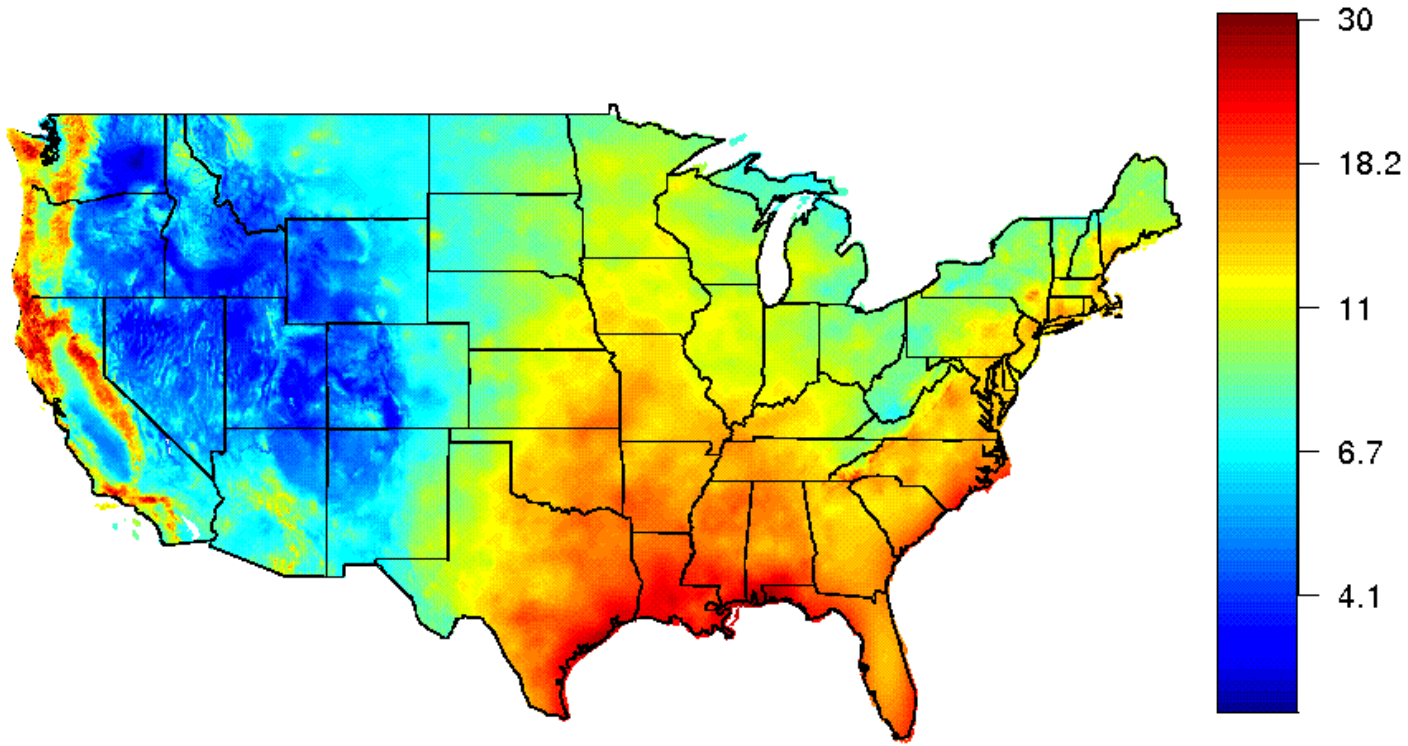
- ϕ by REML
- β given ϕ by GLS
- Predict Z at observed and unobserved sites by kriging
- Apply separately to 19 regions; interpolate across boundaries



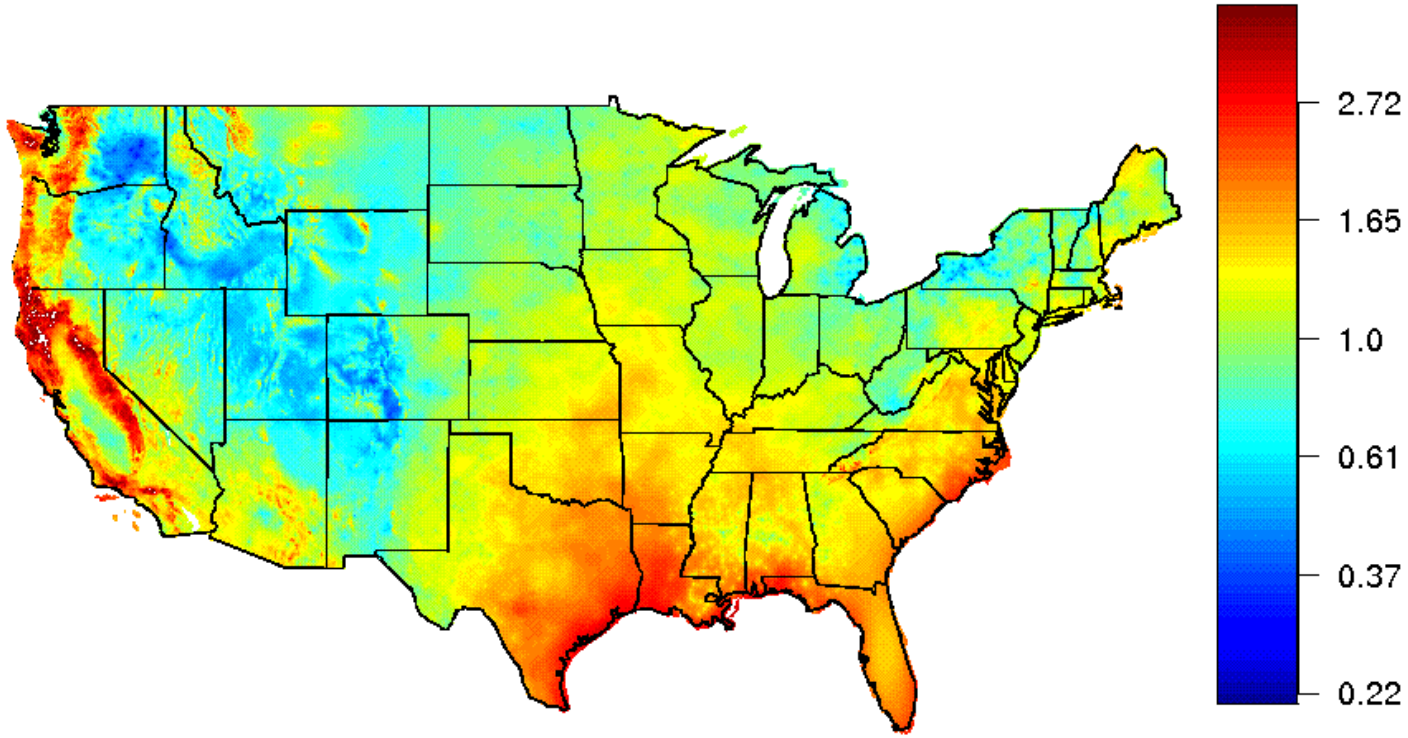
Continental USA divided into 19 regions



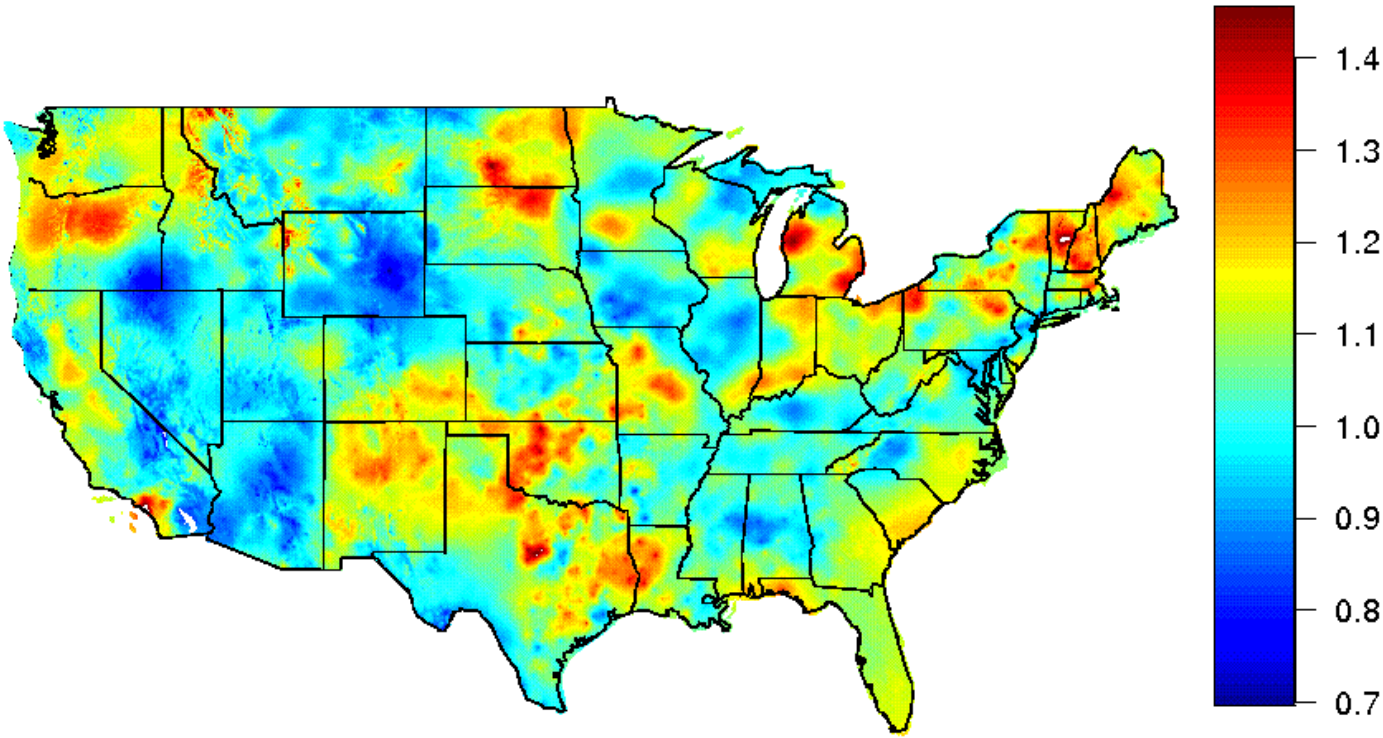
Trends across 19 regions (measured as change in log RV25) for 8 different seasonal models and one non-seasonal model with simple linear trends. Regional averaged trends by geometric weighted average approach.



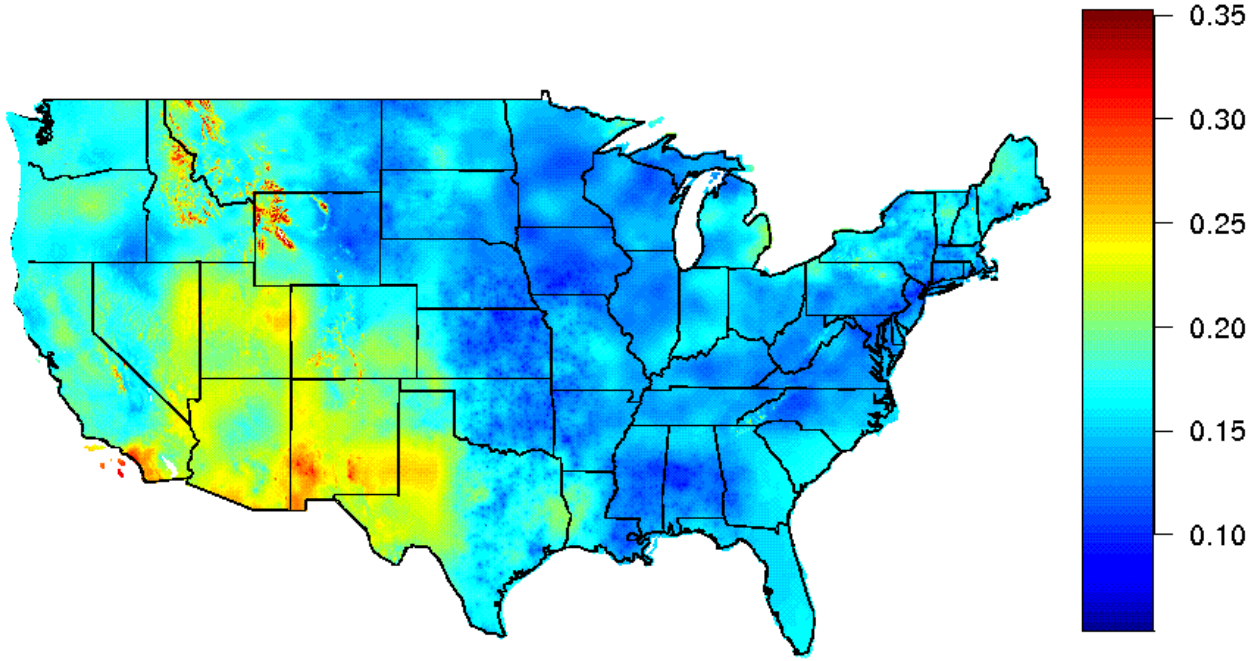
Map of 25-year return values (cm.) for the years 1970–1999



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



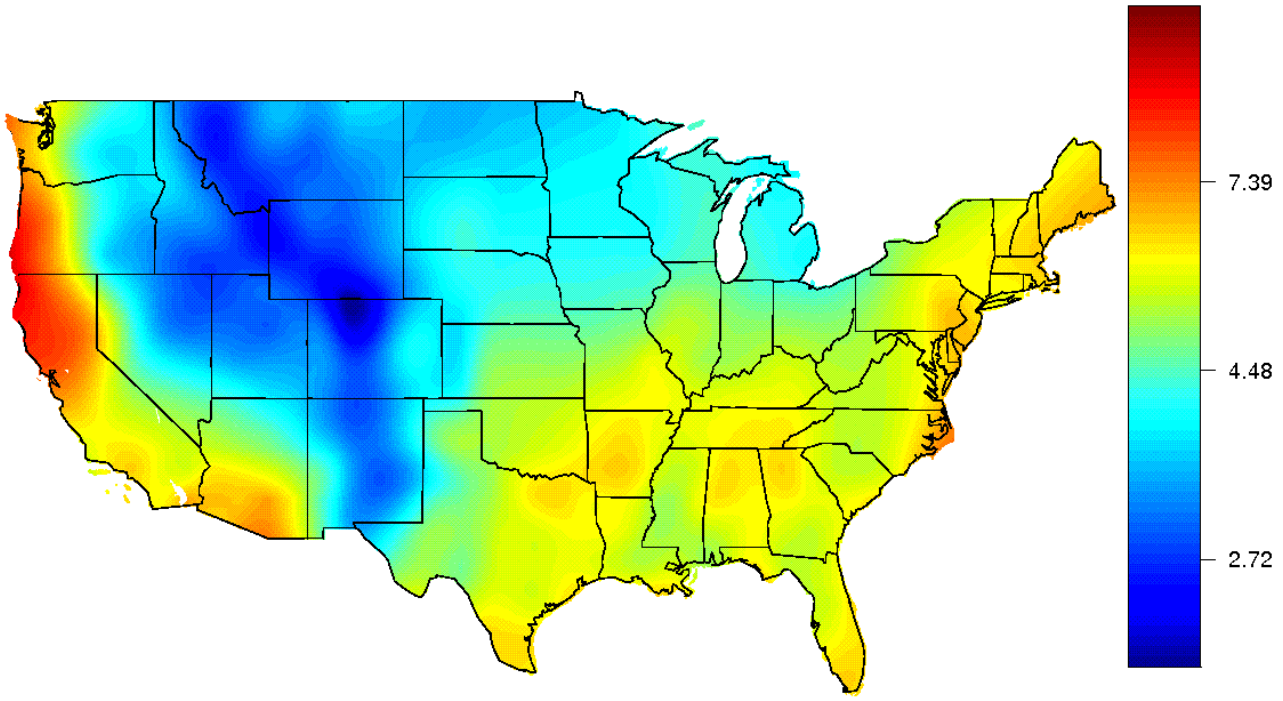
Root mean square prediction errors for map of ratios of 25-year return values in 1999 to those in 1970

	Change	RMSPE		Change	RMSPE
A	−0.01	.03	K	0.08***	.01
B	0.07**	.03	L	0.07***	.02
C	0.11***	.01	M	0.07***	.02
D	0.05***	.01	N	0.02	.03
E	0.13***	.02	O	0.01	.02
F	0.00	.02	P	0.07***	.01
G	−0.01	.02	Q	0.07***	.01
H	0.08***	.01	R	0.15***	.02
I	0.07***	.01	S	0.14***	.02
J	0.05***	.01			

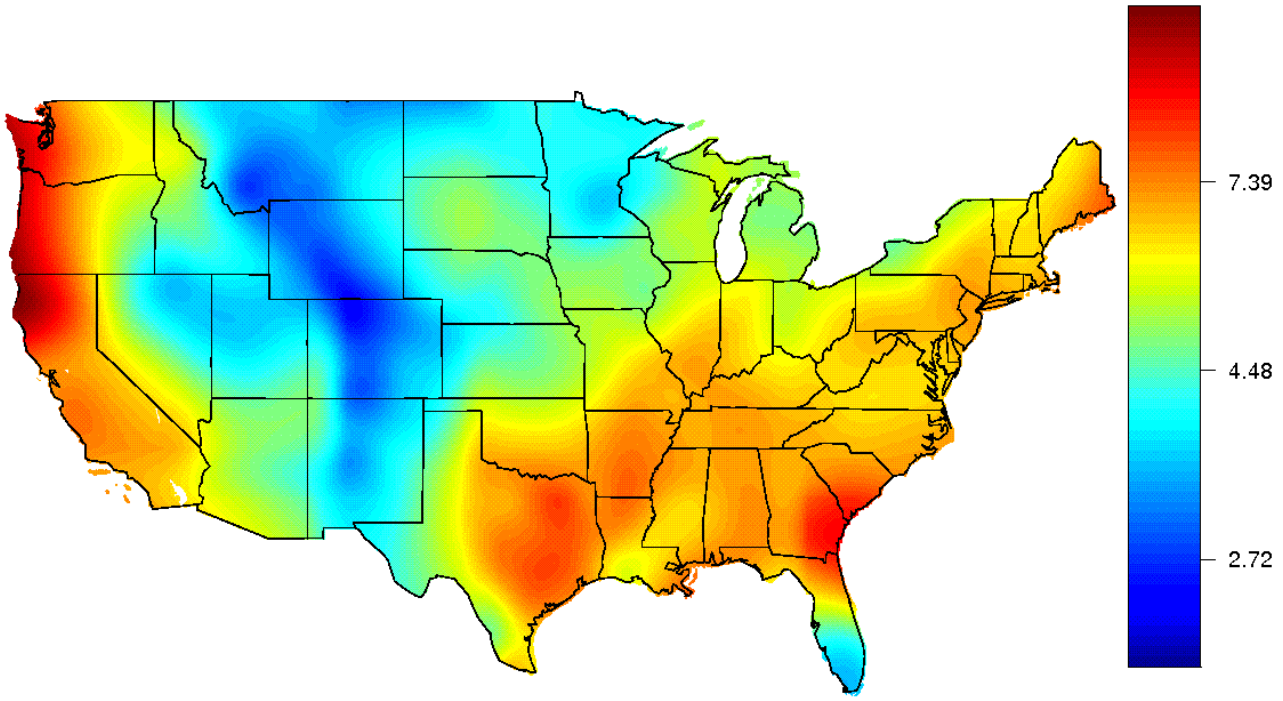
For each grid box, we show the mean change in log 25-year return value (1970 to 1999) and the corresponding standard error (RMSPE)

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

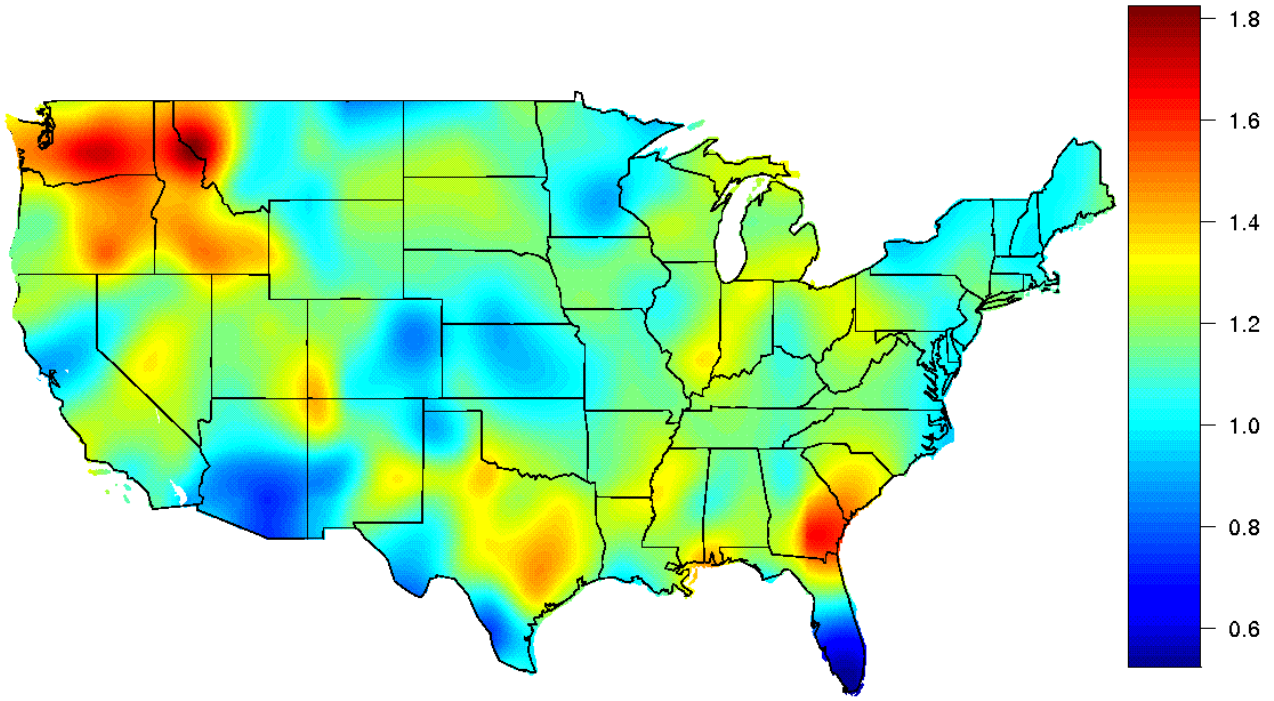
14 of 19 regions are statistically significant increasing: the remaining five are all in western states



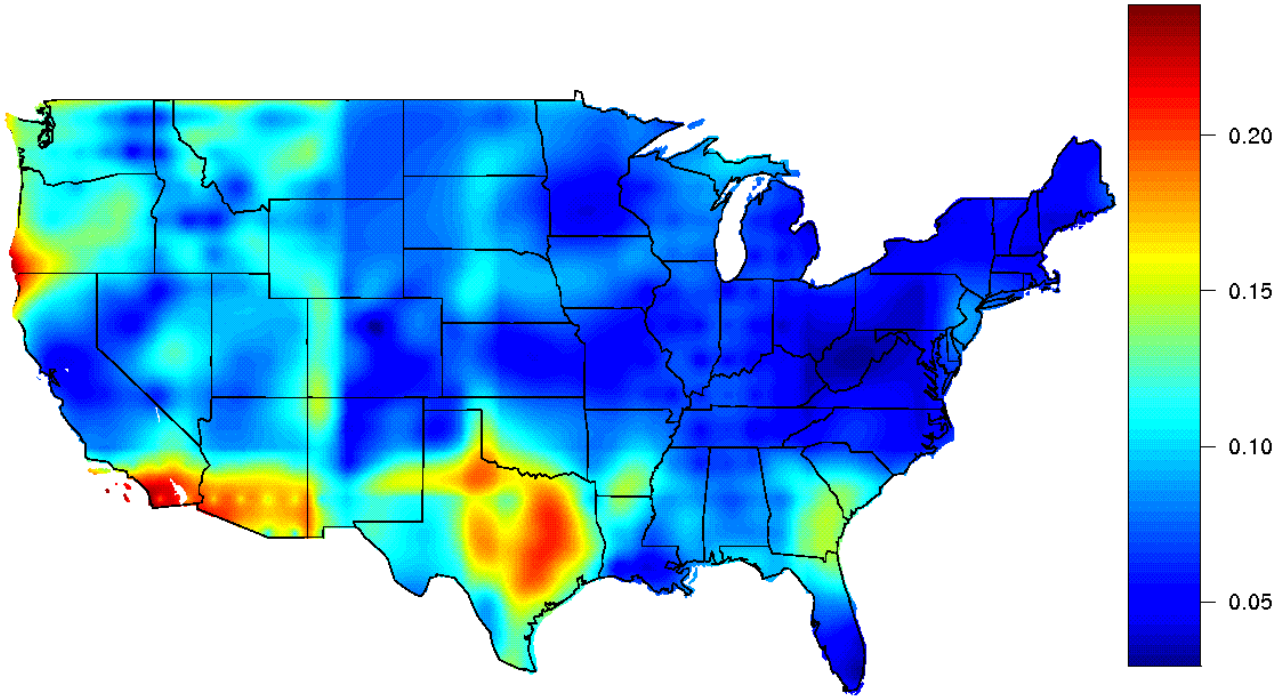
Return value map for CCSM data (cm.): 1970–1999



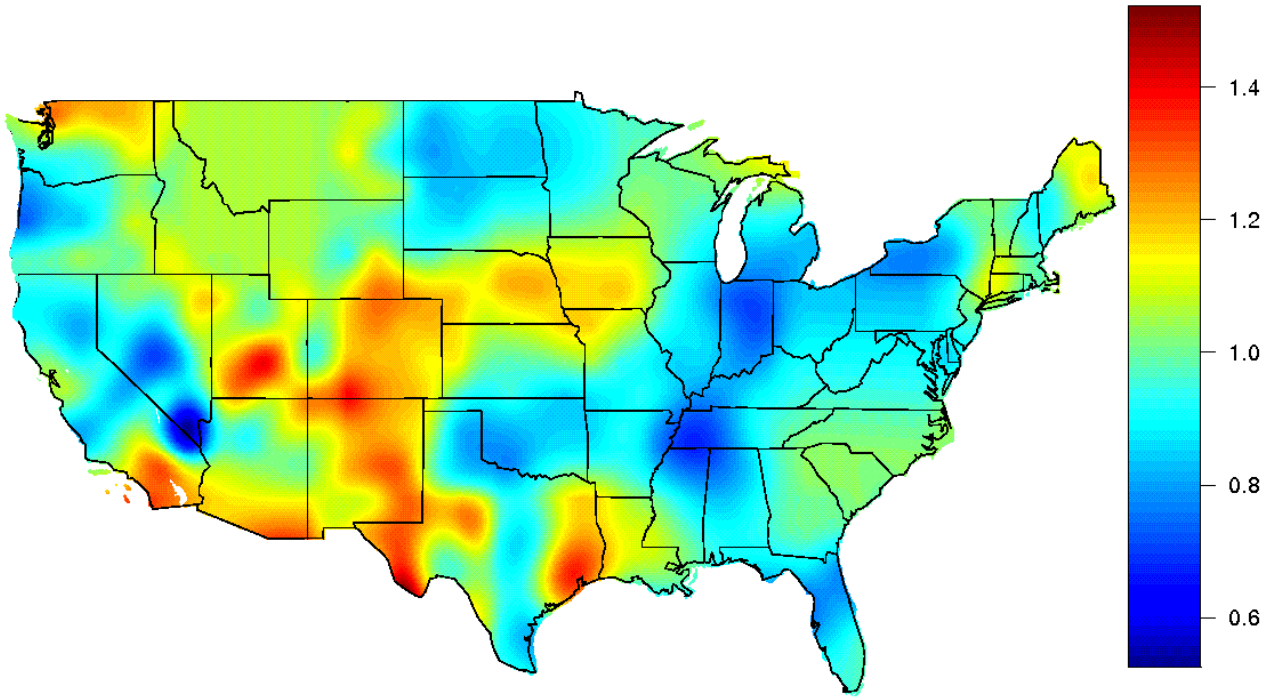
Return value map for CCSM data (cm.): 2070–2099



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

CONCLUSIONS FOR PRECIPITATION ANALYSIS

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. *But...* based on CCSM data there is a completely different spatial pattern and no overall increase

5. Projections to 2070–2099 show further strong increases but note caveat based on point 5

III. HEATWAVE ANALYSIS

(Joint with Michael Wehner, LBLL, and members of the SAMSI Spatial Extremes working group)

Methodology

Assume the probability distribution for the observation at time t is given by

$$F_t(x) \;=\; \exp\left\{-\left(1+\xi_t\frac{x-\mu_t}{\psi_t}\right)_+^{-1/\xi_t}\right\}$$

valid in $x \geq u$. Here $\mu_t,\ \psi_t,\ \xi_t$ are time-dependent parameters assumed to be defined in terms of covariates $x_{t,j},\ 0 \leq j \leq q$ by

$$\mu_t \;=\; \sum_{j=0}^{q} x_{t,j}\beta_j, \quad \log\psi_t \;=\; \sum_{j=0}^{q} x_{t,j}\gamma_j, \quad \xi_t \;=\; \sum_{j=0}^{q} x_{t,j}\delta_j,$$

where we define $x_{t,0} = 1$ for all t and $\beta_j, \gamma_j, \delta_j, j = 0, ..., q$ are the parameters that define the model. Usually at least some of the $\beta_j, \gamma_j, \delta_j$ parameters are pre-set to 0.

Suppose we are trying to use this model to calculate the probability of exceeding a *design value* u^* at some as yet unobserved time t^*. In the "single-event attribution" problem, this corresponds to an actual observed value of u^* in time t^*. Assume the associated GEV parameters for this time point are μ^*, ψ^*, ξ^*.

$$\mu^* = \sum_{j=0}^{q} x_j^* \beta_j, \quad \log \psi^* = \sum_{j=0}^{q} x_j^* \gamma_j, \quad \xi^* = \sum_{j=0}^{q} x_j^* \delta_j,$$

The associated probability p^* is given by

$$p^* = 1 - \exp\left\{ - \left(1 + \xi^* \frac{u^* - \mu^*}{\psi^*} \right)^{-1/\xi^*} \right\}, \qquad (\dagger)$$

so the obvious point estimator $\widehat{p^*}$ is obtained by substituting the maximum likelihood estimators $\widehat{\mu^*}$, $\widehat{\psi^*}$ and $\widehat{\xi^*}$ into equation (†).

To get interval estimates of p^*, we consider here the *profile likelihood* method. In this method, for a large number of candidate values of p^*, we repeat the estimation of the maximum likelihood estimators under the constraint (†). From this *profile likelihood function*, an asymptotically correct confidence interval for p^* is derived from the χ^2 distribution of a likelihood ratio test.

An alternative approach would be to assume prior distributions for all the unknown model parameters and then calculate the posterior distribution of p^* through a Bayesian calculation.

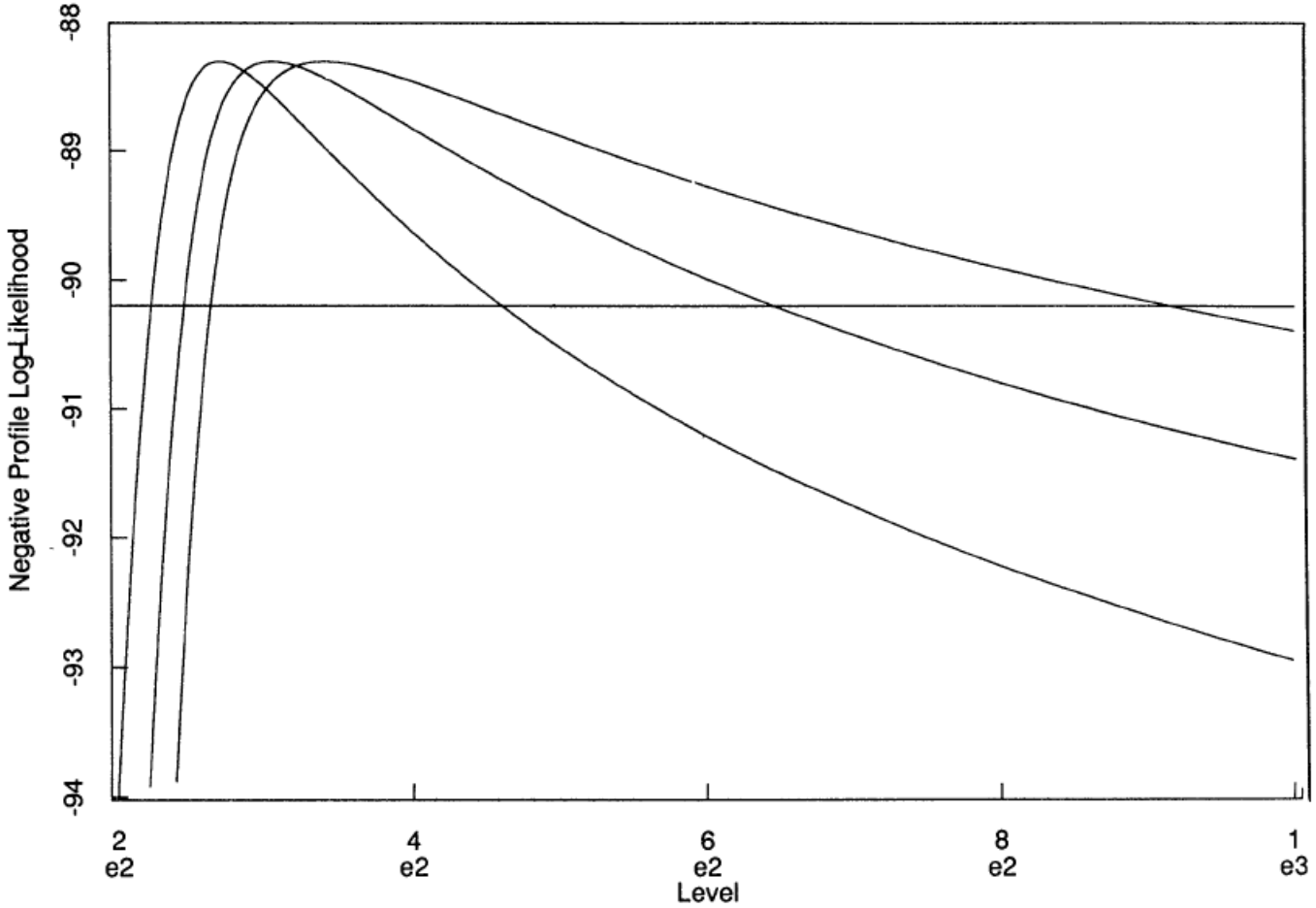


Fig. 9. Profile log-likelihood for 25-, 50- and 100-year return levels for the Nidd data (the horizontal line gives confidence intervals based on an asymptotic χ^2_1 distribution)

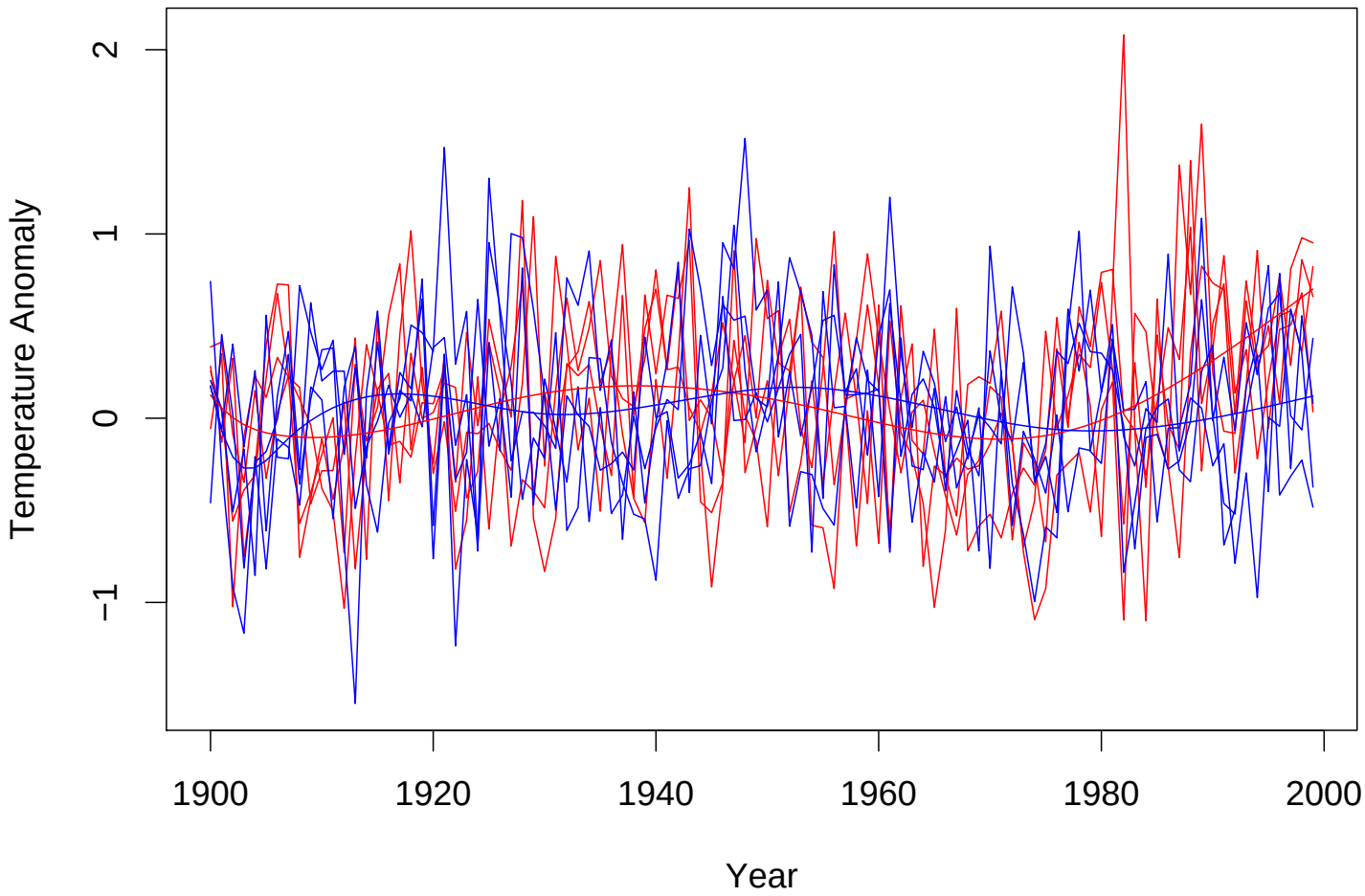
Illustration of profile likelihood method for calculating confidence intervals for a return value. From Davison and Smith (1990)

Results for UKMO data

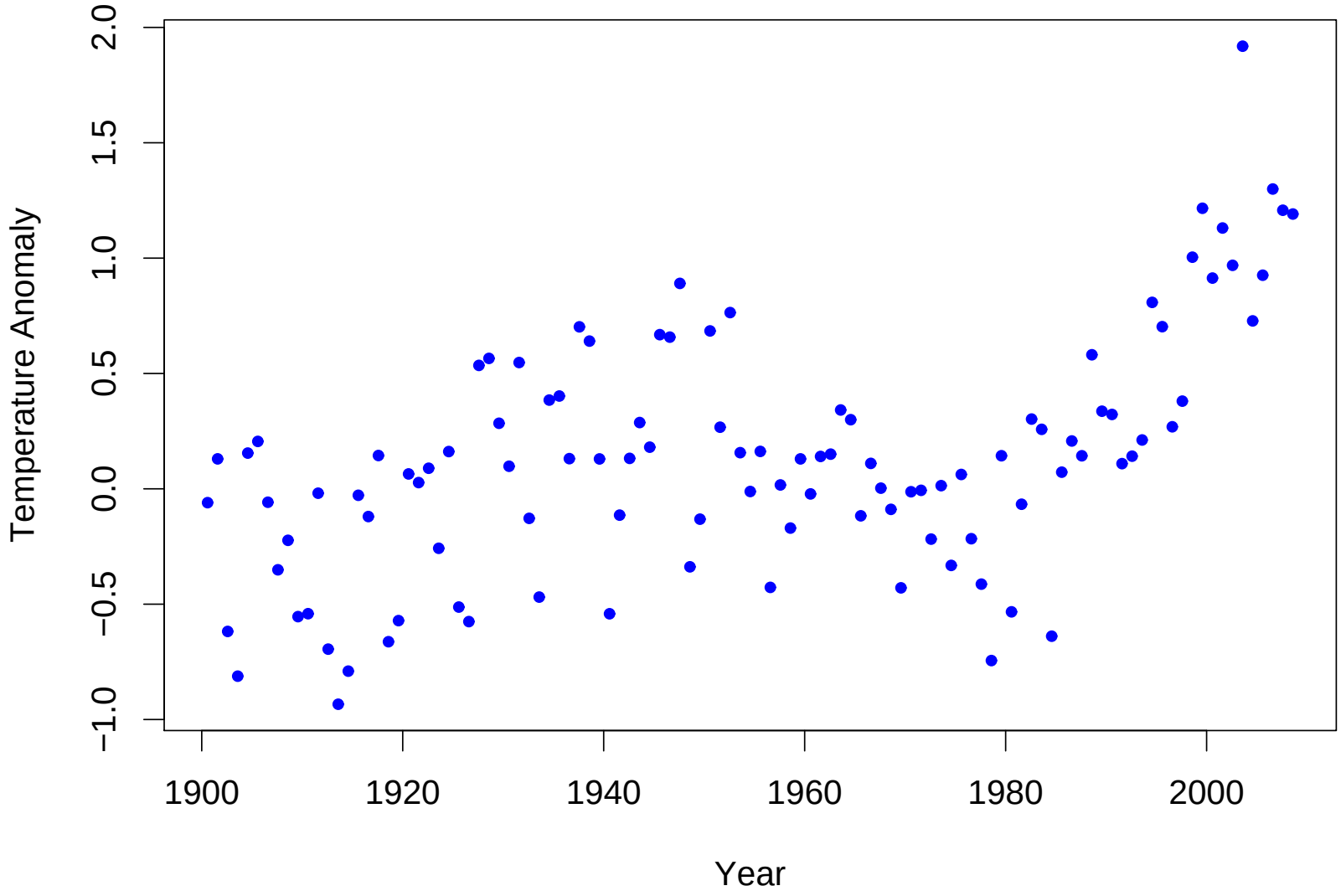
Data:

- 4 runs of UKMO model run data under an "all forcings" scenario, 1900–1999.
- 4 runs of UKMO model run data under an "natural forcings" scenario, 1900–1999.
- Real data from CRU, specifically, the "HADCRUT3VM" data series, 1900–2008.

For each series, I compute JJA means for each year, for the region 30–50^{o}N, 10^{o}W–40^{o}E, same as in Stott, Stone and Allen (SSA). All computed as anomalies from 1960–1999. No "data mask".



Sample traces for JJA averages for 1900–1999 over the spatial region 30–50^{o}N, 10^{o}W–40^{o}E, from four runs of a natural forcings model (blue) and four runs of an anthropogenic model (red). Smoothed curves represent spline-smoothed averages of the four model runs, computed separately for the natural-forcings and anthopogenic-forcings data.



Observational data; JJA averages for 1900–2008 over the spatial region 30–50^{o}N, 10^{o}W–40^{o}E.

Fitting an Extreme Value Model

- Two thresholds, 90th and 95th percentiles
- Regression models for μ_t; hold ψ_t, ξ_t fixed.
- Parametric trends for μ_t: polynomials or splines

Model	90% Threshold		95% Threshold	
	ALL	NAT	ALL	NAT
No Trend	232.3	231.7	149.5	142.8
Poly2	223.4	232.3	149.5	142.1
Poly4	219.4	232.9	144.2	139.9
Poly6	219.8	235.9	147.0	144.0
Spl2	223.8	232.0	149.5	142.2
Spl4	217.2	232.1	144.1	140.8
Spl6	215.7	234.3	145.2	143.7

Table 1: AIC scores for several extreme value models. I choose the Spl4 model, i.e. a natural splines expansion with 4 degrees of freedom.

Selection of a Design Value

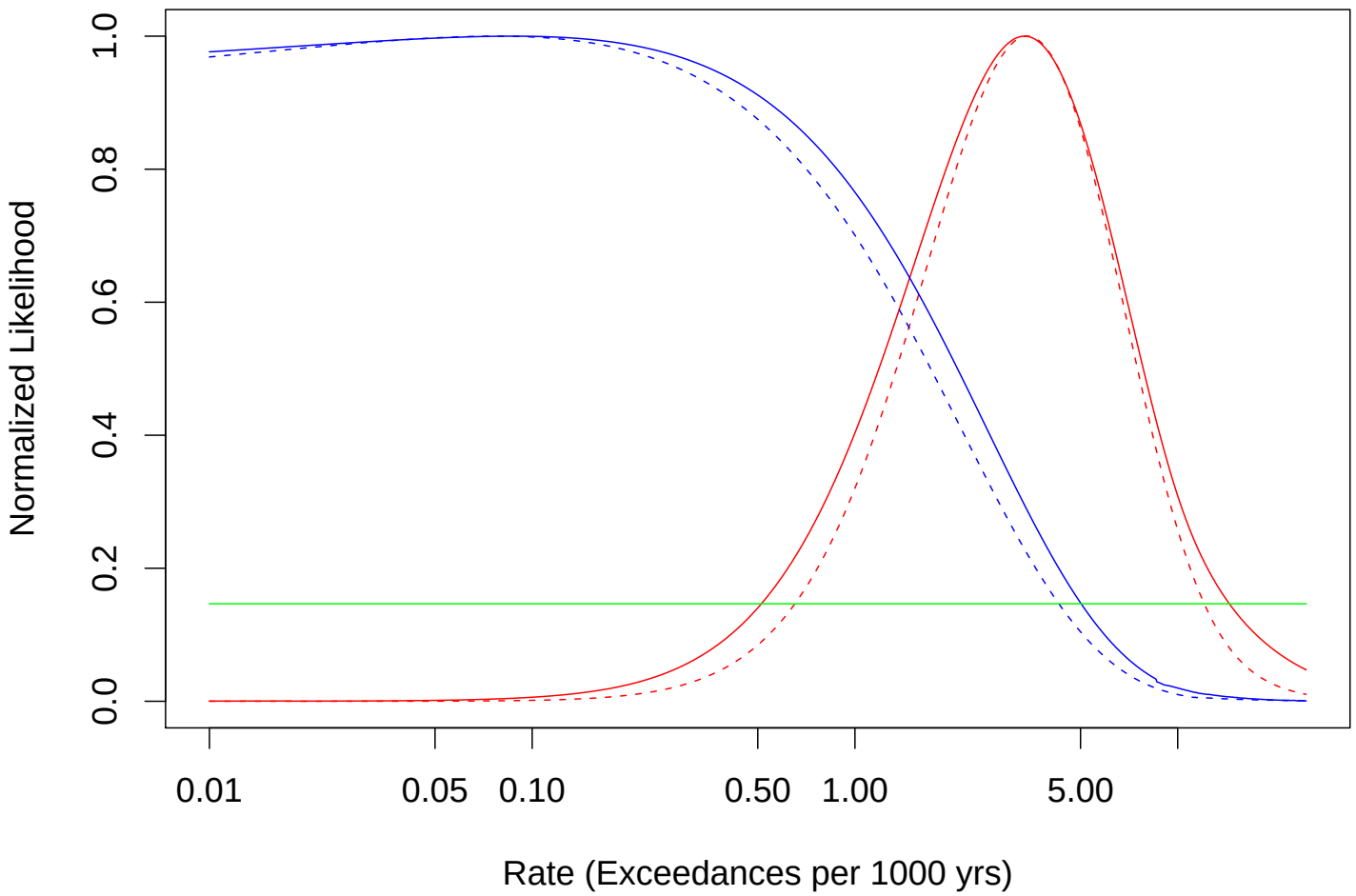
Two issues:

- which level in the observational series we take as best representing the 2003 heatwave
  - I take 1.6K (same as SSA)

- how to derive a corresponding value for the model series, taking into account possible scale mismatches between the observational and model series.
  - Based on variance matching, 1.6K for the observational series translates to 1.84K for the anthropogenic model runs and 1.71 K for the natural model runs

Analysis of Profile Likelihood I

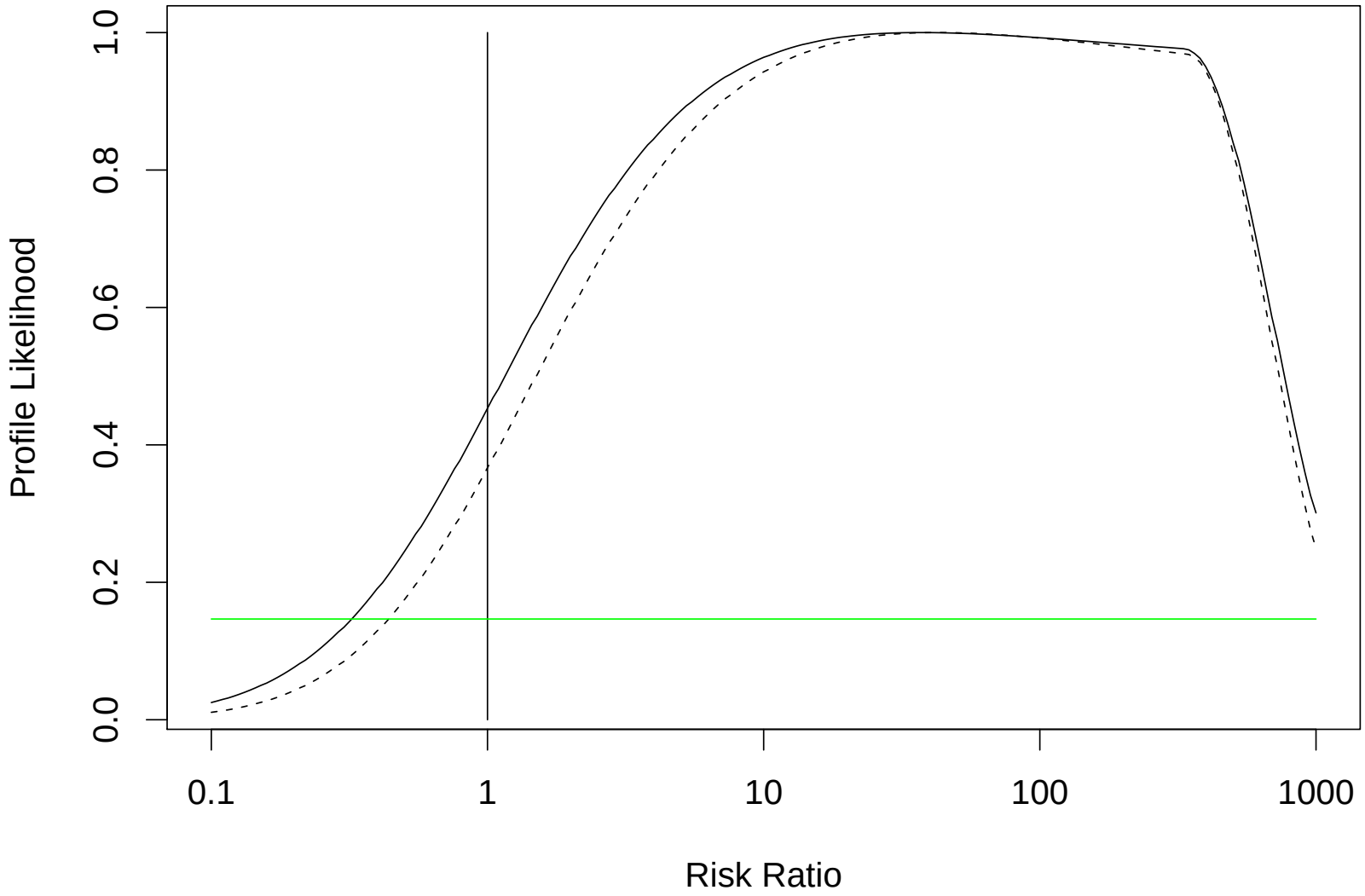
- Profile likelihoods associated with the two design values for the natural and anthropogenic models, for two different thresholds.
  - Discrepancies between profile likelihoods for the two thresholds, same model, are not important.
  - *But* 95% confidence intervals for the exceedance probabilities (determined by green cutoff line) are much wider than in SSA.



Normalized profile likelihoods for the rate of crossing the design value, computed for the natural forcings (blue curves) and the anthropogenic forcings (red curves) models, and for two different thresholds: 95% threshold (solid curve), 90% threshold (dashed curve).

Analysis of Profile Likelihood II

- Profile likelihood plot for the risk ratio, $\frac{P_1}{P_2}$, where P_1 and P_2 are the probability of exceeding the design value for the anthropogenic and natural forcings models respectively. SSA computed FAR= $1 - \frac{P_1}{P_2}$.
  - We *can* get a reasonable "likelihood" for the RR. But any reasonable calculation of a 95% confidence interval is going to be much wider than that presented by SSA.
  - SSA claim "...we estimate it is very likely (confidence level $> 90\%$) that human influence has at least doubled the risk of exceeding a threshold magnitude" ...
  - An equivalent statement would be that a 90% confidence interval for the RR is contained within $(2, \infty)$.
  - I believe that such an assessment understates the uncertainty in estimating RR.



Normalized profile likelihood for the relative risk; 95% threshold (solid curve), 90% threshold (dashed curve).

CONCLUSIONS

- This approach does not fully resolve data mismatch problem between observations and model output
- Profile likelihood method not only approach — consider also Bayesian methods
- How to model time trends is important
- Current results indicate that it's possible to estimate relative risks along the same lines as SSA, but that realistic confidence intervals would be much wider than they derived
- Probably need much more intensive model runs to derive usable interval estimates for the RR

SOME REFERENCES

Coles, S.G. (2001), *An Introduction to Statistical Modeling of Extreme Values*. Springer Verlag, New York.

Smith, R.L. (2003), Statistics of extremes, with applications in environment, insurance and finance. Chapter 1 of *Extreme Values in Finance, Telecommunications and the Environment*, edited by B. Finkenstadt and H. Rootzén, Chapman and Hall/CRC Press, London, pp. 1–78.
http://www.stat.unc.edu/faculty/rs/papers/RLS_Papers.html

Stott, P.A., Stone, D.A. and Allen, M.R. (2004), Human contribution to the European heatwave of 2003. *Nature* **432**, 610–614.