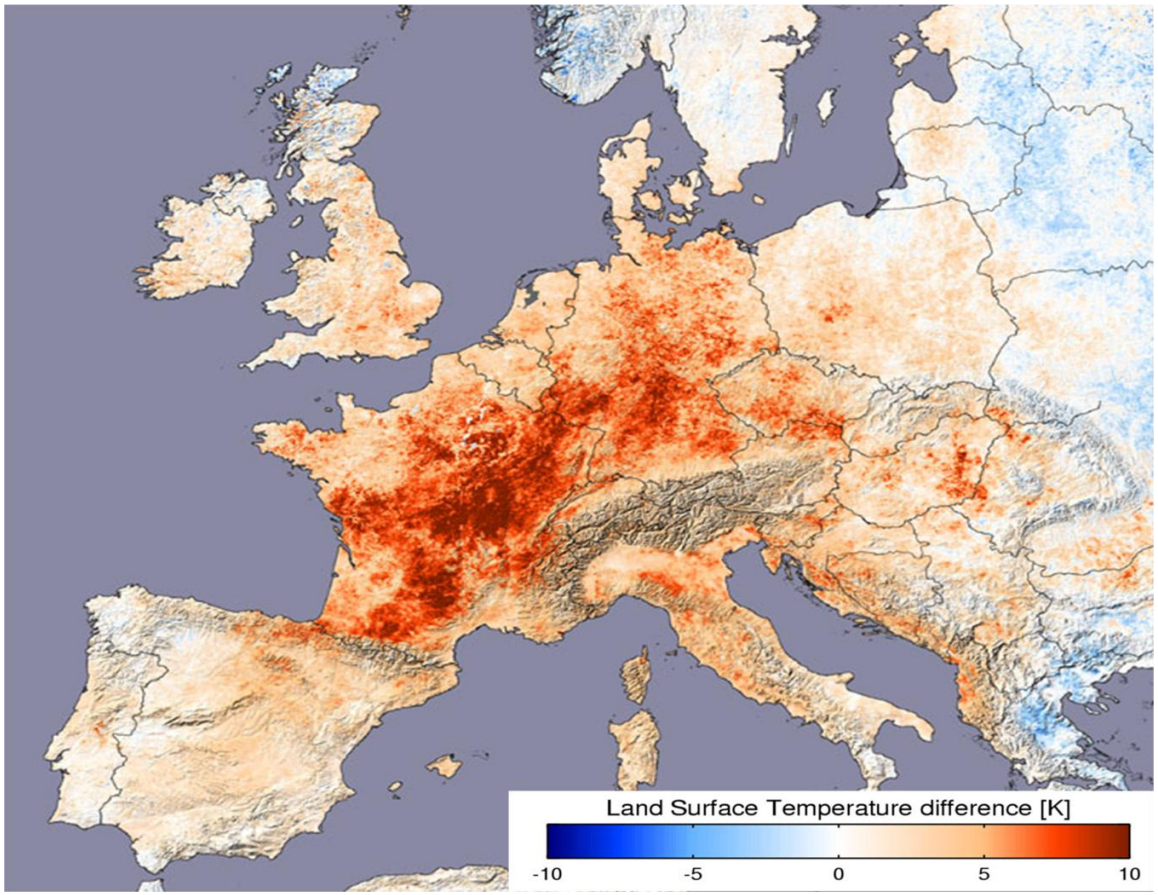
EXTREME VALUE THEORY

Richard L. Smith
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
rls@email.unc.edu

AMS Committee on Probability and Statistics
Short Course on Statistics of Extreme Events
Phoenix, January 11, 2009



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]

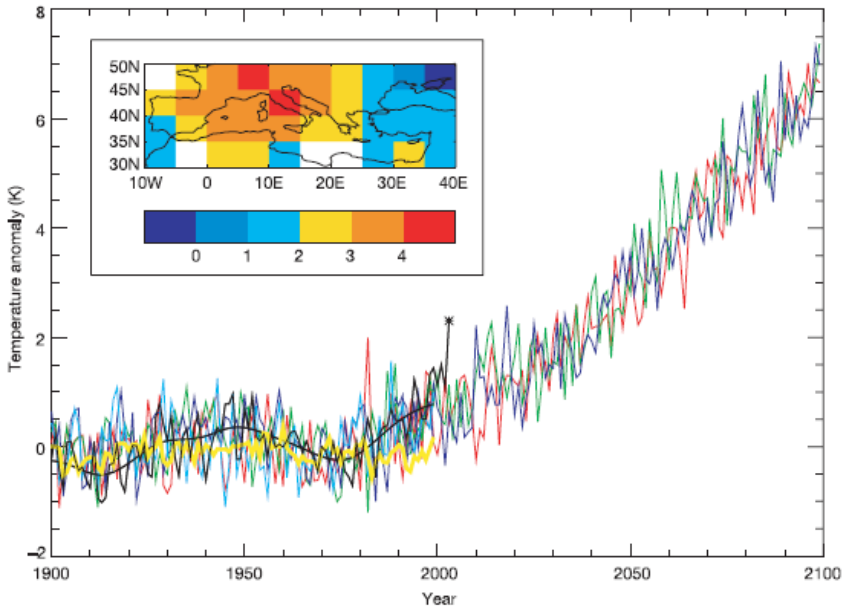


Figure 1 June–August temperature anomalies (relative to 1961–90 mean, in K) over the region shown in inset. Shown are observed temperatures (black line, with low-pass-filtered temperatures as heavy black line), modelled temperatures from four HadCM3 simulations including both anthropogenic and natural 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.

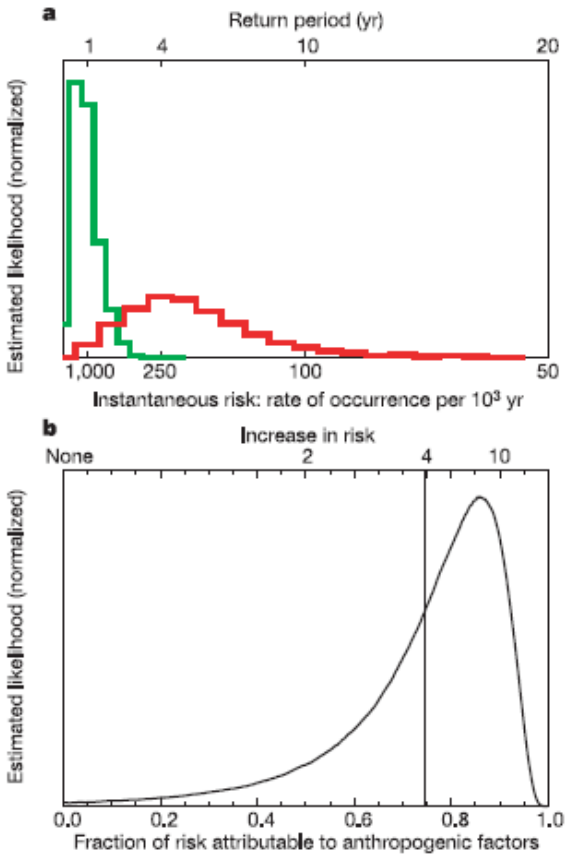


Figure 4 Change in risk of mean European summer temperatures exceeding the 1.6 K threshold. **a**, Histograms of instantaneous return periods under late-twentieth-century conditions in the absence of anthropogenic climate change (green line) and with anthropogenic climate change (red line). **b**, Fraction 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
- Diagnostics

II. Example: North Atlantic Storms

III. Example: European Heatwave

IV. Example: Trends in Extreme Rainfall Events

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

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

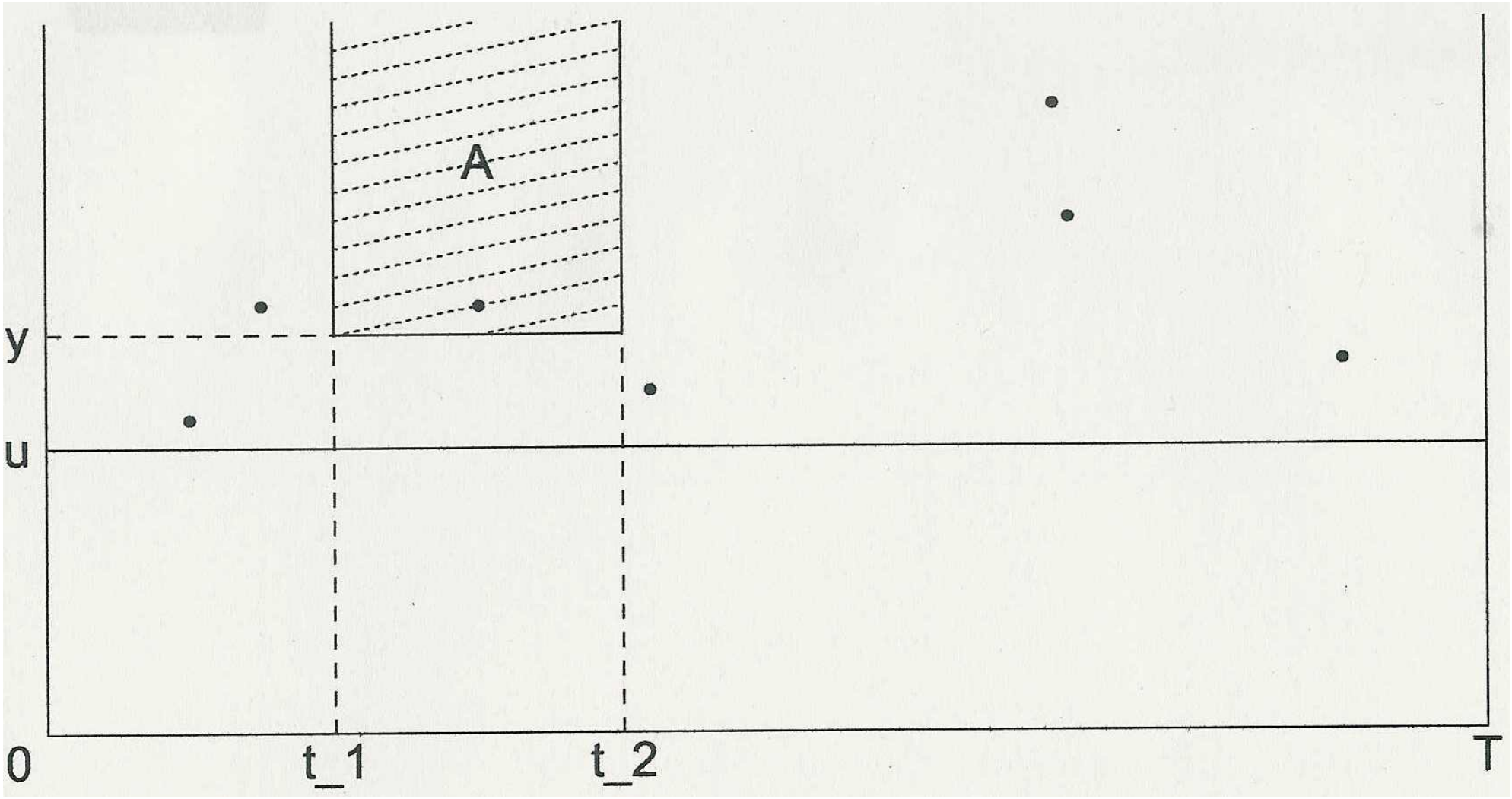


Illustration of point process model.

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

GEV log likelihood:

$$\ell \;=\; -N\log\psi - \left(\frac{1}{\xi}+1\right)\sum_i \log\left(1+\xi\frac{Y_i-\mu}{\psi}\right) - \sum_i \left(1+\xi\frac{Y_i-\mu}{\psi}\right)^{-1/\xi}$$

provided $1+\xi(Y_i-\mu)/\psi > 0$ for each i.

Poisson-GPD model:

$$\ell \;=\; N\log\lambda - \lambda T - N\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{N}\log\left(1+\xi\frac{Y_i}{\sigma}\right)$$

provided $1+\xi Y_i/\sigma > 0$ for all i.

The *method of maximum likelihood* states that we choose the parameters (μ,ψ,ξ) or (λ,σ,ξ) to maximize ℓ. These can be calculated numerically on the computer.

DIAGNOSTICS

Gumbel plots

QQ plots of residuals

Mean excess plot

Z and W plots

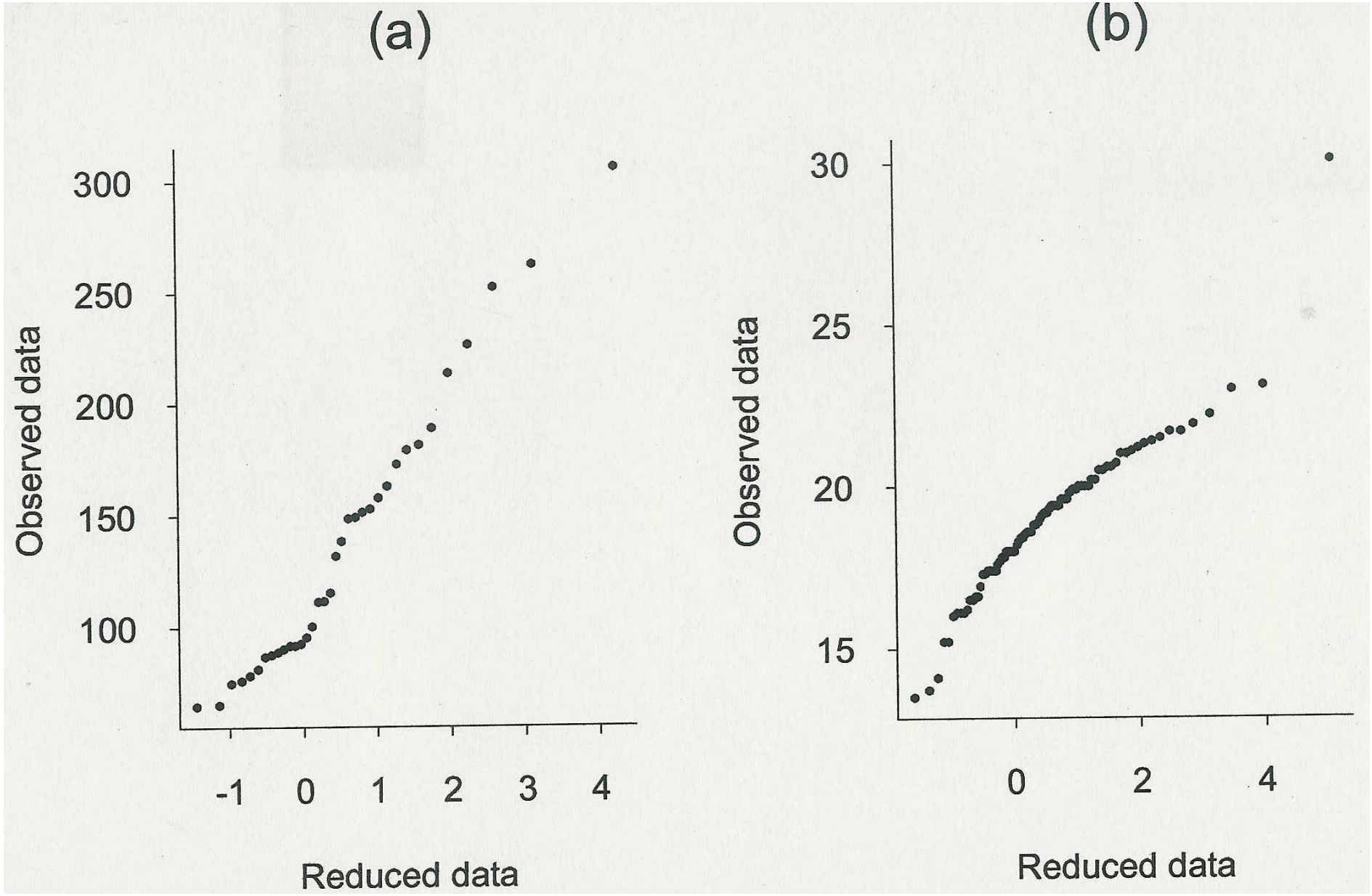
Gumbel plots

Used as a diagnostic for Gumbel distribution with annual maxima data. Order data as $Y_{1:N} \leq ... \leq Y_{N:N}$, then plot $Y_{i:N}$ against *reduced value* $x_{i:N}$,

$$x_{i:N} = -\log(-\log p_{i:N}),$$

$p_{i:N}$ being the i'th *plotting position*, usually taken to be $(i-\frac{1}{2})/N$.

A straight line is ideal. Curvature may indicate Fréchet or Weibull form. Also look for outliers.



Gumbel plots. (a) Annual maxima for River Nidd flow series. (b) Annual maximum temperatures in Ivigtut, Iceland.

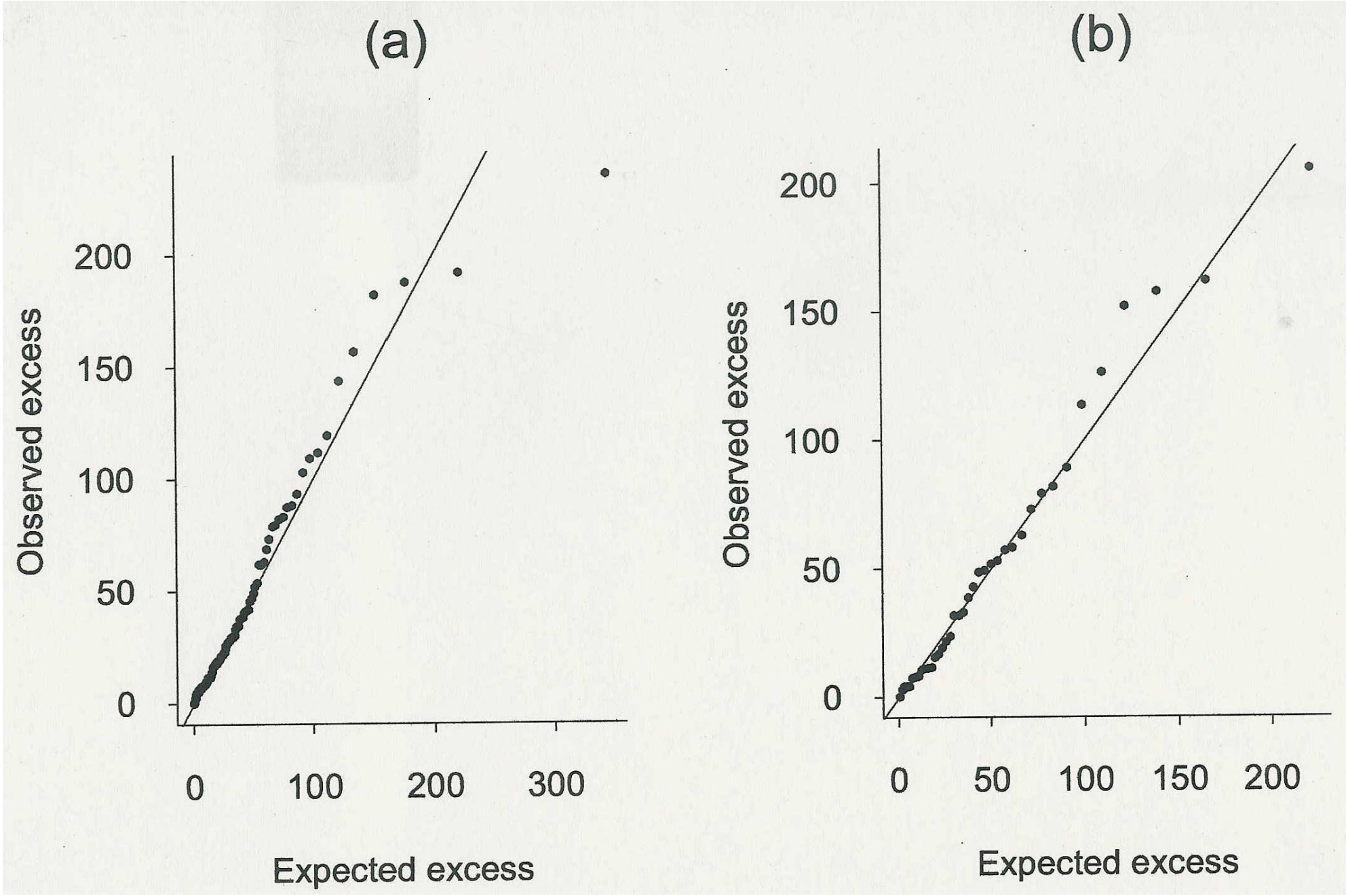
QQ plots of residuals

A second type of probability plot is drawn *after* fitting the model. Suppose $Y_1, ..., Y_N$ are IID observations whose common distribution function is $G(y;\theta)$ depending on parameter vector θ. Suppose θ has been estimated by $\hat{\theta}$, and let $G^{-1}(p;\theta)$ denote the inverse distribution function of G, written as a function of θ. A QQ (quantile-quantile) plot consists of first ordering the observations $Y_{1:N} \leq ... \leq Y_{N:N}$, and then plotting $Y_{i:N}$ against the reduced value

$$x_{i:N} = G^{-1}(p_{i:N};\hat{\theta}),$$

where $p_{i:N}$ may be taken as $(i-\frac{1}{2})/N$. If the model is a good fit, the plot should be roughly a straight line of unit slope through the origin.

Examples...



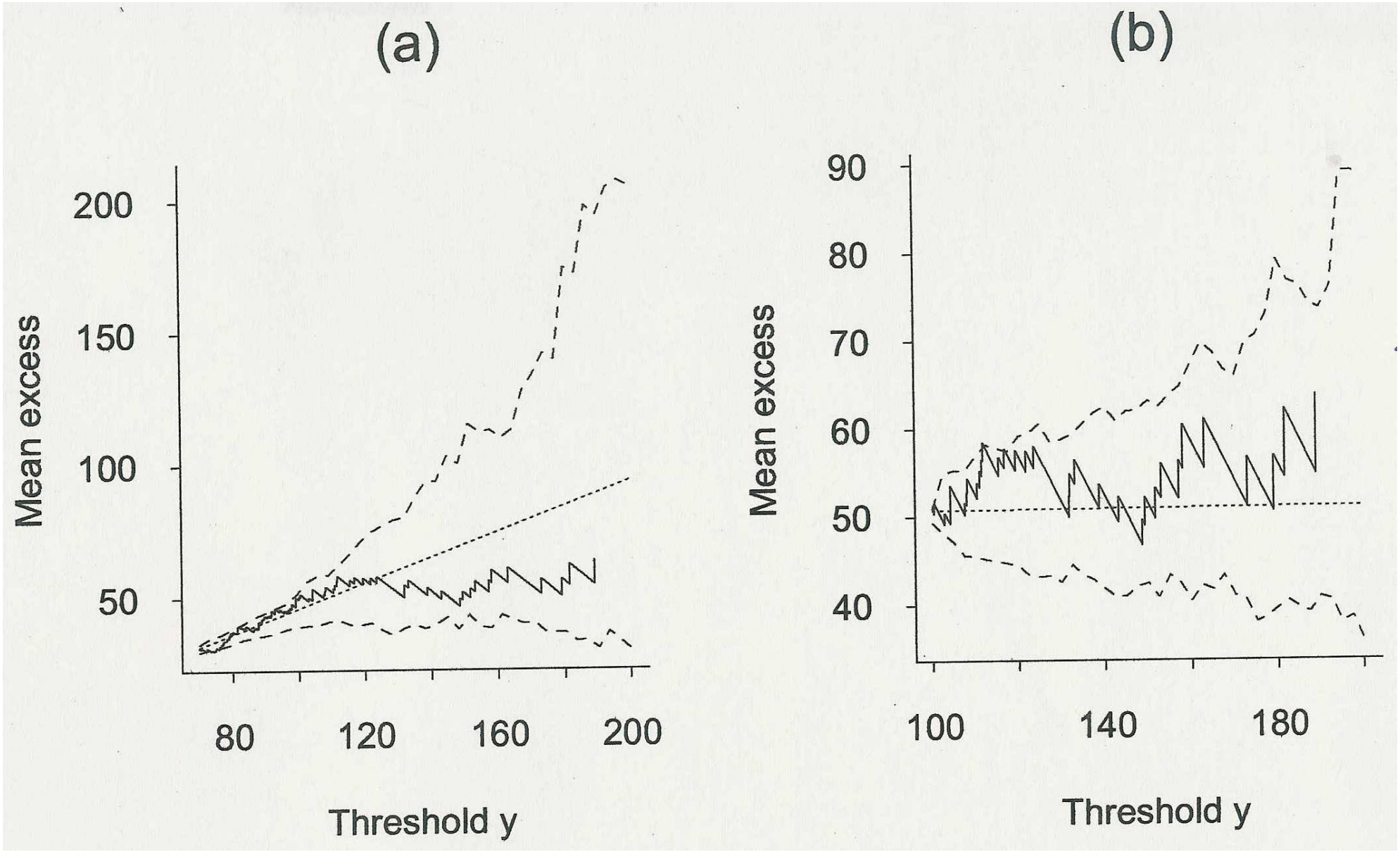
QQ plots for GPD, Nidd data. (a) $u = 70$. (b) $u = 100$.

Mean excess plot

Idea: for a sequence of values of w, plot the mean excess over w against w itself. If the GPD is a good fit, the plot should be approximately a straight line.

In practice, the actual plot is very jagged and therefore its "straightness" is difficult to assess. However, a Monte Carlo technique, *assuming* the GPD is valid throughout the range of the plot, can be used to assess this.

Examples...



Mean excess over threshold plots for Nidd data, with Monte Carlo confidence bands, relative to threshold 70 (a) and 100 (b).

Z- and W-statistic plots

Consider nonstationary model with μ_t, ψ_t, ξ_t dependent on t.

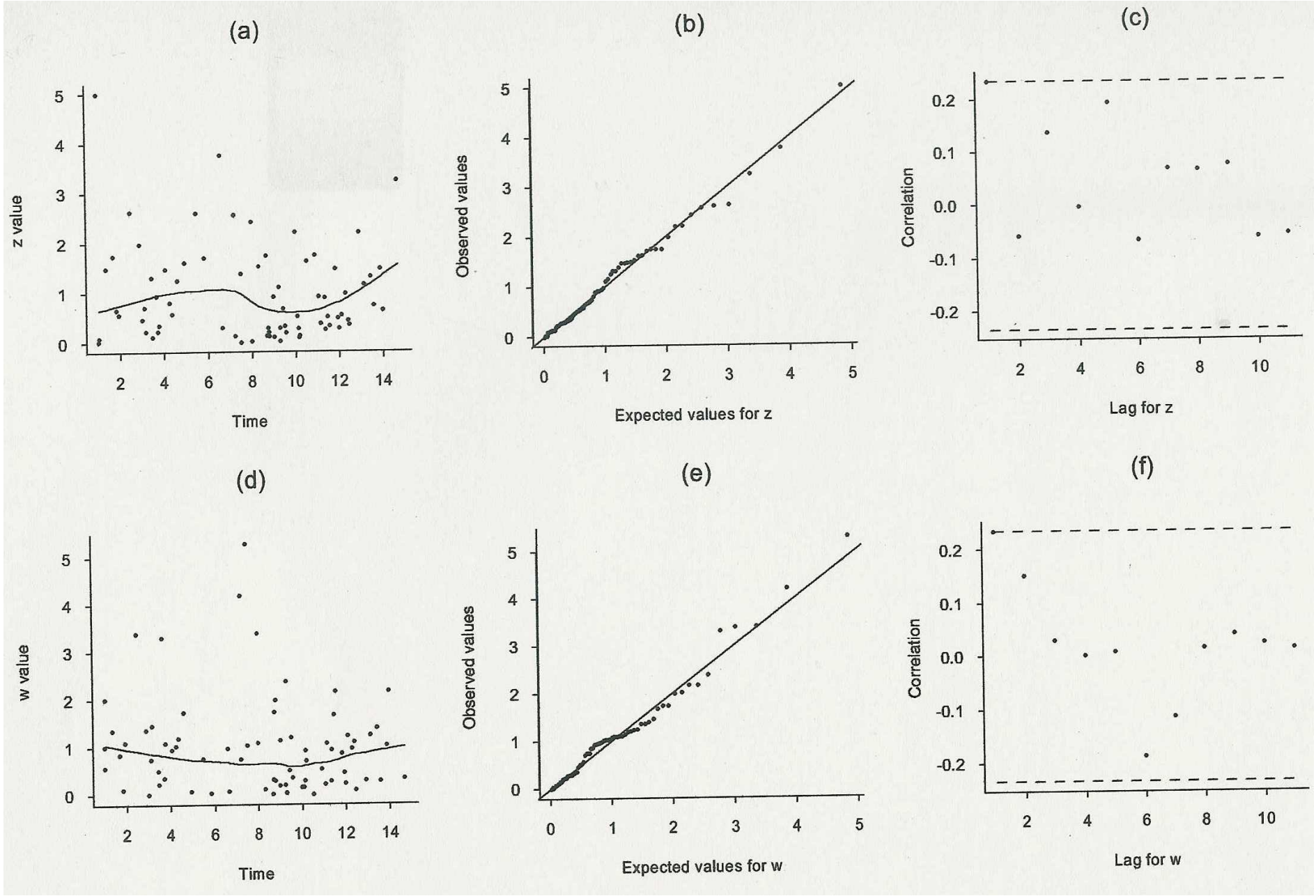
Z statistic based on intervals between exceedances T_k:

$$\begin{aligned} Z_k &= \int_{T_{k-1}}^{T_k} \lambda_u(s)ds, \\ \lambda_u(s) &= \{1 + \xi_s(u - \mu_s)/\psi_s)\}^{-1/\xi_s}. \end{aligned}$$

W statistic based on excess values: if Y_k is excess over threshold at time T_k,

$$W_k = \frac{1}{\xi_{T_k}} \log\left\{1 + \frac{\xi_{T_k} Y_k}{\psi_{T_k} + \xi_{T_k}(u - \mu_{T_k})}\right\}.$$

Idea: if the model is exact, both Z_k and W_k and i.i.d. exponential with mean 1. Can test this with various plots.



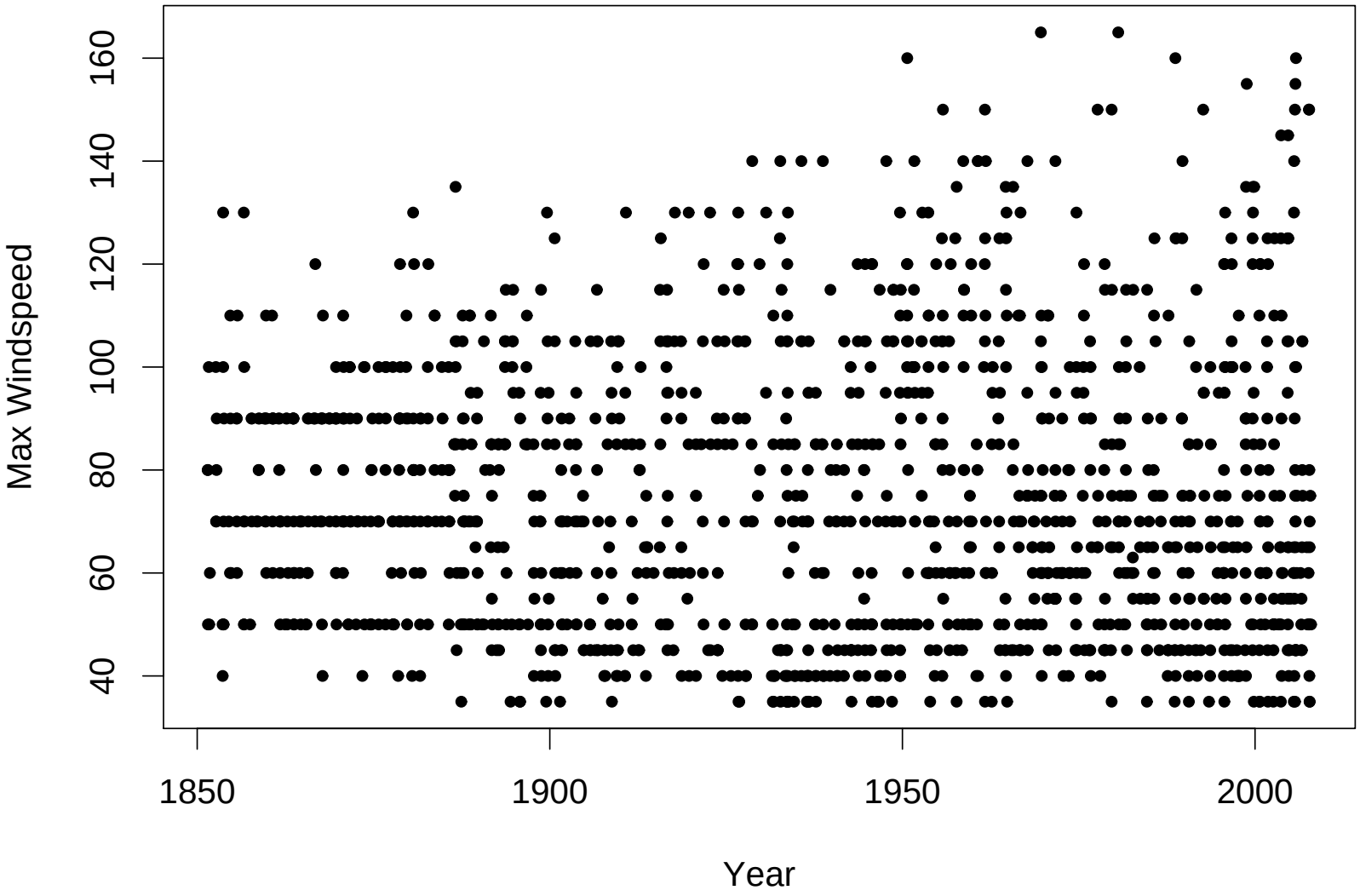
Diagnostic plots based on Z and W statistics for oil company insurance data ($u = 5$)

II. NORTH ATLANTIC CYCLONES

Data from HURDAT

Maximum windspeeds in all North Atlantic Cyclones from 1851–2007

TROPICAL CYCLONES FOR THE NORTH ATLANTIC



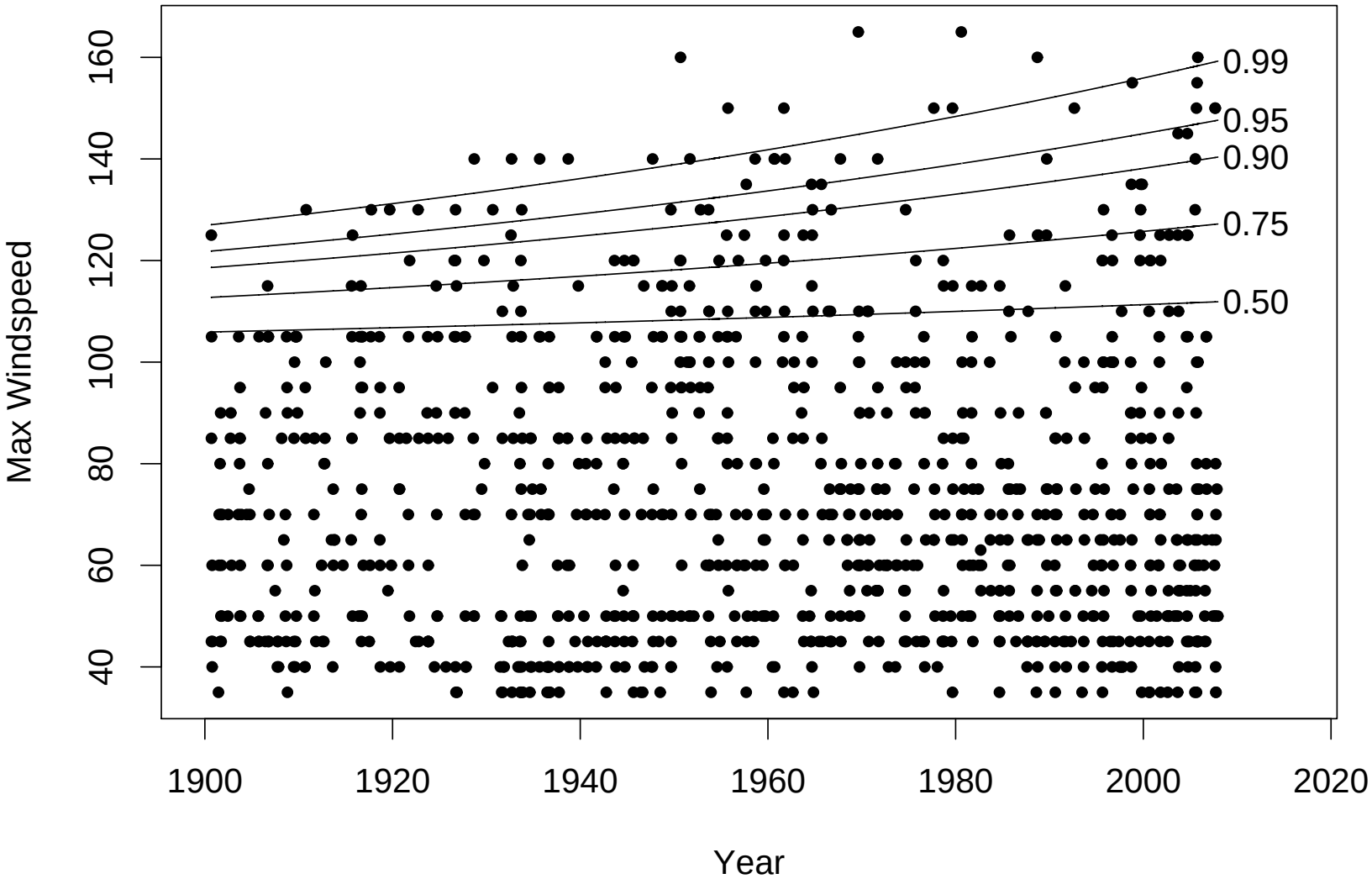
POT MODELS 1900–2007, u=102.5

Model	p	NLLH	NLLH+p
Gumbel	2	847.8	849.8
GEV	3	843.8	846.8
GEV, lin μ	4	834.7	838.7
GEV, quad μ	5	833.4	838.4
GEV, cubic μ	6	829.8	835.8
GEV, lin μ, lin $\log\psi$	5	828.0	833.0
GEV, quad μ, lin $\log\psi$	6	826.8	832.8
GEV, lin μ, quad $\log\psi$	6	827.2	833.2

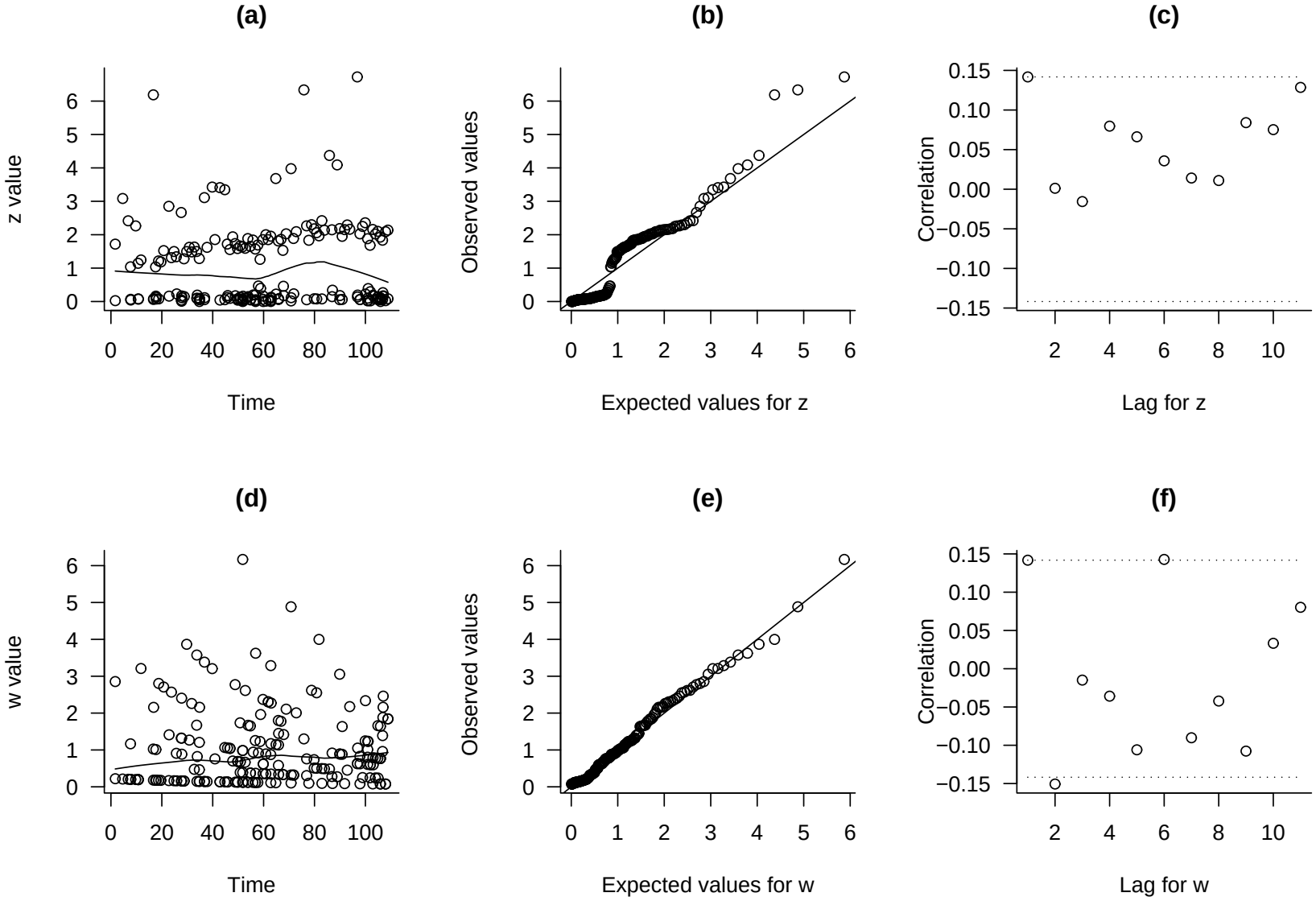
Fitted model: $\mu = \beta_0 + \beta_1 t$, $\log\psi = \beta_2 + \beta_3 t$, ξ const

	β_0	β_1	β_2	β_3	ξ
Estimate	102.5	0.0158	2.284	0.0075	−0.302
S.E.	2.4	0.049	0.476	0.0021	0.066

TROPICAL CYCLONES FOR THE NORTH ATLANTIC



Diagnostic Plots for Atlantic Cyclones



III. EUROPEAN HEATWAVE

Data:

5 model runs from CCSM 1871–2100, including anthropogenic forcing

2 model runs from UKMO 1861–2000, including anthropogenic forcing

1 model runs from UKMO 2001–2100, including anthropogenic forcing

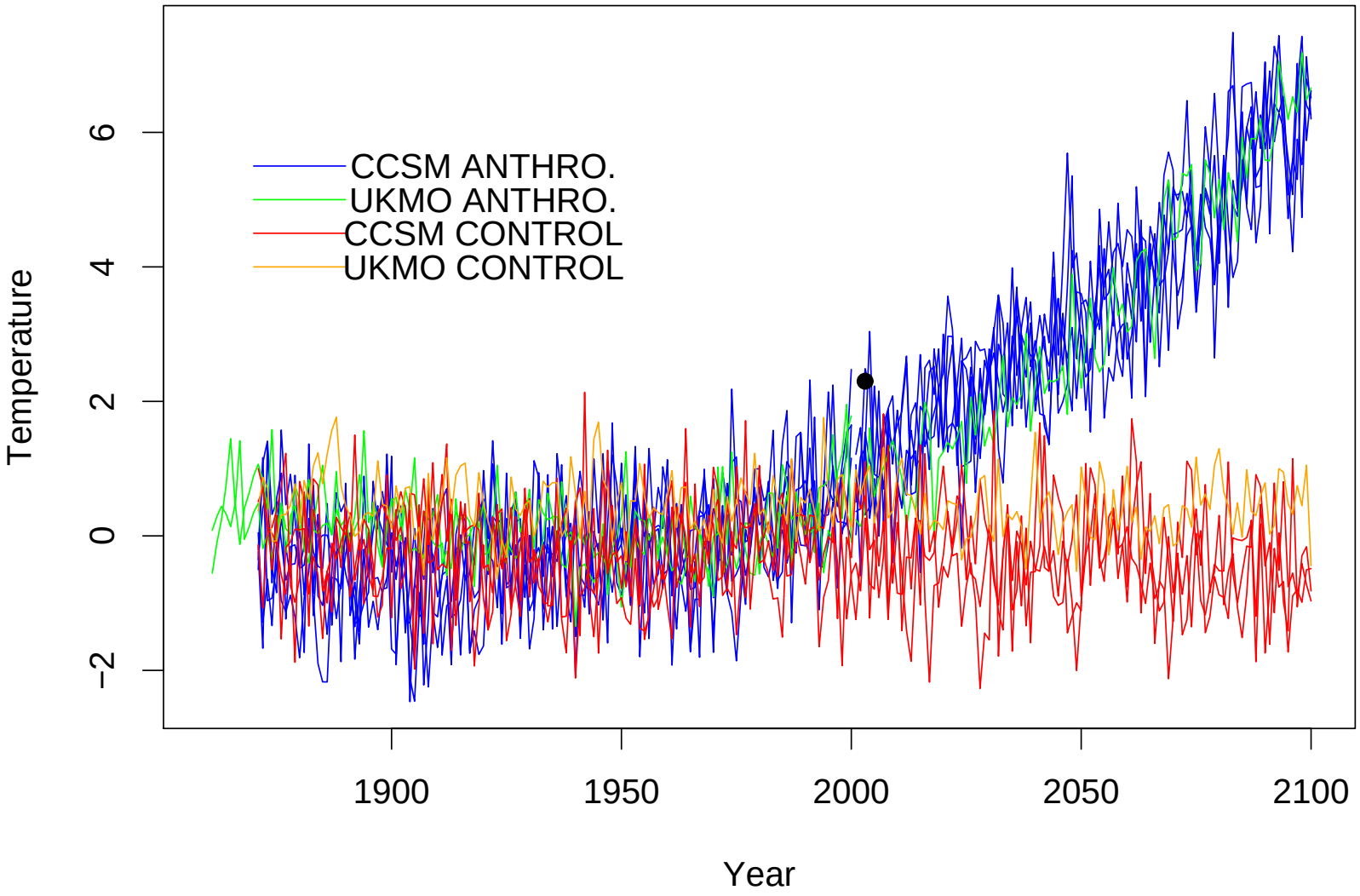
2 control runs from CCSM, 230+500 years

2 control runs from UKMO, 341+81 years

All model data have been calculated for the grid box from 30–50^{o} N, 10^{o} W–40^{o} E, annual average temperatures over June–August

Expressed an anomalies from 1961–1990, similar to Stott, Stone and Allen (2004)

CLIMATE MODEL RUNS: ANOMALIES FROM 1961−1990



Method:

Fit POT models with various trend terms to the anthropogenic model runs, 1861–2010

Also fit trend-free model to control runs ($\mu = 0.176$, $\log \psi = -1.068$, $\xi = -0.068$)

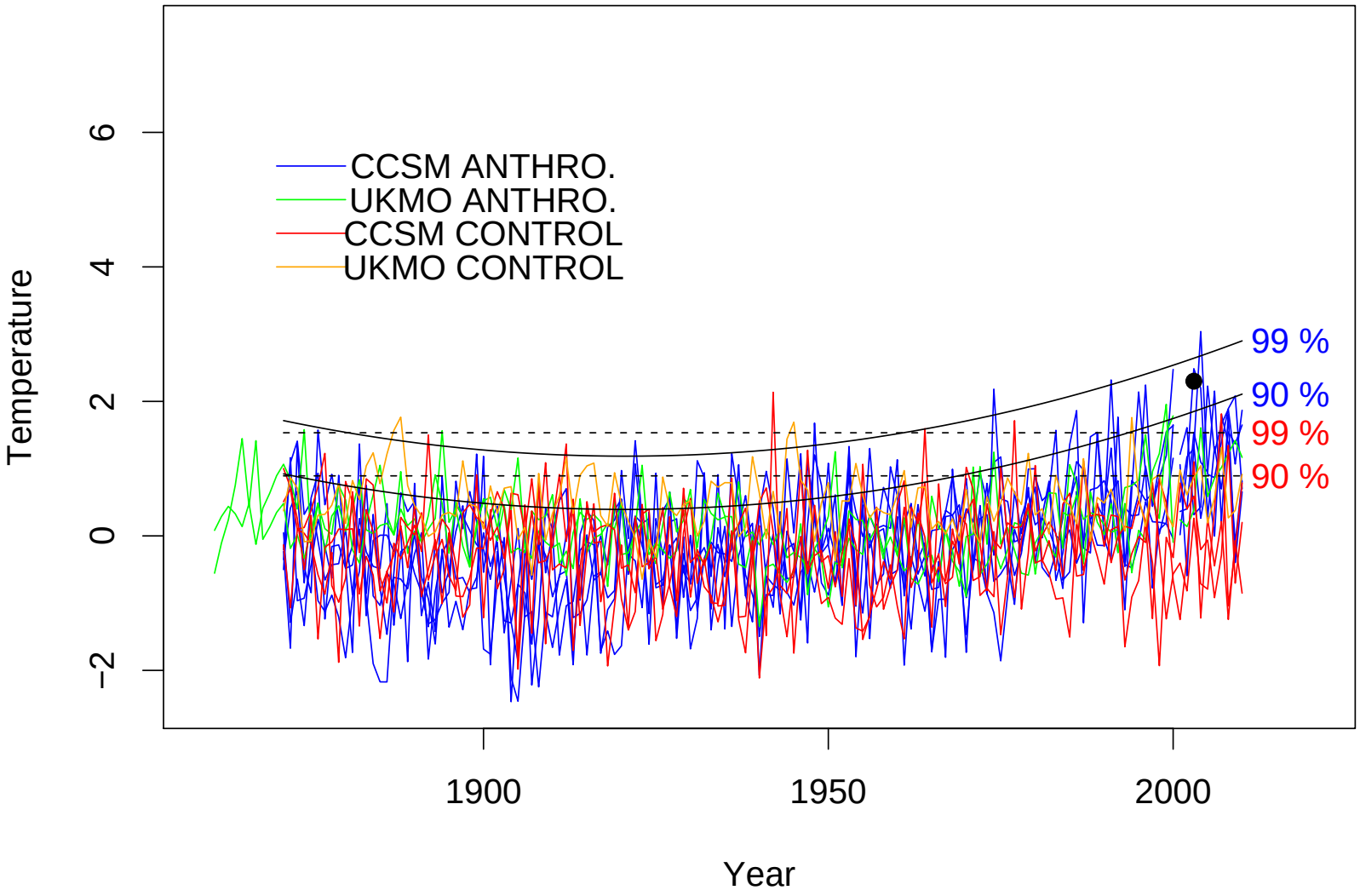
POT MODELS 1861–2010, u=1

Model	p	NLLH	NLLH+p
Gumbel	2	349.6	351.6
GEV	3	348.6	351.6
GEV, lin μ	4	315.5	319.5
GEV, quad μ	5	288.1	293.1
GEV, cubic μ	6	287.7	293.7
GEV, quart μ	7	285.1	292.1
GEV, quad μ, lin $\log\psi$	6	287.9	293.9
GEV, quad μ, quad $\log\psi$	7	287.0	294.9

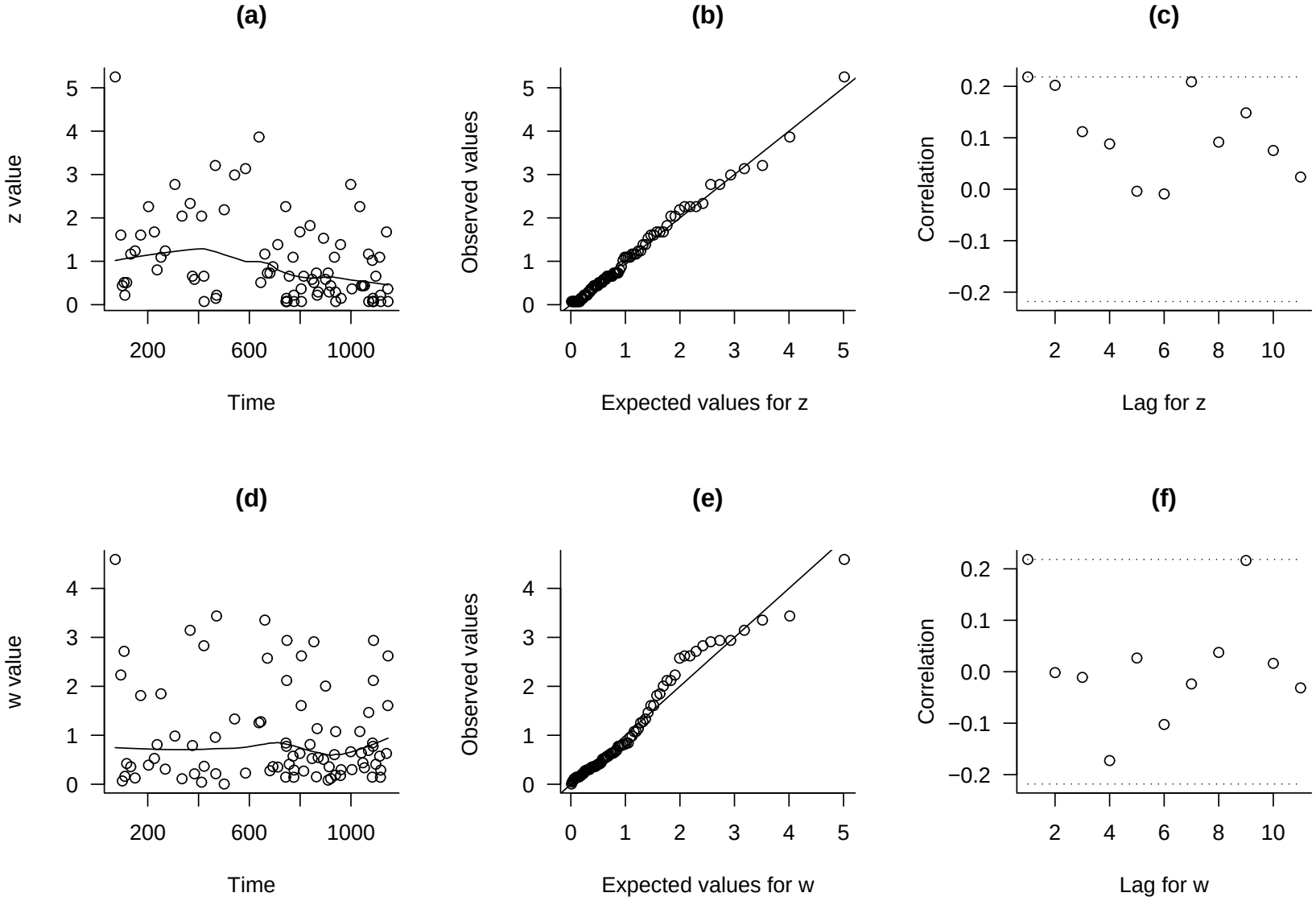
Fitted model: $\mu = \beta_0 + \beta_1 t + \beta_2 t^2$, ψ, ξ const

	β_0	β_1	β_2	$\log\psi$	ξ
Estimate	–0.187	–0.030	0.000215	0.047	0.212
S.E.	0.335	0.0054	0.00003	0.212	0.067

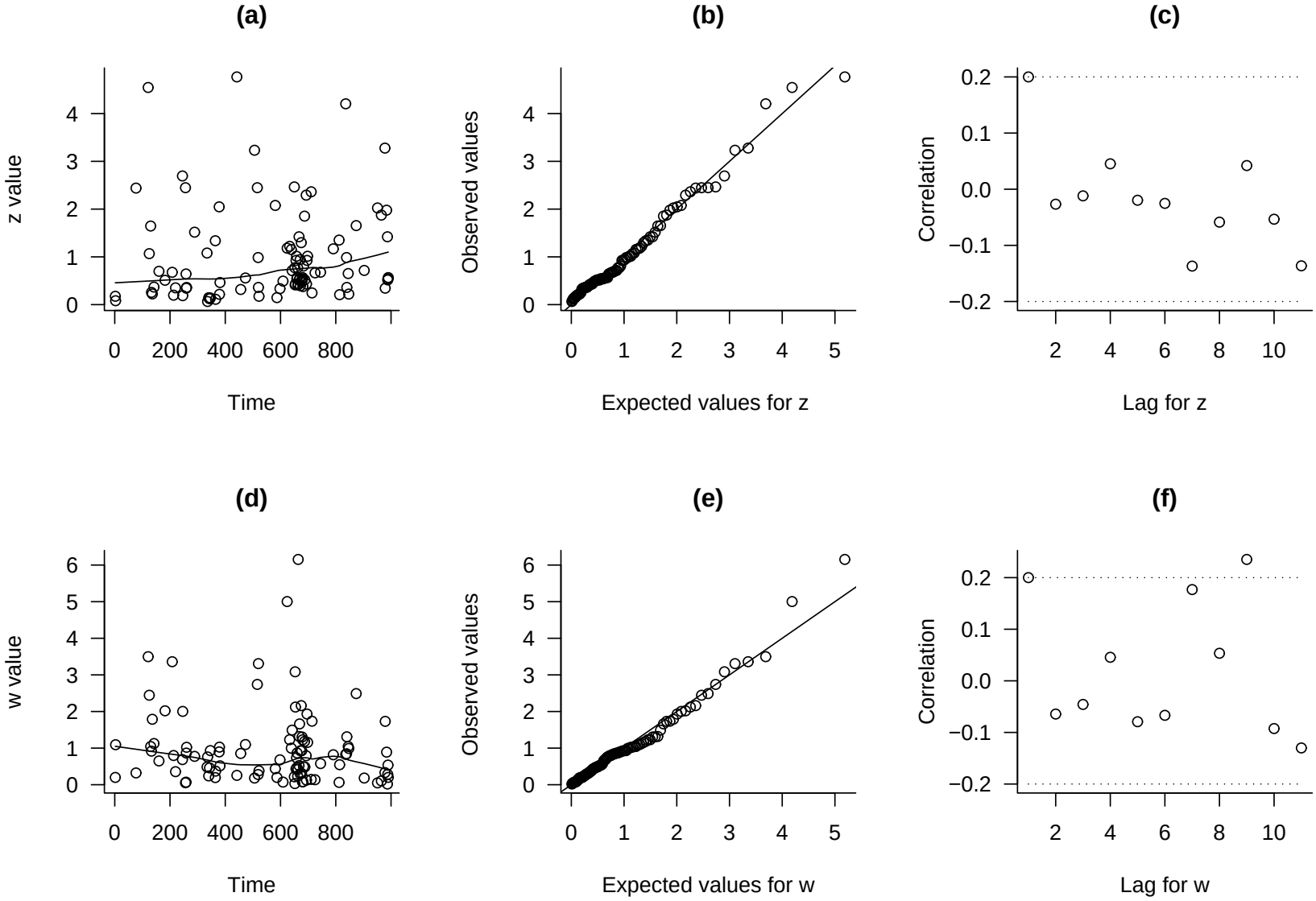
CLIMATE MODEL RUNS: ANOMALIES FROM 1961−1990



Diagnostic Plots for Temperatures (Control)



Diagnostic Plots for Temperatures (Anthropogenic)



We now estimate the probabilities of crossing various thresholds in 2003.

Express answer as N=1/(exceedance probability)

Threshold 2.3:

N=3024 (control), N=29.1 (anthropogenic)

Threshold 2.6:

N=14759 (control), N=83.2 (anthropogenic)

IV. 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, but in observational and model-generated data.

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, to appear *JASA*; Sang and Gelfand, submitted)

This discussion uses extreme value methodology to look for trends

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 (A1B scenario)

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

EXTREME VALUES METHODOLOGY

The essential idea is to fit a probability model to the exceedances over a high threshold at each of ≈5000 data sites, and then to combine data across sites using spatial statistics.

The model at each site is based on the *generalized extreme value distribution*, interpreted as an approximate tail probability in the right hand tail of the distribution.

$$\Pr\{Y \geq y\} \approx \delta_t \left(1 + \xi \frac{y-\mu}{\psi}\right)_+^{-1/\xi} \quad \text{for large } y,$$

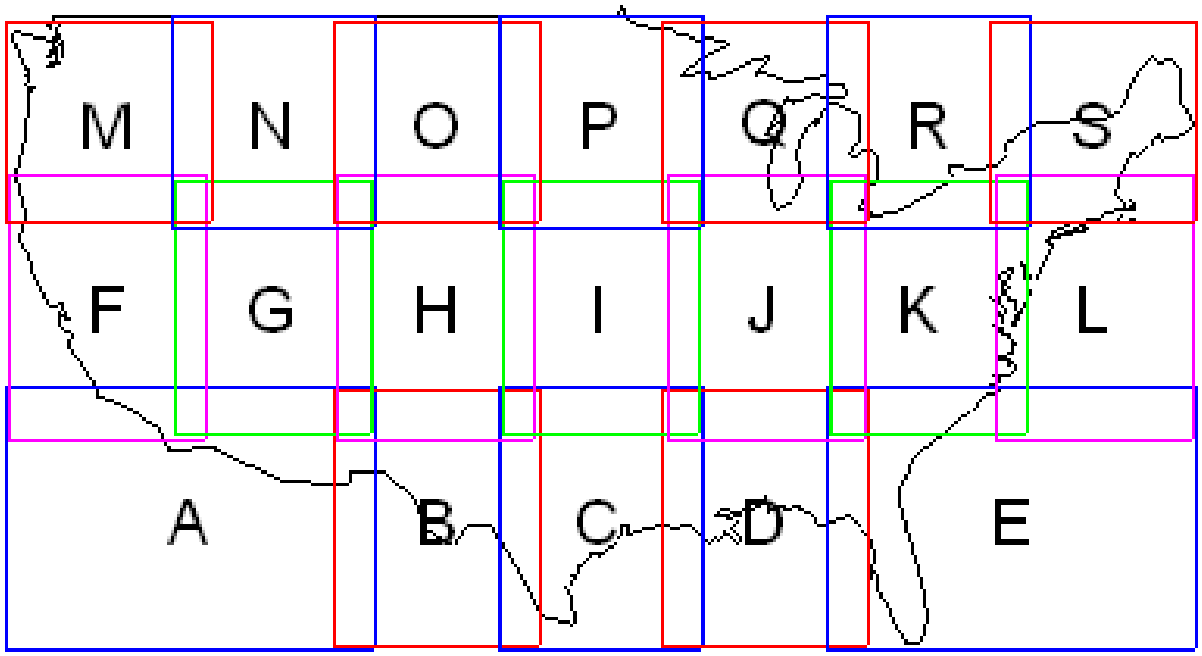
Here $x_+ = \max(x, 0)$, δ_t is a time increment (here 1 day based on a time unit of 1 year) and the parameters μ, ψ, ξ represent the location, scale and shape of the distribution. In particular, when $\xi > 0$ the marginal distributions have a Pareto (power-law) tail with power $-1/\xi$.

TEMPORAL AND SPATIAL DEPENDENCE

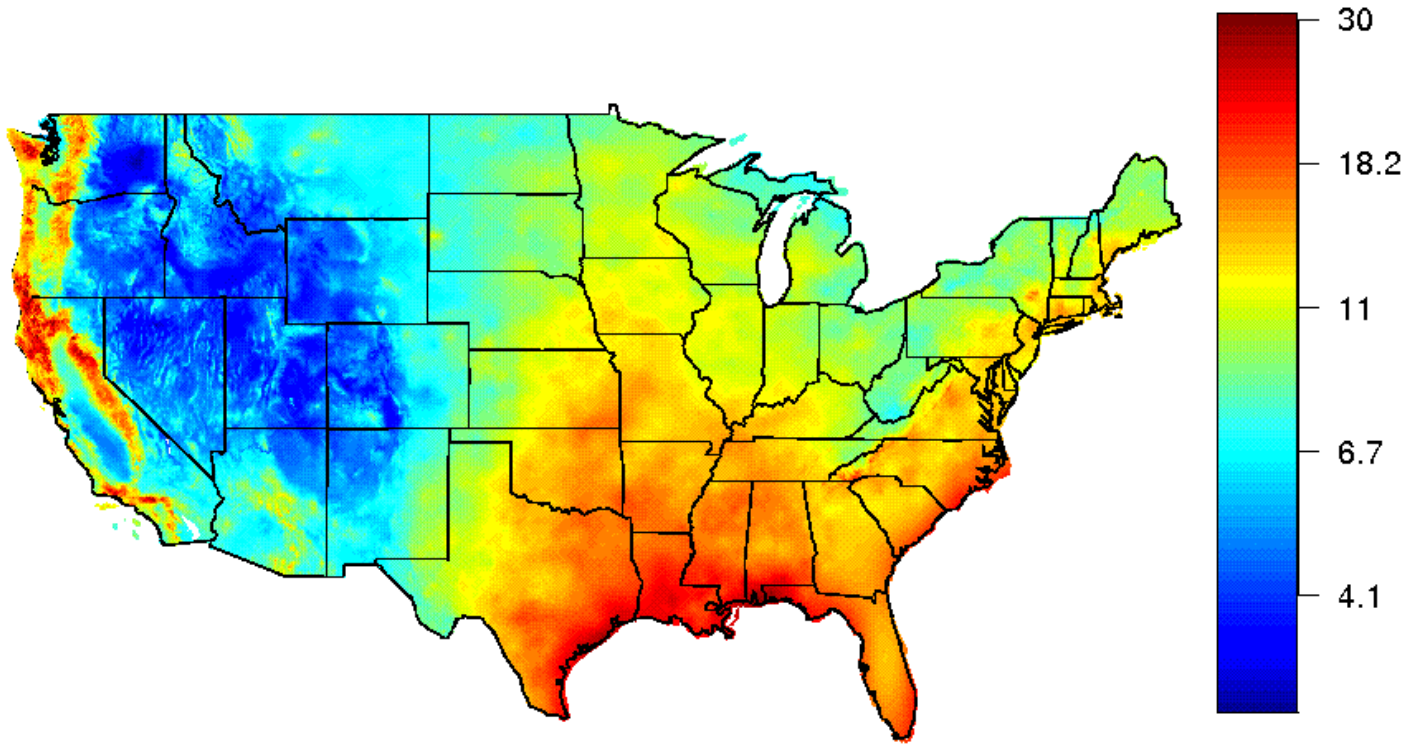
Here, we make two extensions of the basic methodology.

First, the parameters μ, ψ, ξ are allowed to be time-dependent through covariates. This allows a very flexible approach to seasonality, and we can also introduce linear trend terms to examine changes in the extreme value distribution over the time period of the study.

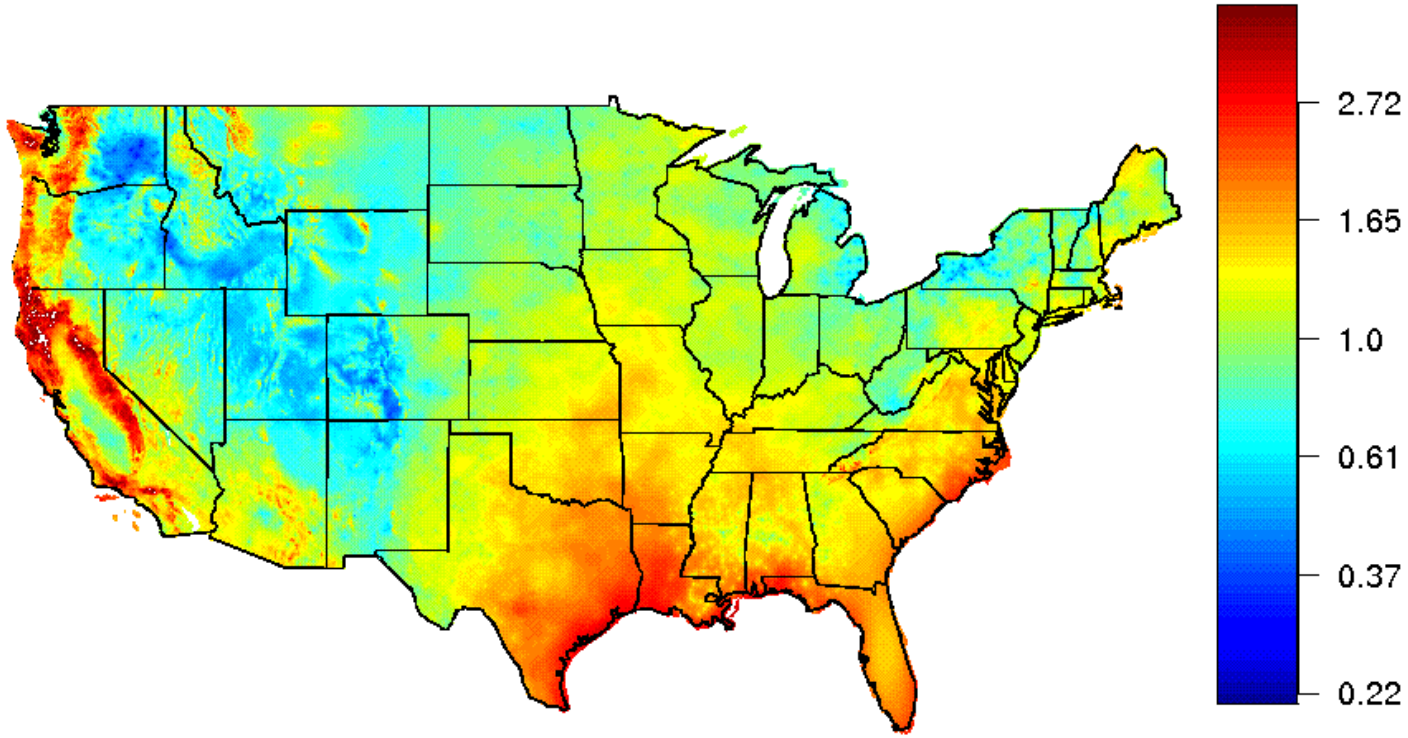
The second extension is *spatial smoothing*: after estimating the 25-year return value at each site, we smooth the results across sites by a technique similar to kriging. We allow for spatial nonstationarity by dividing the US into 19 overlapping boxes, and interpolating across the boundaries.



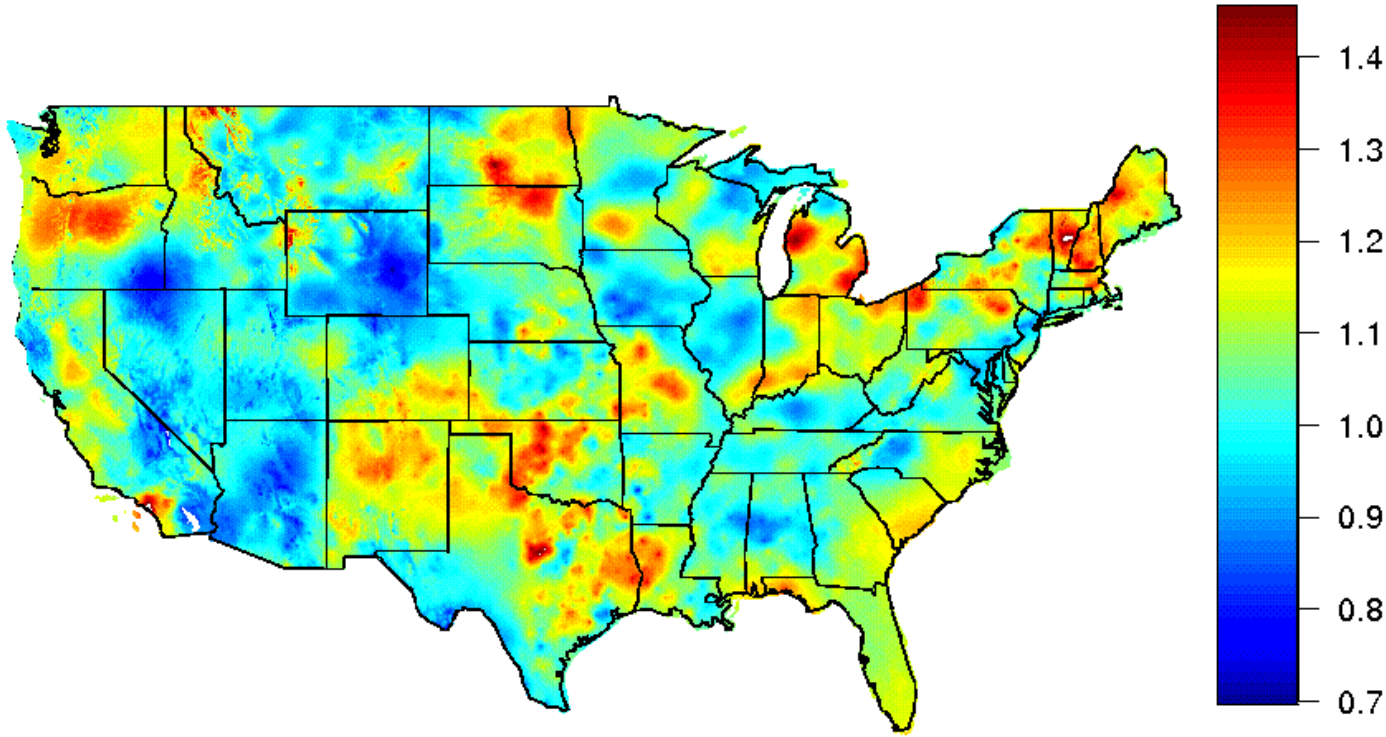
Continental USA divided into 19 regions



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, using a statistical model that assumes a linear trend in the GEV model parameters

	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			

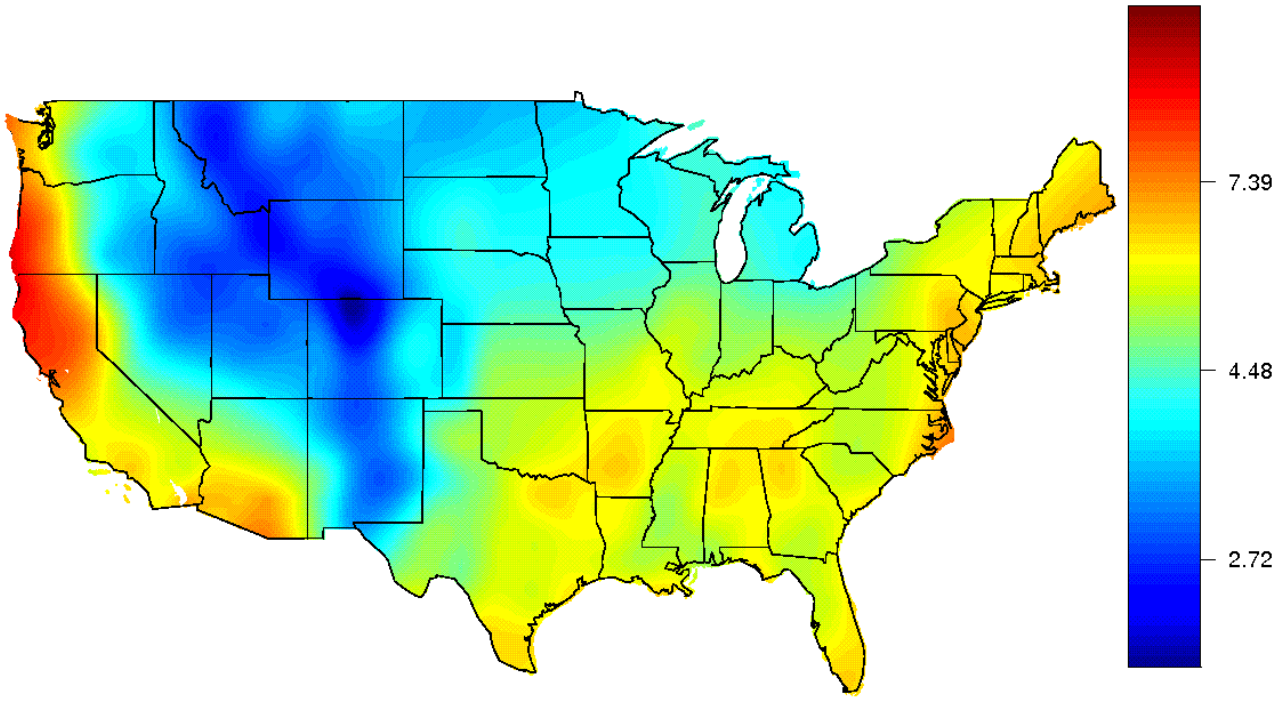
For each grid box, we show the mean change in log 25-year return value (1970 to 1999) and the corresponding standard error (RMSPE)

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

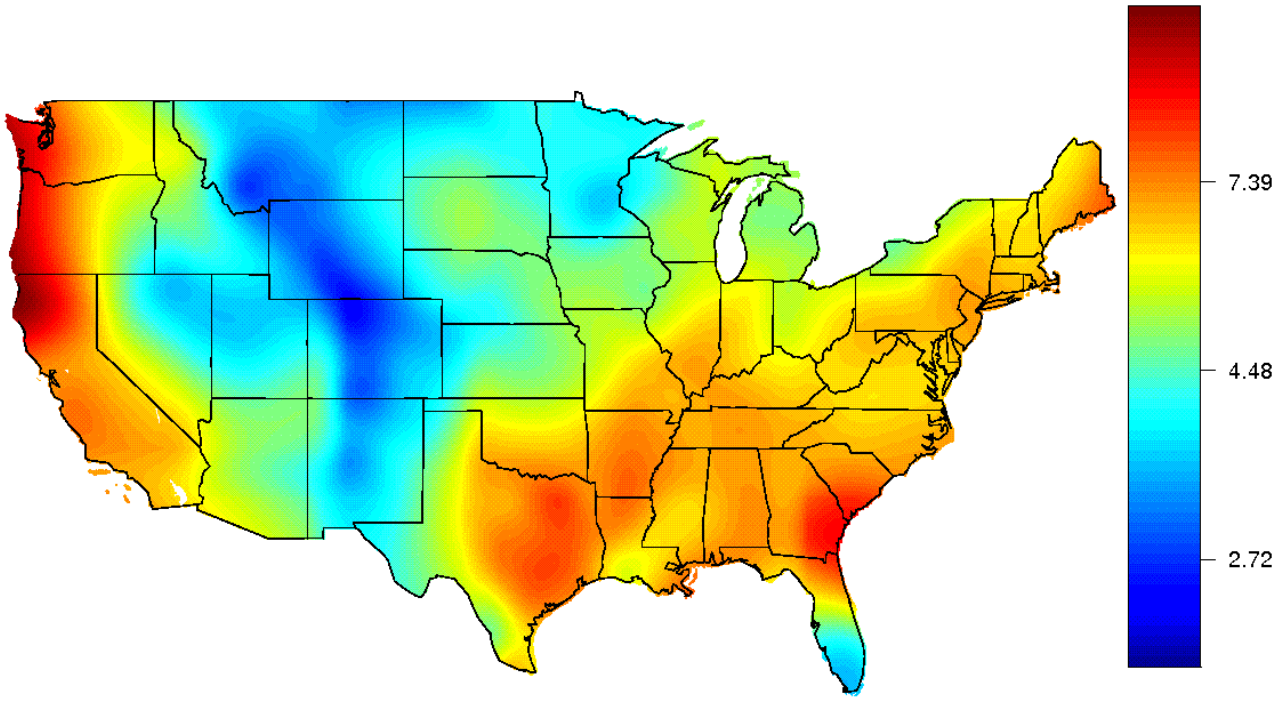
14 of 19 regions are statistically significant increasing: the remaining five are all in western states

We can use the same statistical methods to project future changes by using data from climate models.

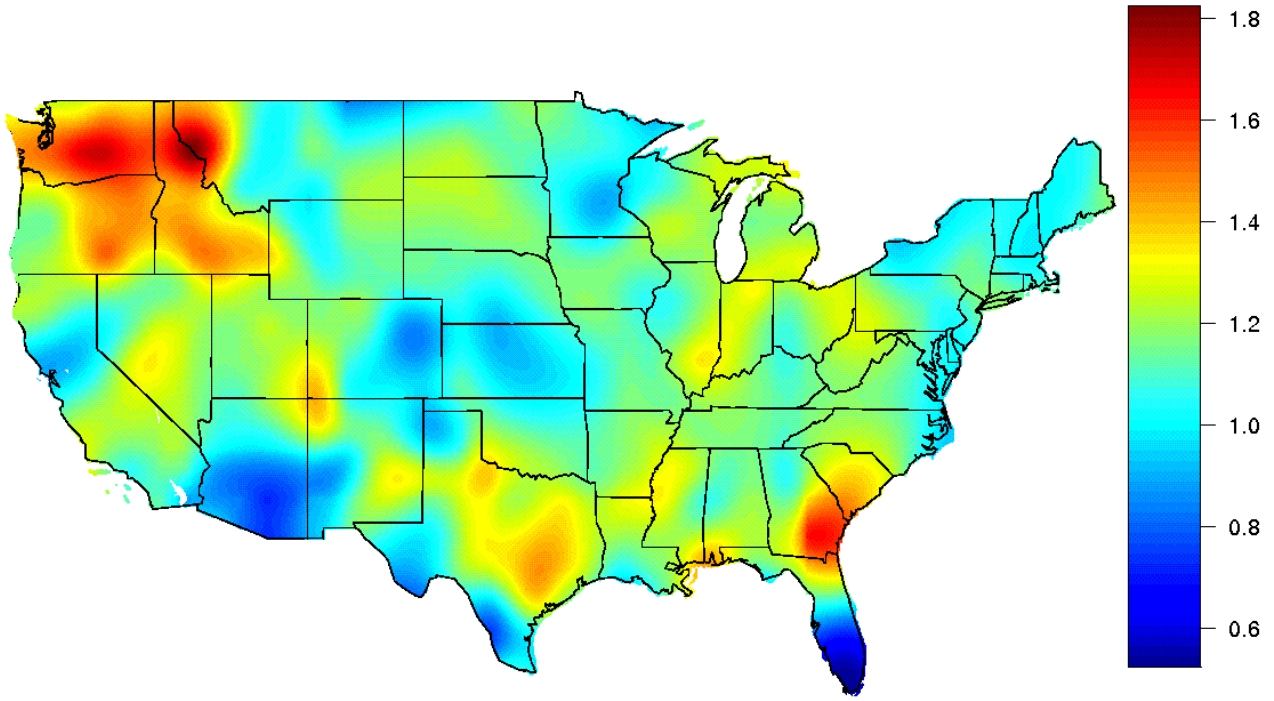
Here we use data from CCSM, the climate model run at NCAR.



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



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



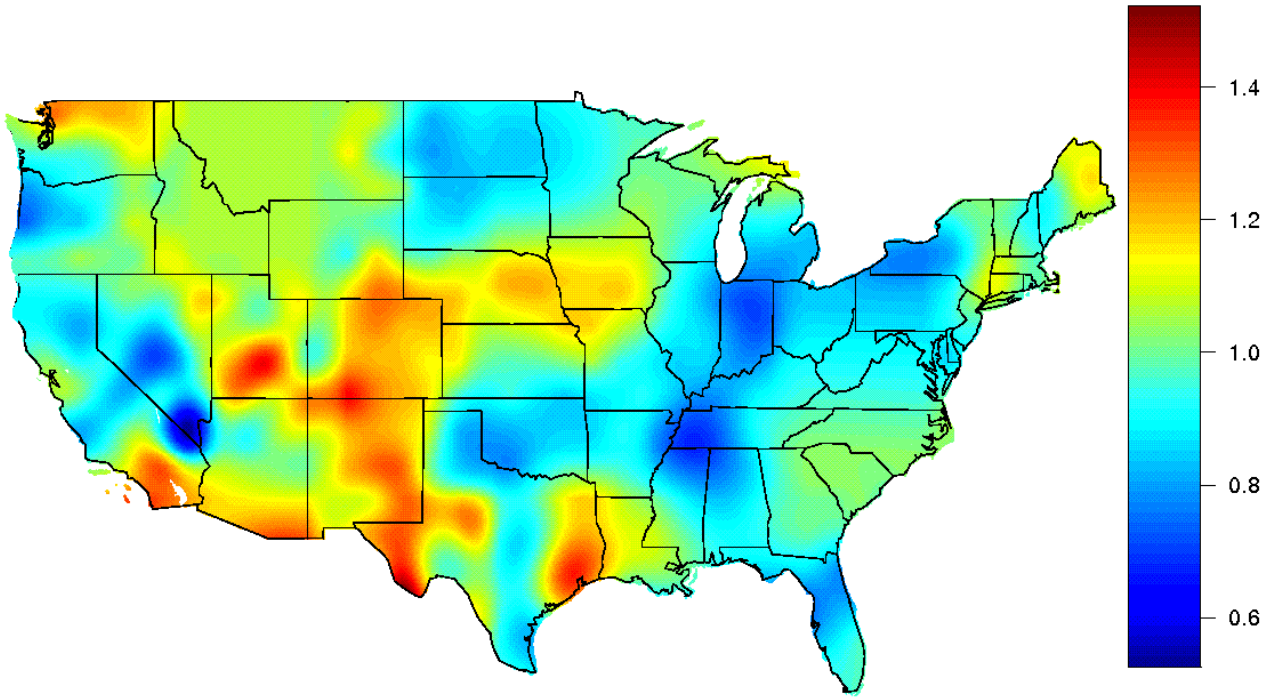
Estimated ratios of 25-year return values for 2070–2099 to those of 1970–1999, based on CCSM data, A1B scenario

The climate model data show clear evidence of an increase in 25-year return values over the next 100 years, as much as doubling in some places.

A caveat...

Although the overall increase in observed precipitation extremes is similar to that stated by other authors, the spatial pattern is completely different. There are various possible explanations, including different methods of spatial aggregation and different treatments of seasonal effects.

Even when the *same* methods are applied to CCSM data over 1970–1999, the results are different.



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

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

5. *But...* based on CCSM data there is a completely different spatial pattern and no overall increase — still leaves some doubt as to overall interpretation.

FURTHER READING

Finkenstadt, B. and Rootzén, H. (editors) (2003), *Extreme Values in Finance, Telecommunications and the Environment.* Chapman and Hall/CRC Press, London.
(See http://www.stat.unc.edu/postscript/rs/semstatrls.pdf)

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

THANK YOU FOR YOUR ATTENTION!