Analysing trends in the magnitude & frequency of extreme events

Seminar at **Heriot-Watt University**, May 2006

Adam Butler, Biomathematics & Statistics Scotland

Joint work with:

Janet Heffernan & Jonathan Tawn, Lancaster University

Roger Flather & Kevin Horsburgh, Proudman Oceanographic Laboratory

Storm surge pushing water
ashore during hurricane
Historic NWS Collection   Credit: U.S. Navy

Motivating example: storm surges

- **Sea level** = Mean sea level + Tide + Surge + Waves
- **Surges** are currents generated by winds & air pressure; They distort sea levels, especially in shallow waters, & are often a major contributory factor in coastal flooding
- **Increasing trend** in global mean sea levels is established, but there may be different trends in extreme sea levels
- e.g. **Changing patterns of storminess** could lead to changes in magnitude & frequency of storm surge events

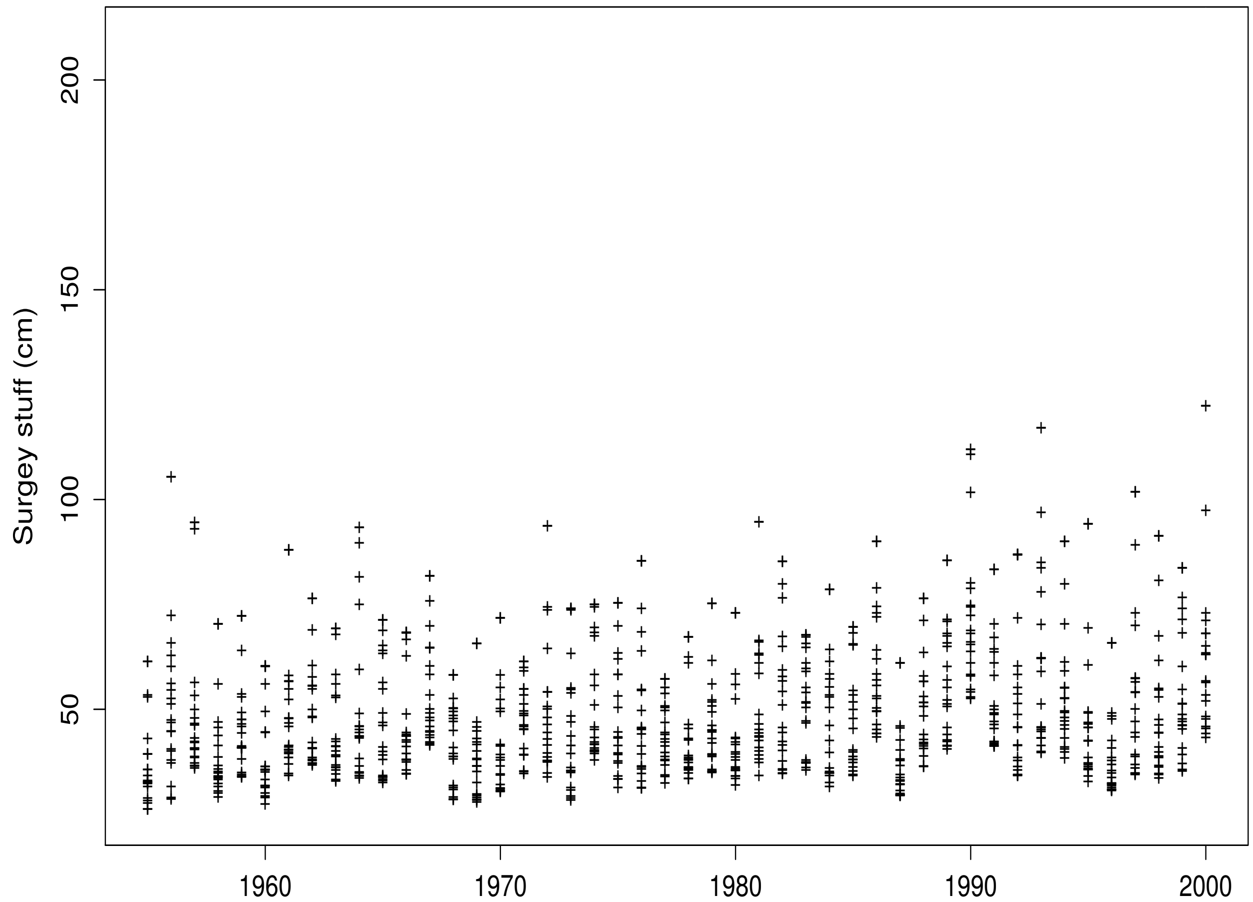
North Sea data

- **North Sea** coasts are at high risk from coastal flooding
- e.g. **Catastrophic floods** of 1362, 1634, 1953
- We **analyse surge elevations** for the period 1955-2000
- **Hourly data** come from:

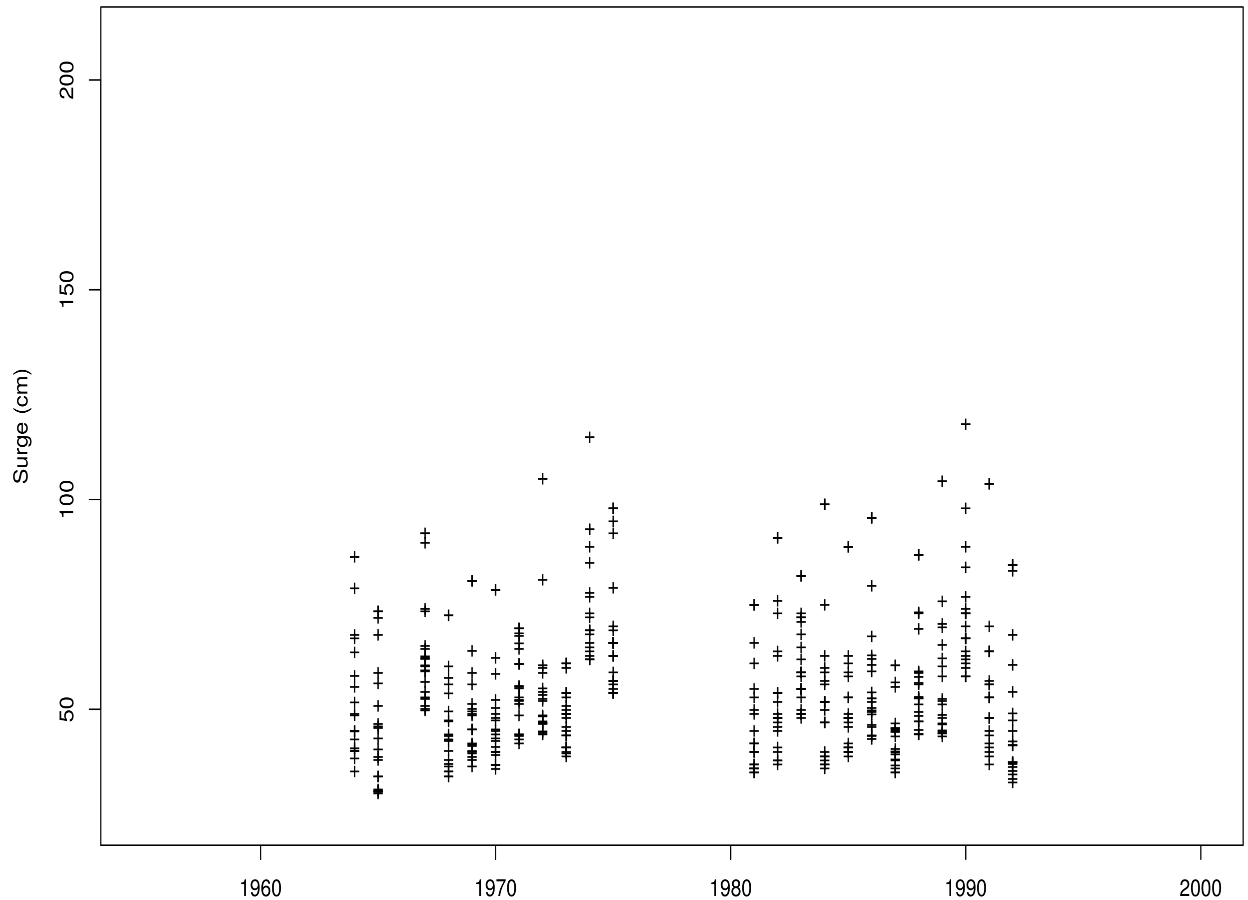
  - tide gauge records for coastal ports (sparse)

  - hindcast output from a numerical storm surge model
- **Synthetic data** refer to 35×35km grid cells

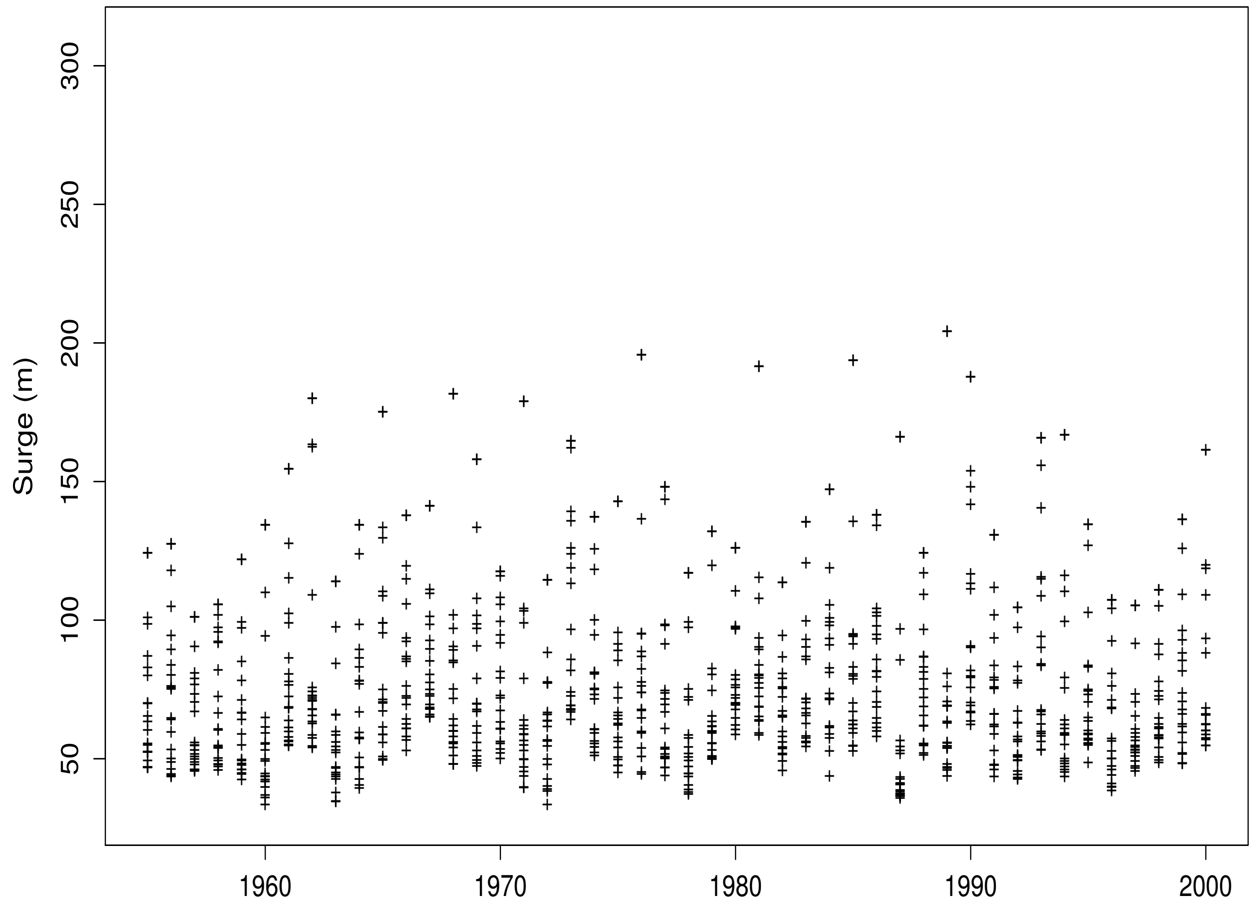
Synthetic data for Aberdeen: 20 largest storm surges per year



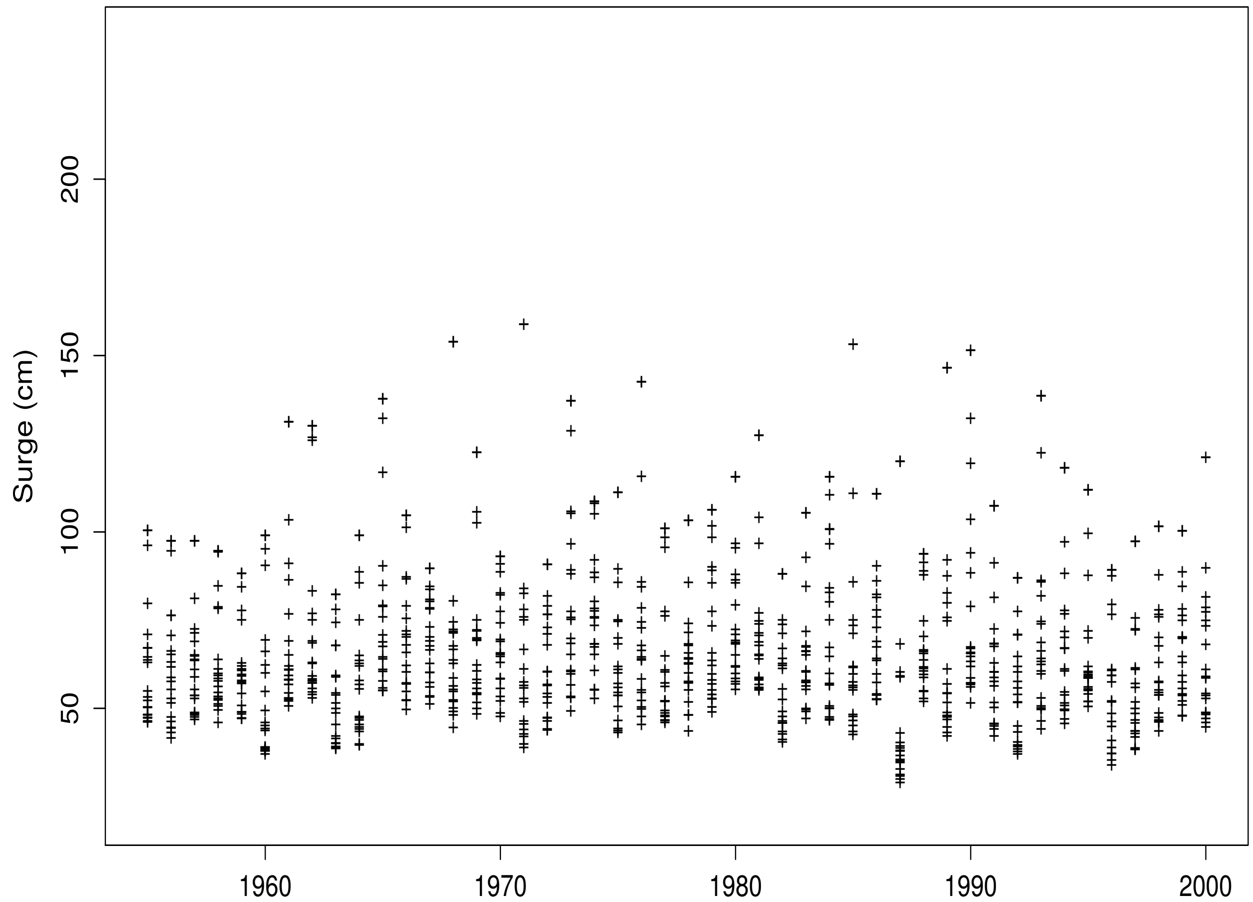
Observational data for Aberdeen: 20 largest storm surges per year



Synthetic data for Lowestoft: 20 largest storm surges per year



Synthetic data for Dover: 20 largest storm surges per year



Rest of the talk

1 : Extreme value theory

2 : Trends in extremes

3 : Dealing with dependence

1 - Extreme value theory

- **Typical extreme value problems**:

- Exceedance probabilities, $\mathbb{P}(X > u)$
  e.g. what is the probability that a sea wall of height u is overtopped during one year ?

- N-year return levels, $q_N = \{u : \mathbb{P}(X > u) = 1/N\}$
  e.g. what height should we build a sea wall such that it is overtopped with probability $1/N$ in a particular year ?

- **Common features**:

  We are interested in an event in the tail of the distribution

  There is little or no data available on such events -

  we are often extrapolating

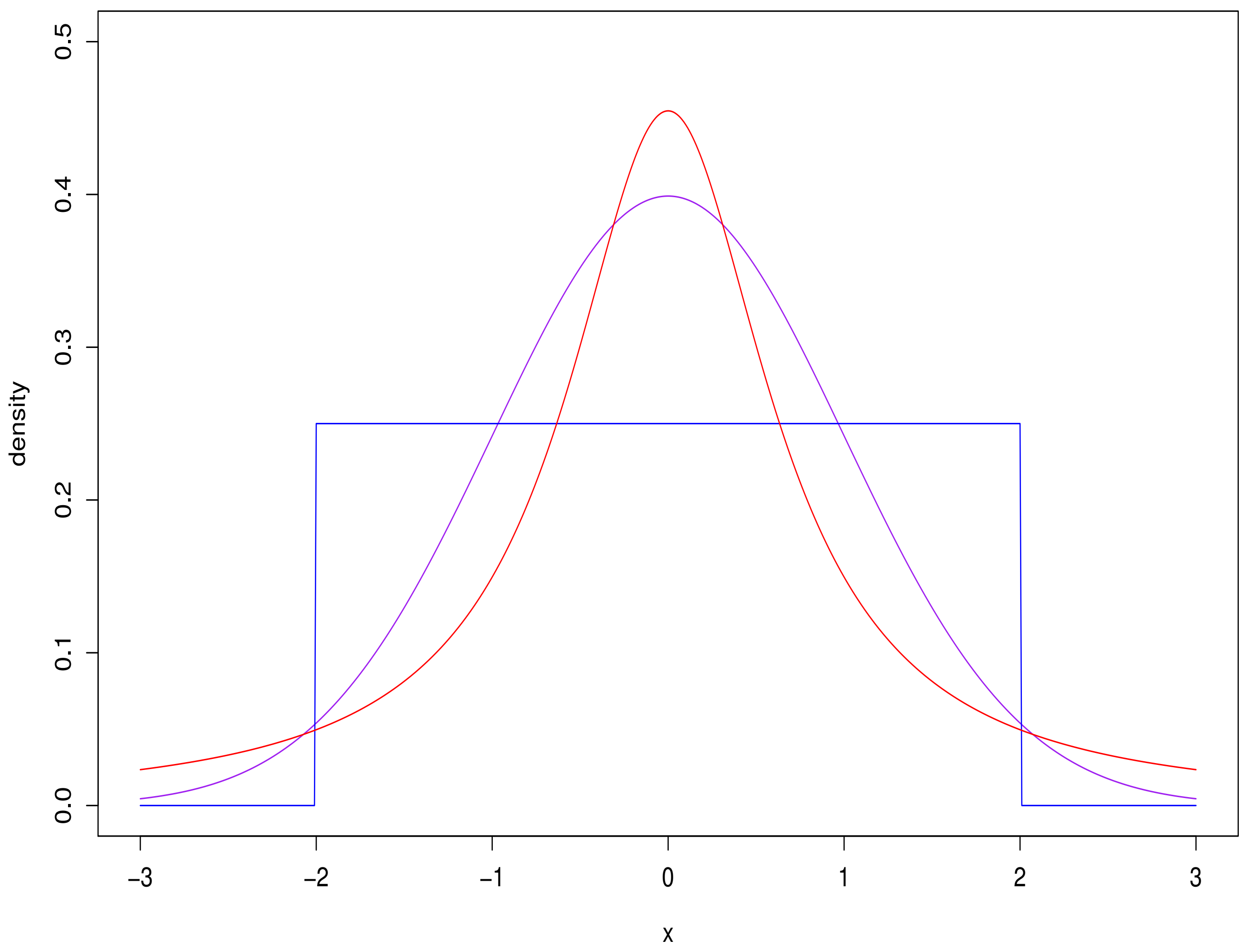
- **A worrying example**:

  The trouble with standard models is that inferences about the tails are *highly* sensitive to model mis-specification

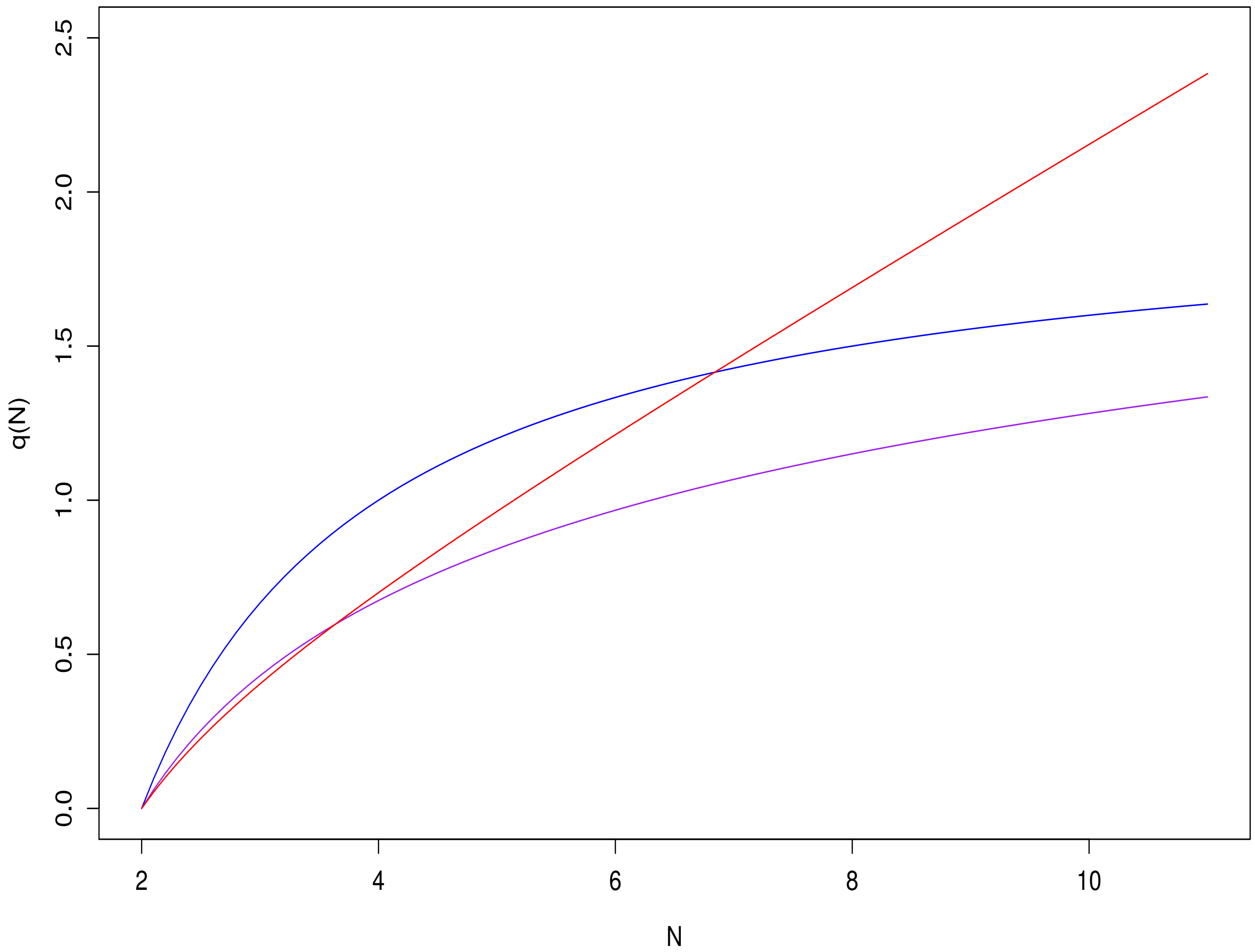
  e.g. consider the following distributions:

  Normal(0,1), Cauchy(0,0.7) and Uniform(-2,2)...

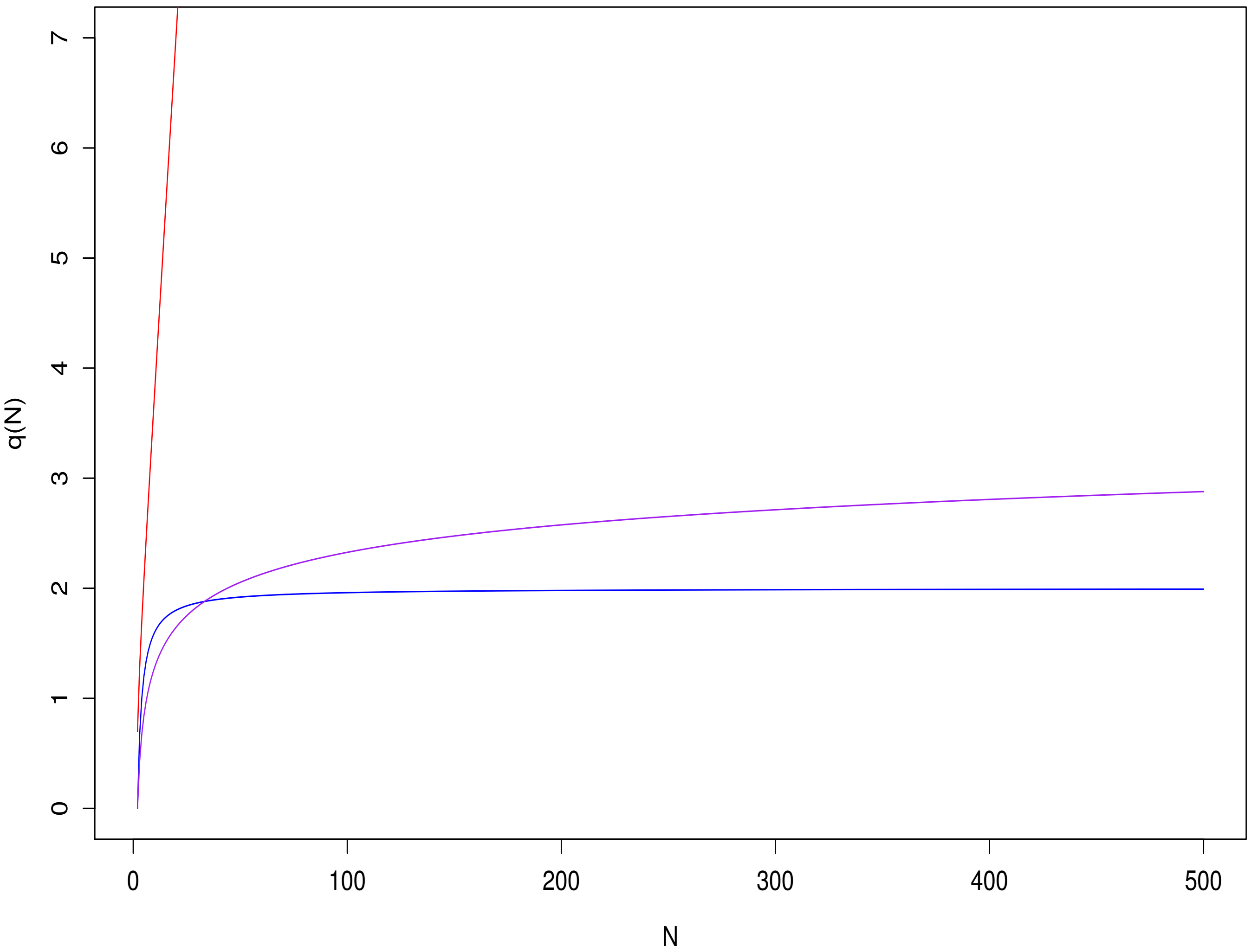
Density functions: Cauchy (red), Normal (purple), Uniform (blue)



Return levels: Cauchy (red), Normal (purple), Uniform (blue)



Return levels, again: Cauchy (red), Normal (purple), Uniform (blue)



- **The extreme value approach**:

  Base inferences about extreme events

  only on data at relatively extreme levels

  Models for extreme data motivated by asymptotic theory

  Provides a relatively robust basis for extrapolation

- **Key assumptions**:

  We are looking at a sufficiently extreme level

  Extreme events are drawn from a common population

  The data are free from outliers

- **Threshold exceedance approach**:

  Describes exceedances of a sufficiently high threshold u

  As $u \longrightarrow \infty$ the conditional distribution $(X - u)|X > u$ tends to a generalised Pareto distribution,

$$F(y) = 1 - \left(1 + \frac{\xi y}{\phi}\right)^{-1/\xi},$$

  with scale parameter ϕ and shape parameter ξ

- **Block maxima approach approach**:

  Describes the maxima of sufficiently long blocks

  As $n \longrightarrow \infty$ the normalised maxima

$$[\max(X_1, \ldots, X_n) - b_n]/a_n$$

  tend to a GEV distribution

$$F(z) = \exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}^{-1/\xi}\right],$$

  with location, scale and shape parameters μ, σ & ξ.

- **Parameters**:

  GEV & GPD parameters linked by $\phi = \sigma + \xi(u - \mu)$

  The shape parameter ξ is identical in both models

  $\xi = 0$: (Gumbel case) normal tails

  $\xi > 0$: heavy tails

  $\xi < 0$: light tails

  N-year return level is related to GEV parameters via

$$q(N) = \mu - \frac{\sigma}{\xi}\left[1 - \left\{-\log\frac{N-1}{N}\right\}^{-\xi}\right]$$

- **r-largest value approach**:

  Describes the r largest events within a block (e.g. year)

  Generalisation of the GEV model,

  with the same parameters μ, σ and ξ

  Makes more efficient use of the available data

  Asymptotic motivation relies upon r being sufficiently small

- **Poisson process framework**:

  As $n \longrightarrow \infty$ the point process

$$\left\{ \frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right\}$$

  converges on intervals of the form $[0, 1] \times [u, \infty)$ to a Poisson process with intensity determined by μ, σ and ξ

- **Statistical inference & modelling**:

  Assume asymptotic models are valid at some finite level

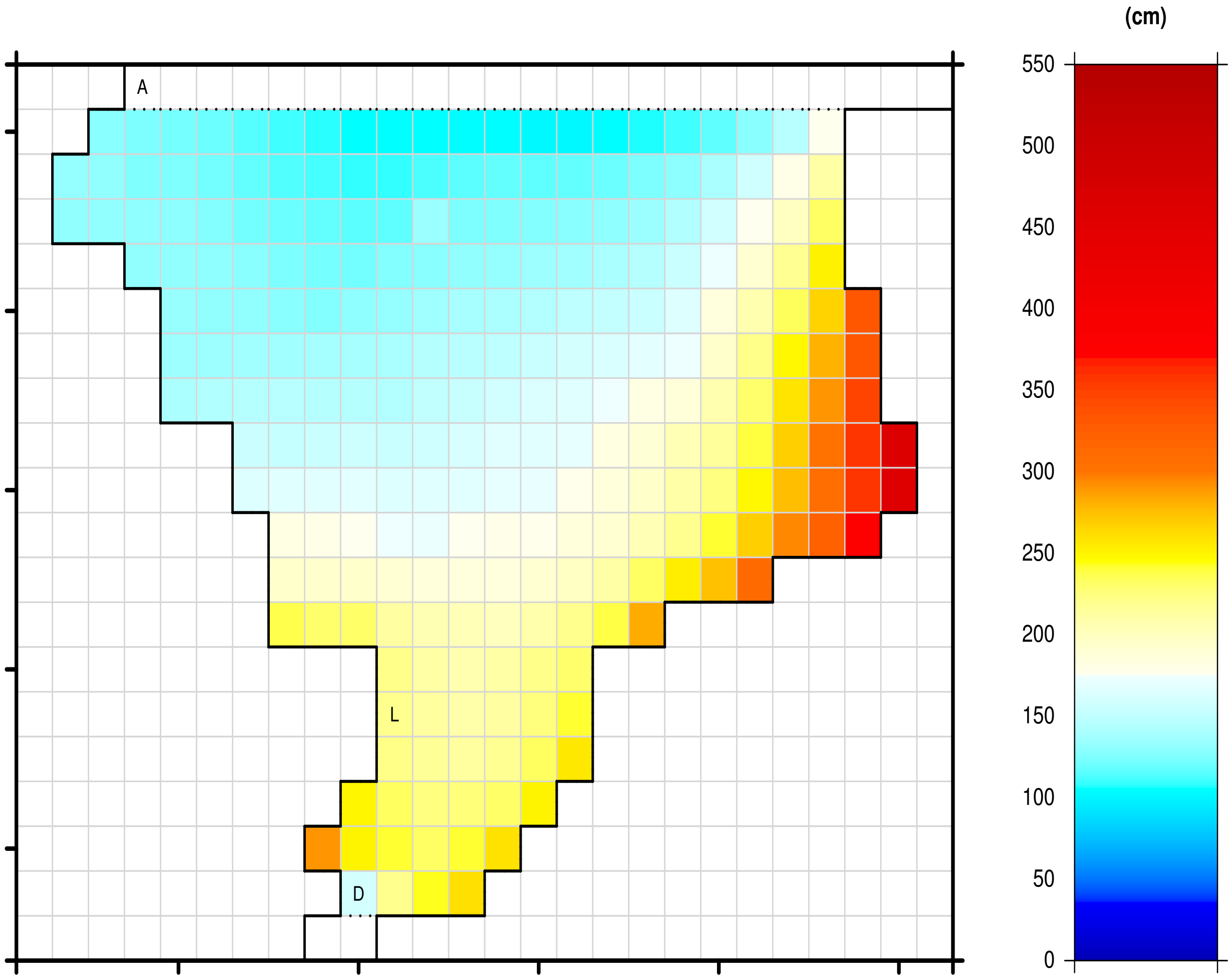
  Selection of an appropriate level (u, n or r) is *difficult*

  but parameter stability plots are a useful tool

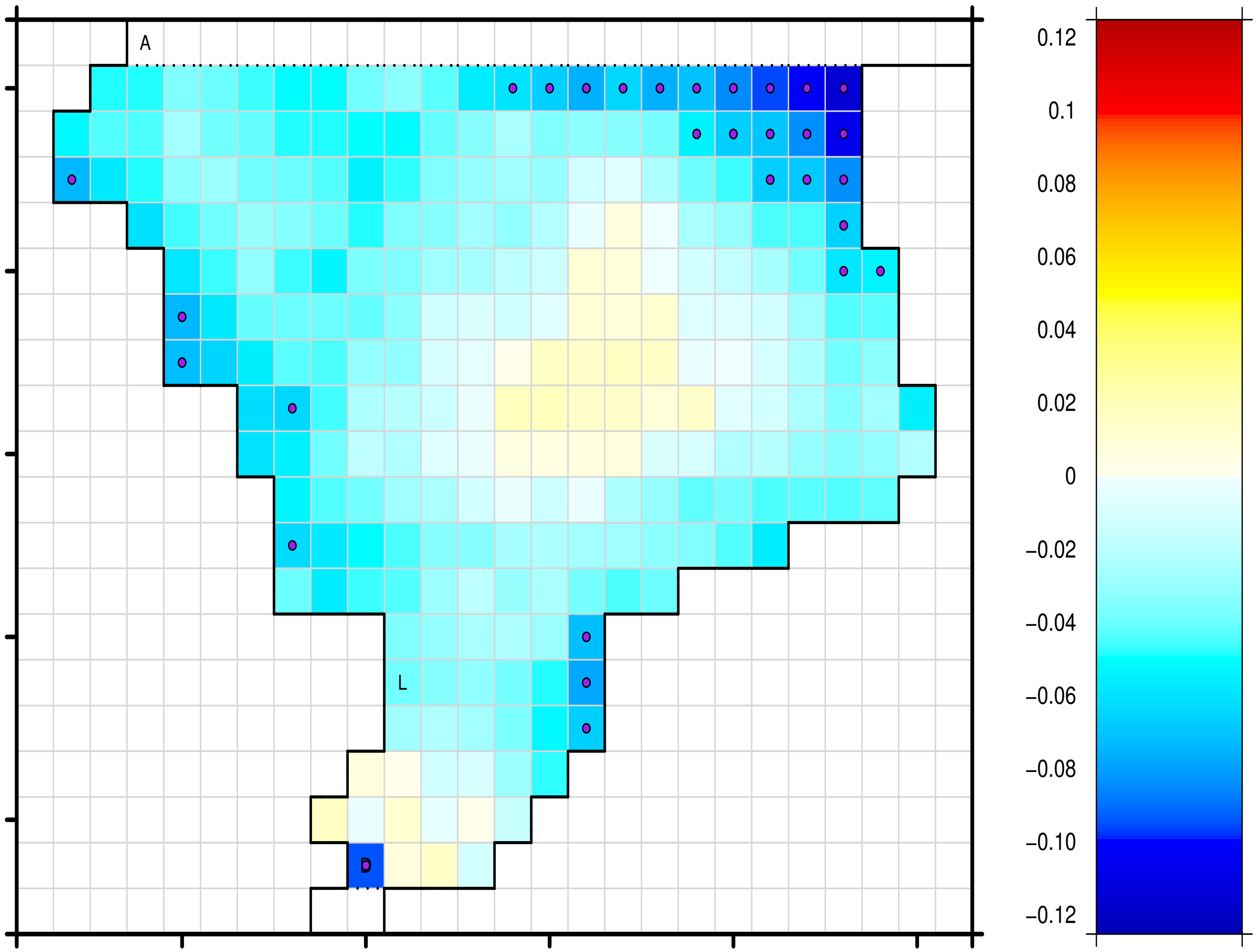
  Estimate model parameters via e.g. maximum likelihood

  Thereby estimate return levels & exceedance probabilities

Estimated 50y return levels, based on an r-largest fit at each site



Shape parameter estimates, based on an r-largest fit at each site



- **Introductory book**:

  Coles, S.G. (2001) An Introduction to Statistical Modelling of Extreme Values, Springer.

- **Software** (http://cran.r-project.org/):

  R packages `evd`, `evdbayes`, `evir`, `ismev`

- **Webpage**:

  www.cru.uea.ac.uk/cru/projects/mice/html/extremes.html

2 - Trends in extremes

- Z_j is the maximum of process X in year t_j
- Assume that Z_j has a GEV distribution
  with parameters $\theta(t_j) = (\mu(t_j), \sigma(t_j), \xi(t_j))$
- Model these parameters as functions of time t
- Identical approach for r-largest model
  & for point process model of threshold exceedances

Some recent applications

Annual minimum temperatures at Arosa weather station, in the Swiss Alps, for 1954-1997 (Chavez-Demoulin, 1999)

Acres burnt by large wildfires in the USA since 1825 (Ramesh, 2001)

'Extreme datamining' of terrabyte-sized logfiles for internet applications (Chavez-Demoulin *et al.*, 2003)

Parametric trends

Model components of $\theta(t)$ as parametric functions of time

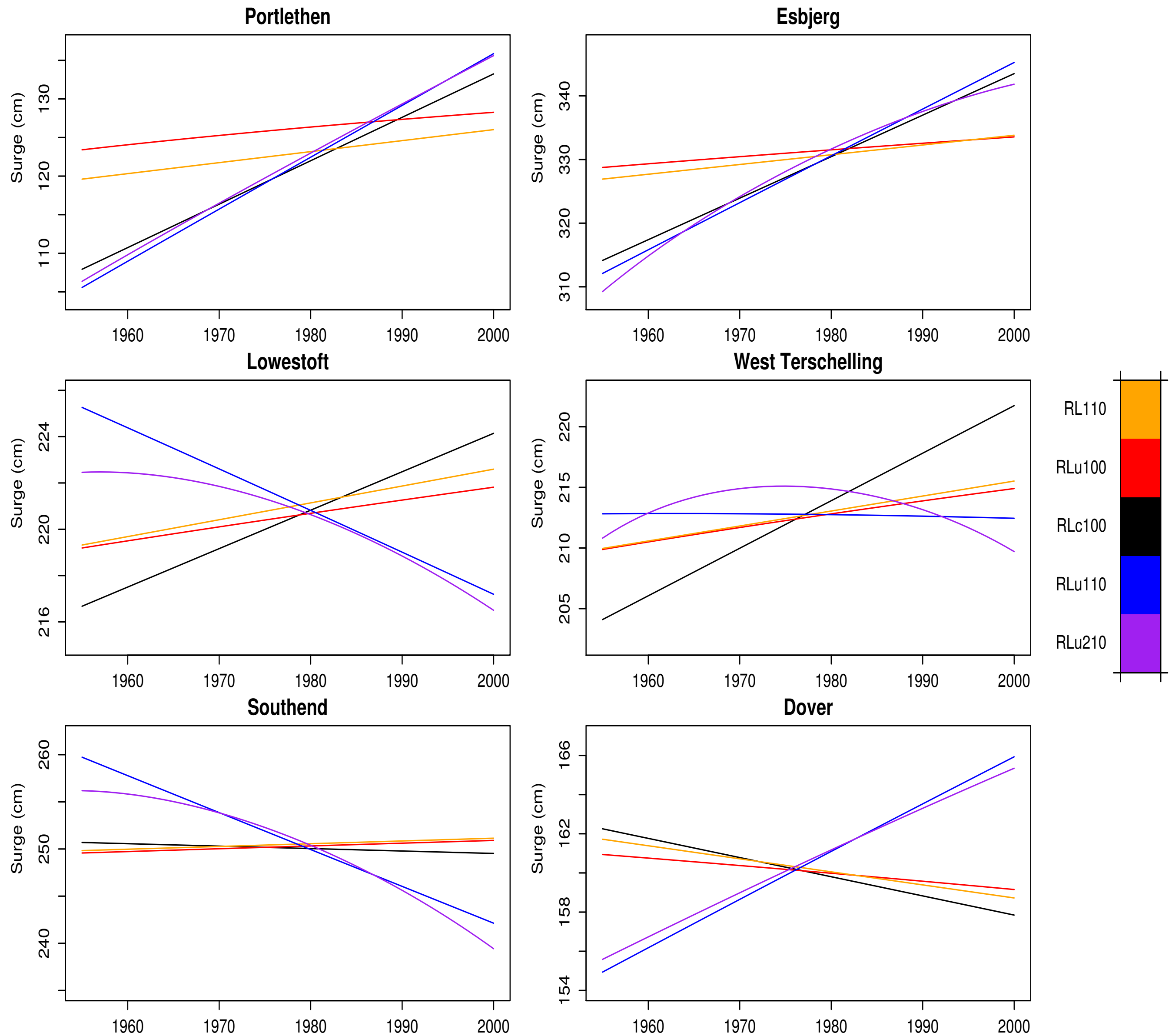
e.g. polynomial $\theta(t) = (\mu_0 + t\mu_1, \sigma^*(\mu_0 + t\mu_1), \xi)$

or changepoint $\theta(t) = (\mu_l \mathbb{I}[t \leq T] + \mu_u \mathbb{I}[t > T], \sigma, \xi)$

Estimate parameters by maximisation of the log-likelihood

$$\sum_j \log f\left(z_j; \theta(t_j)\right)$$

or by Bayesian methods



Portlethen
Esbjerg
Lowestoft
West Terschelling
Southend
Dover
Surge (cm)
1960
1970
1980
1990
2000
RL110
RLu100
RLc100
RLu110
RLu210

Local likelihood

Davison and Ramesh (2000); Hall and Tajvidi (2000)

Estimate $\theta(t_j)$ by numerically maximising the local likelihood

$$\sum_J w_{Jj} \log f\left(z_J; \theta(t_j)\right)$$

The weights w_{Jj} determine the degree of smoothing

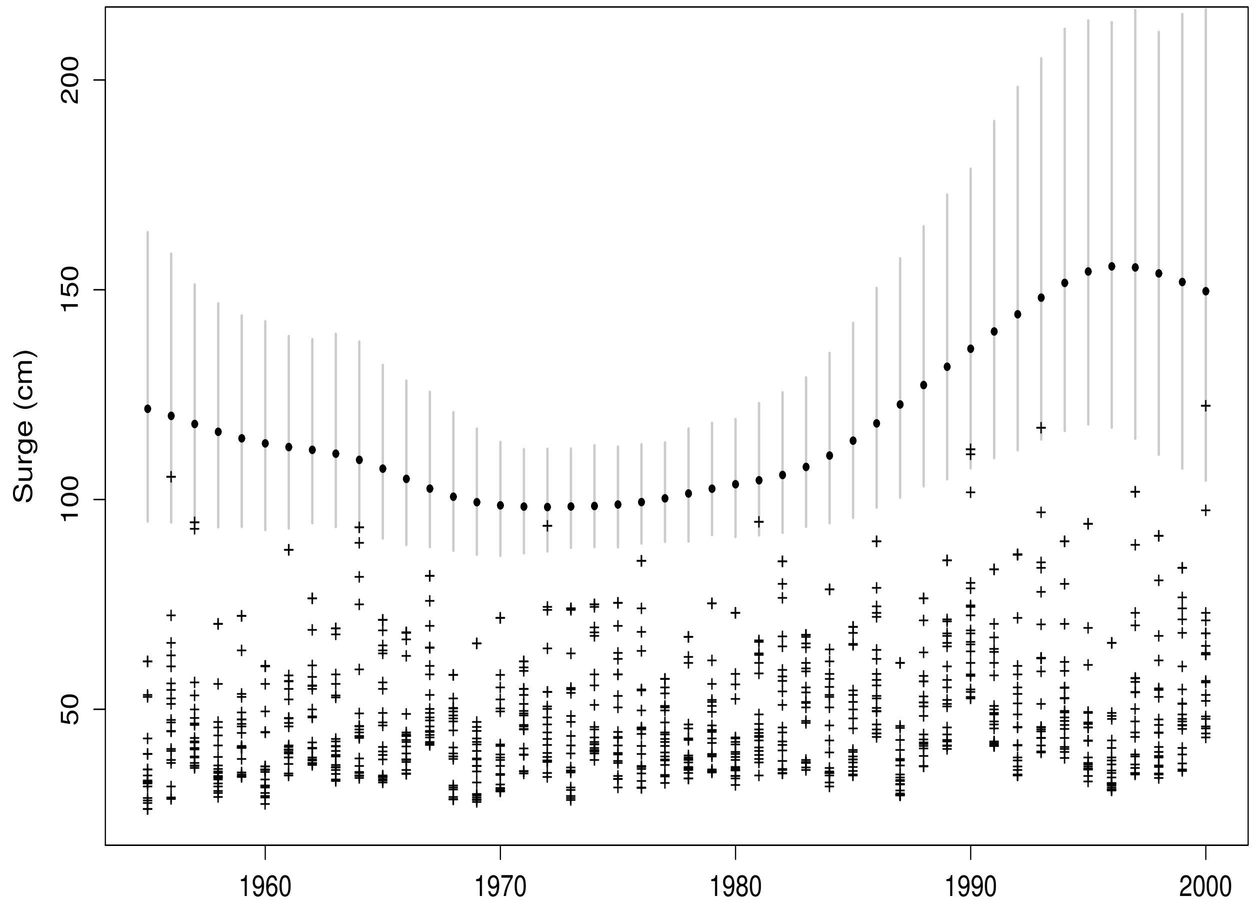
Choice of weight function - e.g. use a Gaussian kernel:

$$w_{Jj} = \phi\left(\frac{t_J - t_j}{h}\right)$$

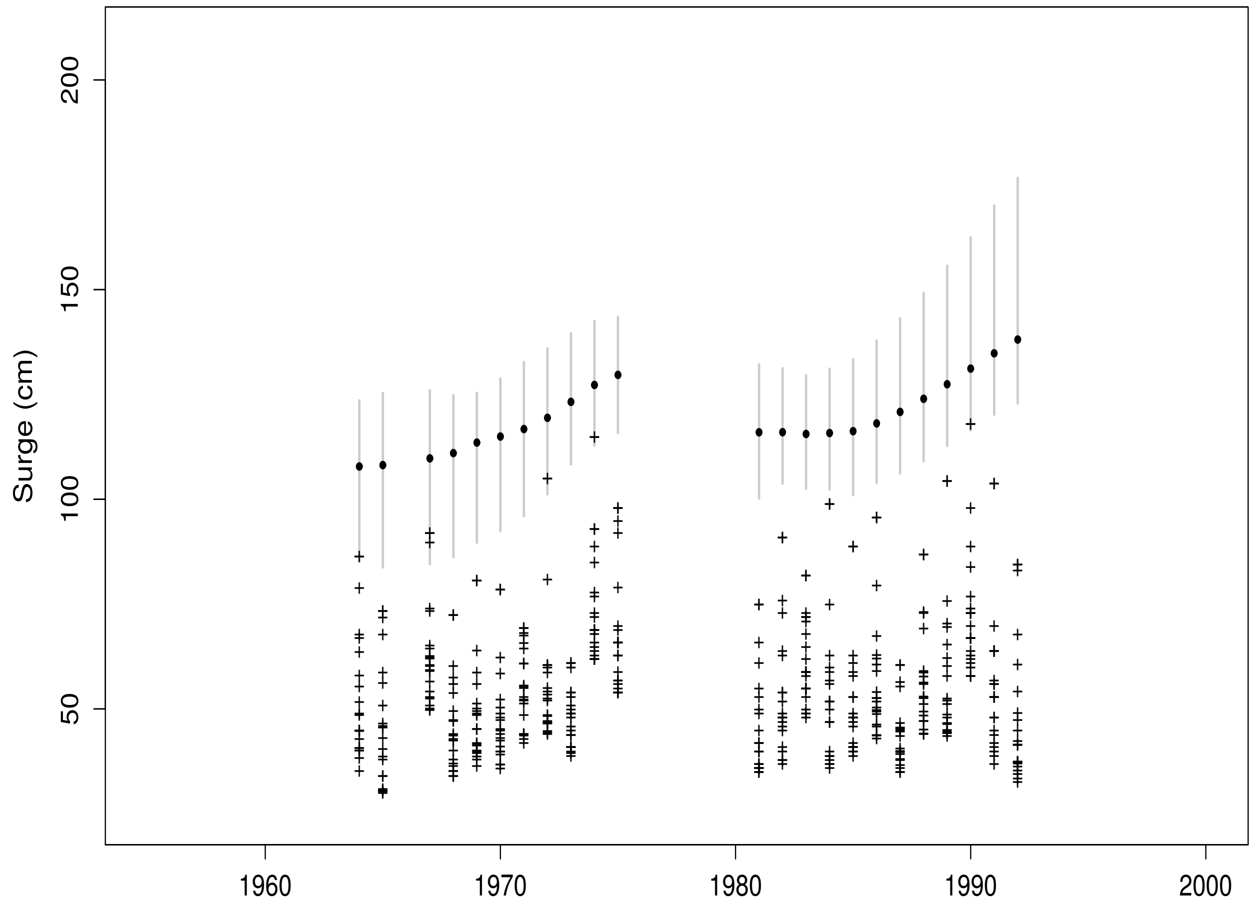
Select the bandwidth h either:

- subjectively, based on scientific knowledge; or
- using an automatic criterion e.g. cross-validation

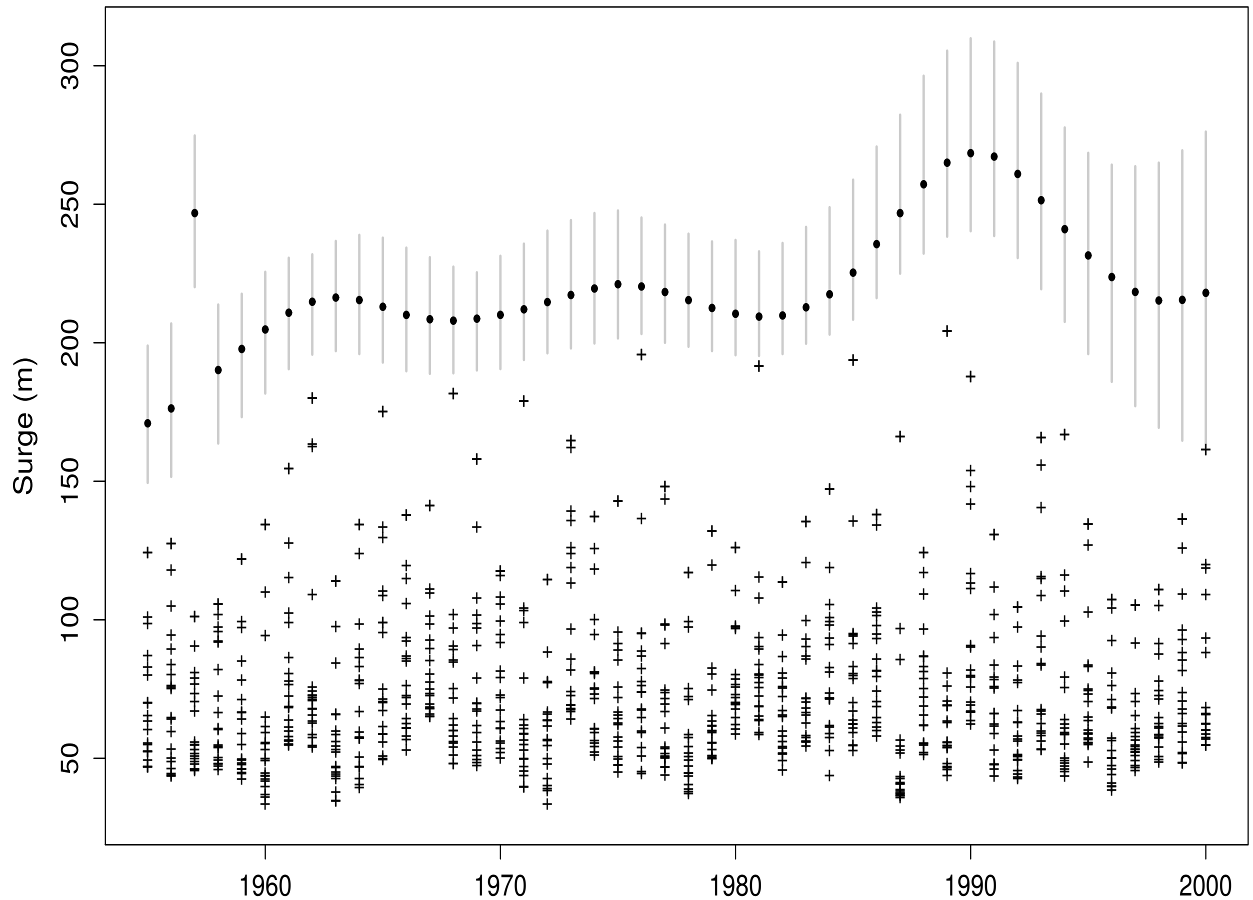
50y surge return levels: model output for Aberdeen



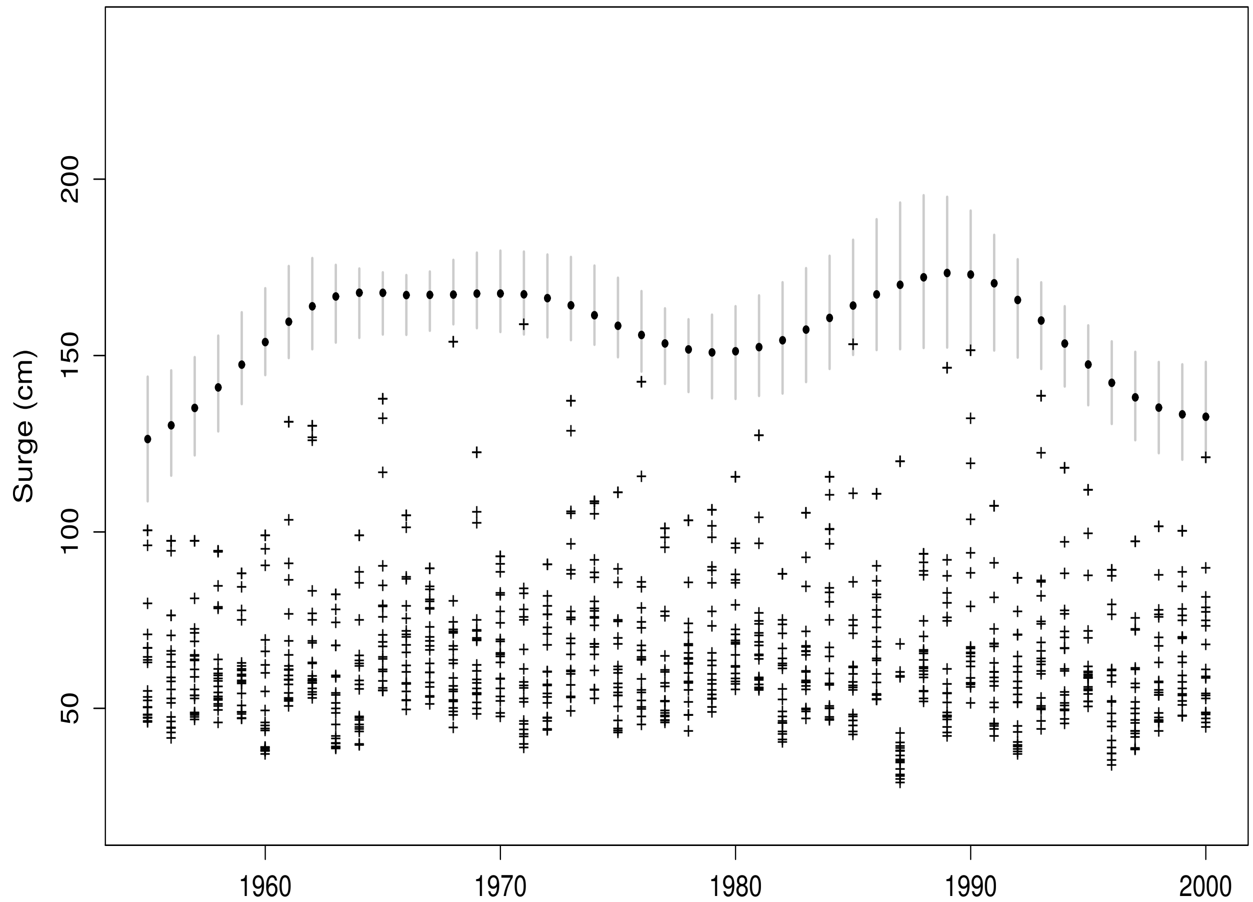
50y surge return levels: observational data for Aberdeen



50y surge return levels: model output for Lowestoft



50y surge return levels: model output for Dover



Local likelihood with constraints

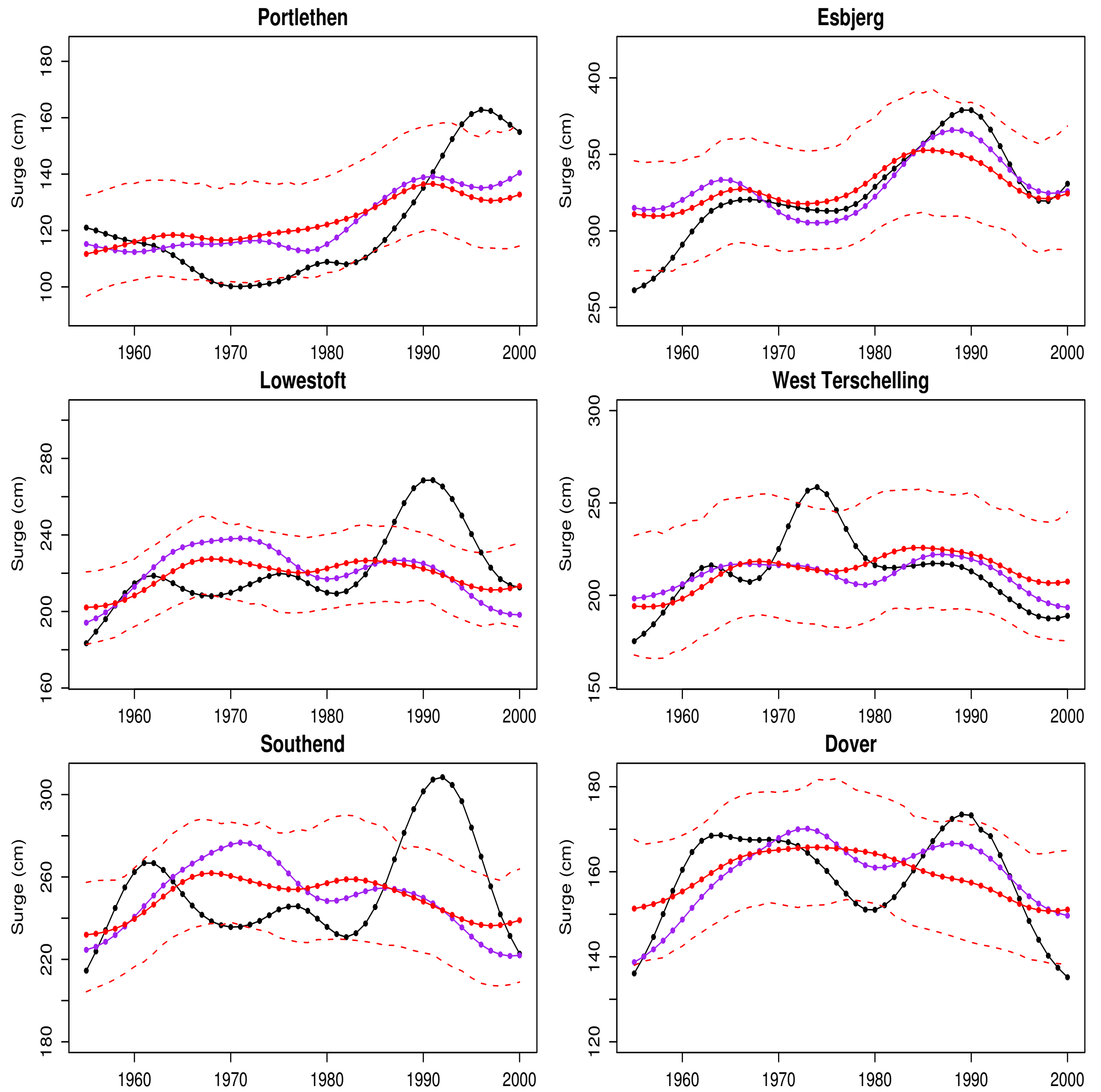
Partition the parameters as $\theta(t) = (\beta, \theta^*(t))$

Iterative procedure for inference, alternating maximisation of

$$\sum_J w_{Jj} \log f\left(z_J; (\hat{\beta}, \theta^*(t_j))\right)$$

for each j with maximisation of

$$\sum_j \log f\left(z_j; (\beta, \hat{\theta}^*(t_j))\right)$$


Portlethen
Esbjerg
Lowestoft
West Terschelling
Southend
Dover
Surge (cm)
1960
1970
1980
1990
2000

Penalised likelihood

Chavez-Demoulin (1999); Pauli and Coles (2001)

Estimate $\theta(t)$ by maximising the penalised likelihood

$$\sum_{j} \log f\left(z_j; \theta(t_j)\right) + \lambda \int_t \theta''(t)^2 dt$$

using a Fisher scoring algorithm

Known as spline smoothing as solutions are cubic splines

The penalty terms λ determine the degree of smoothing

Nonlinear dynamic models (Gaetan and Grigoletto, 2004)

Assume that $\theta(t_j) = M_j \alpha_j$ for a known matrix M_j

Assume that $\alpha_j | \alpha_{j-1} \sim N(T\alpha_{j-1}, Q)$ & $\alpha_0 \sim N(a, B)$

Matrices T and M_j control the complexity of the model

Use McMC to infer values of unobserved states $\alpha_1, \ldots, \alpha_N$

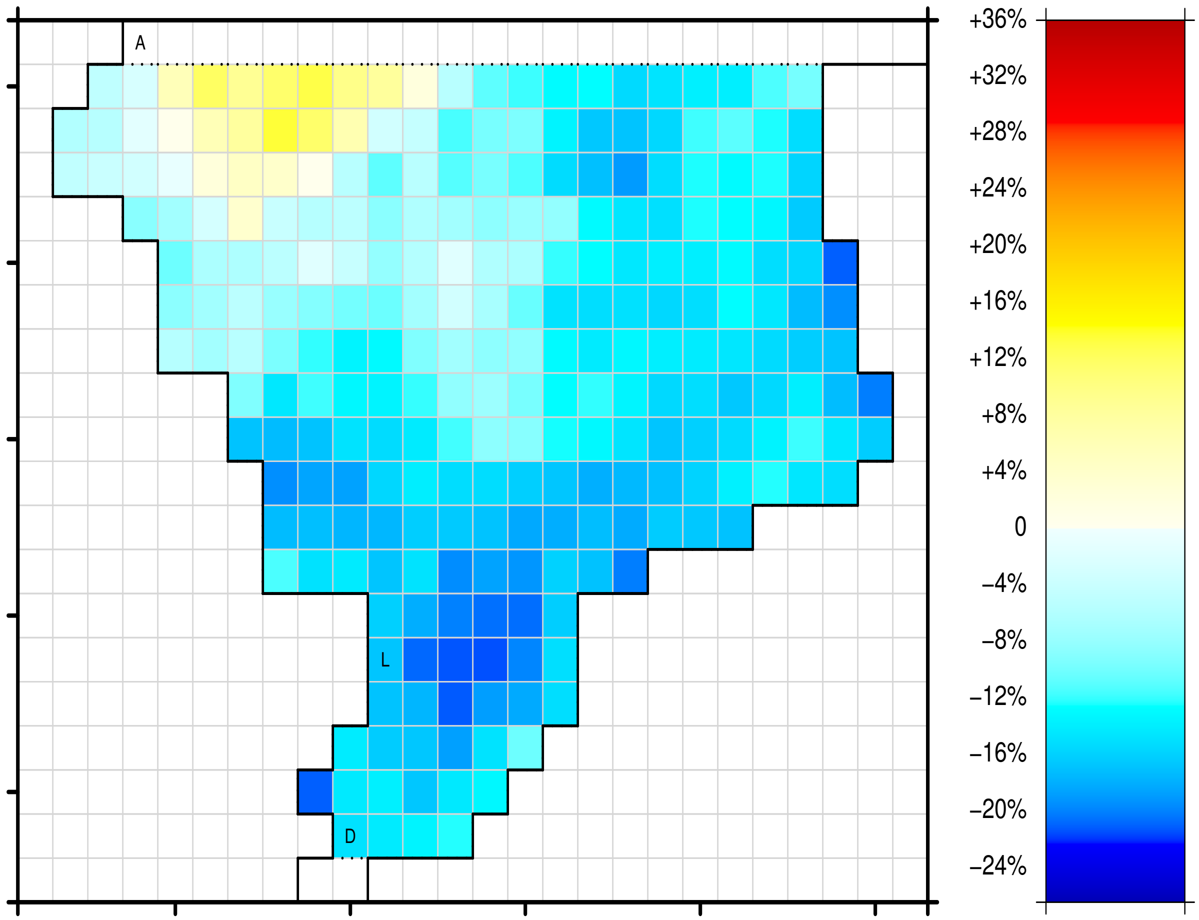
Hyperparameters a, B and Q control degree of smoothing:

select using posterior predictive assessment

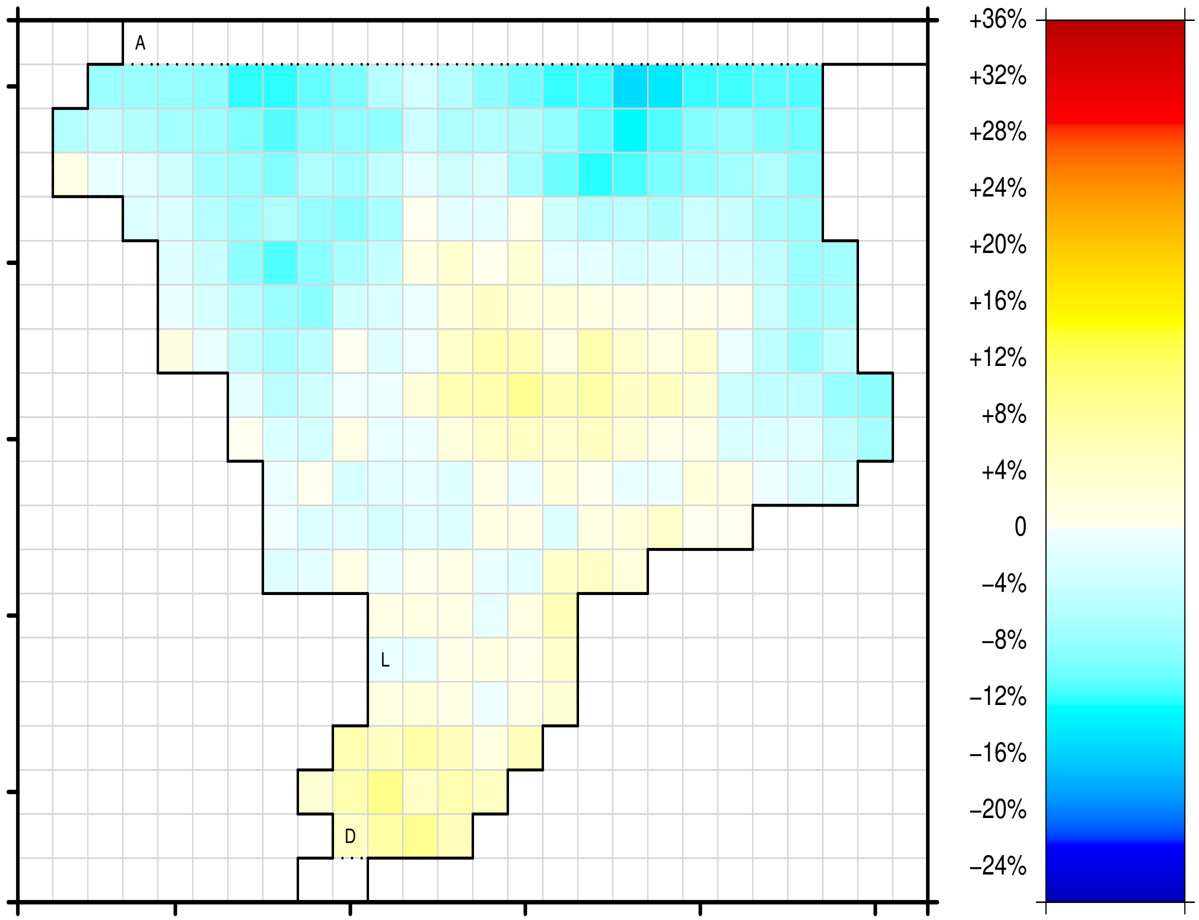
3: Dealing with dependence

- **Dependence within process** X leads to clustering & complicates the definition of an extreme event
- **Commonality** of parameter values over different variables
- **Dependence between variables** at extreme levels

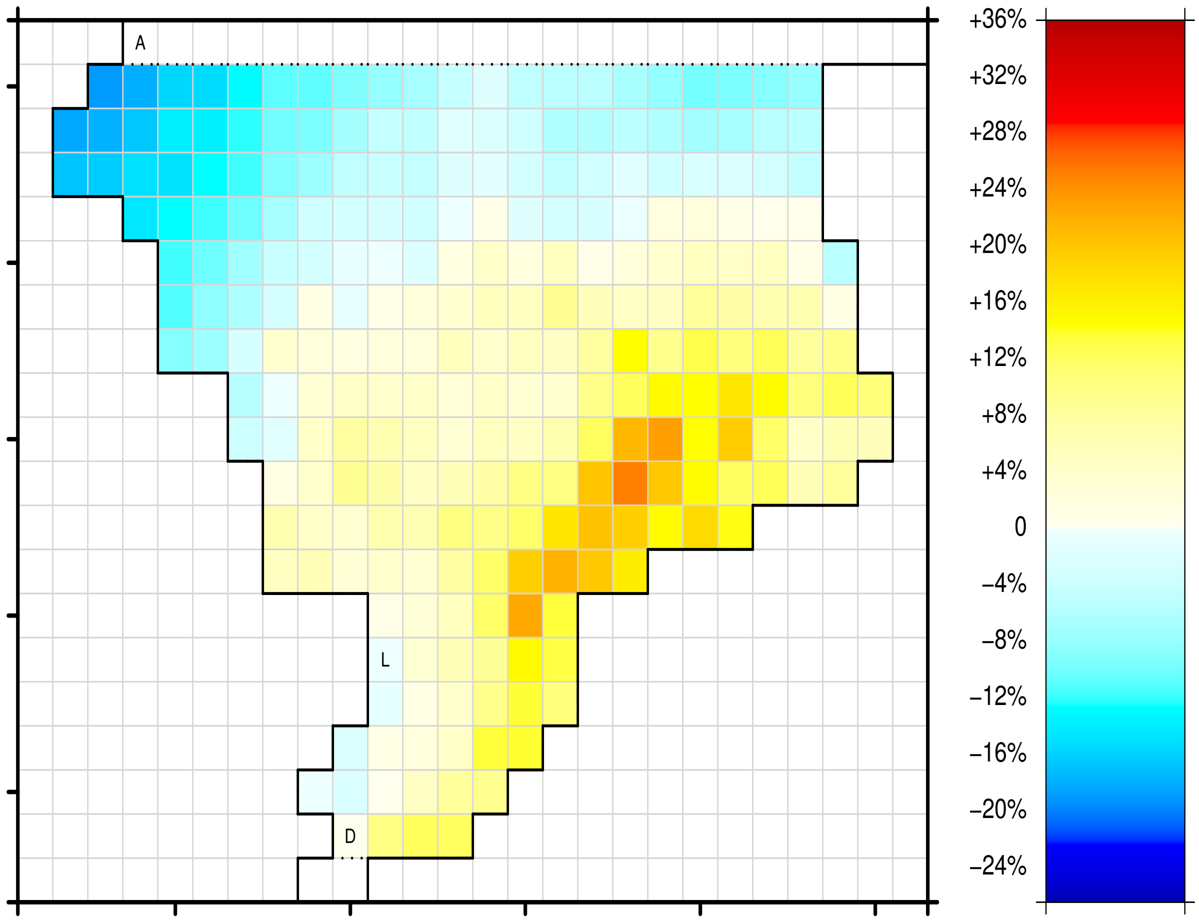
Deviations in 50y return levels from long-term means: 1955



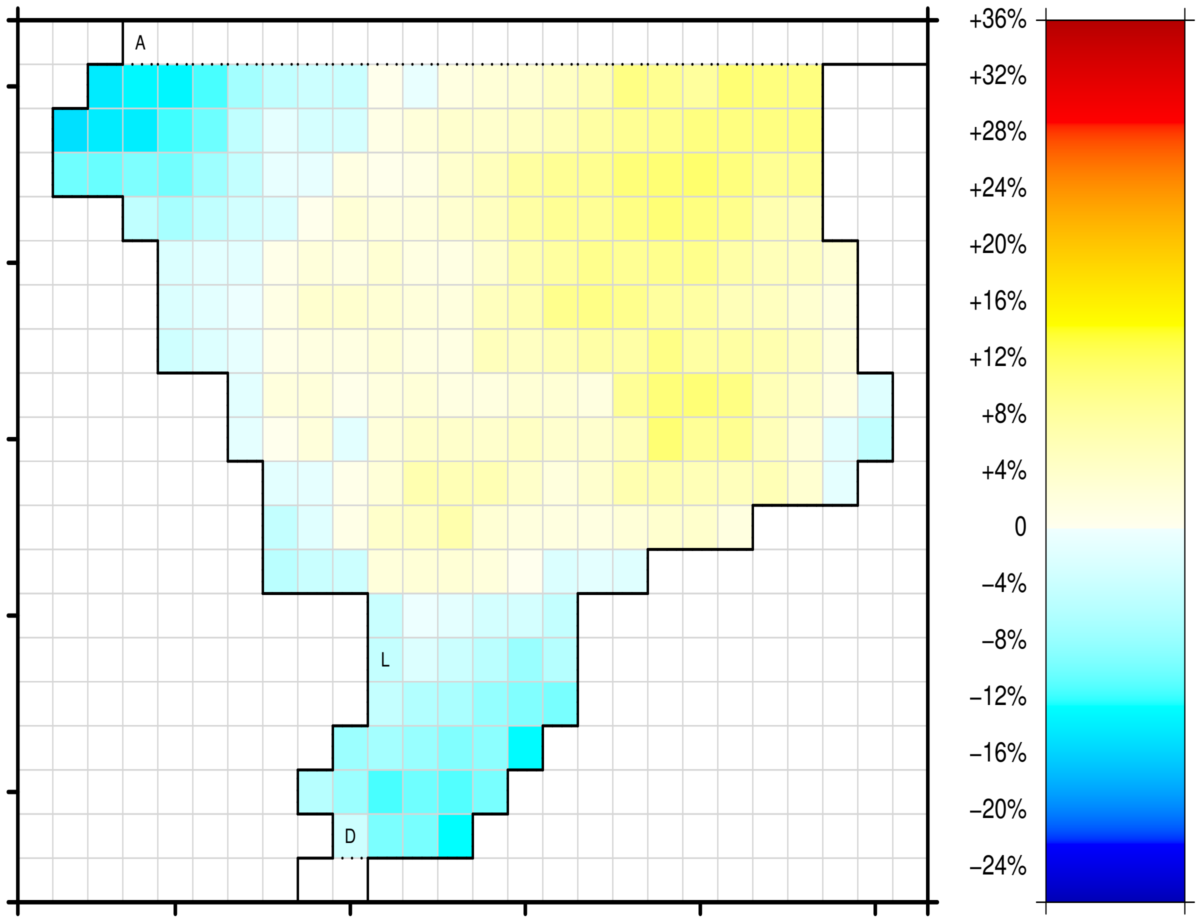
Deviations in 50y return levels from long-term means: 1962



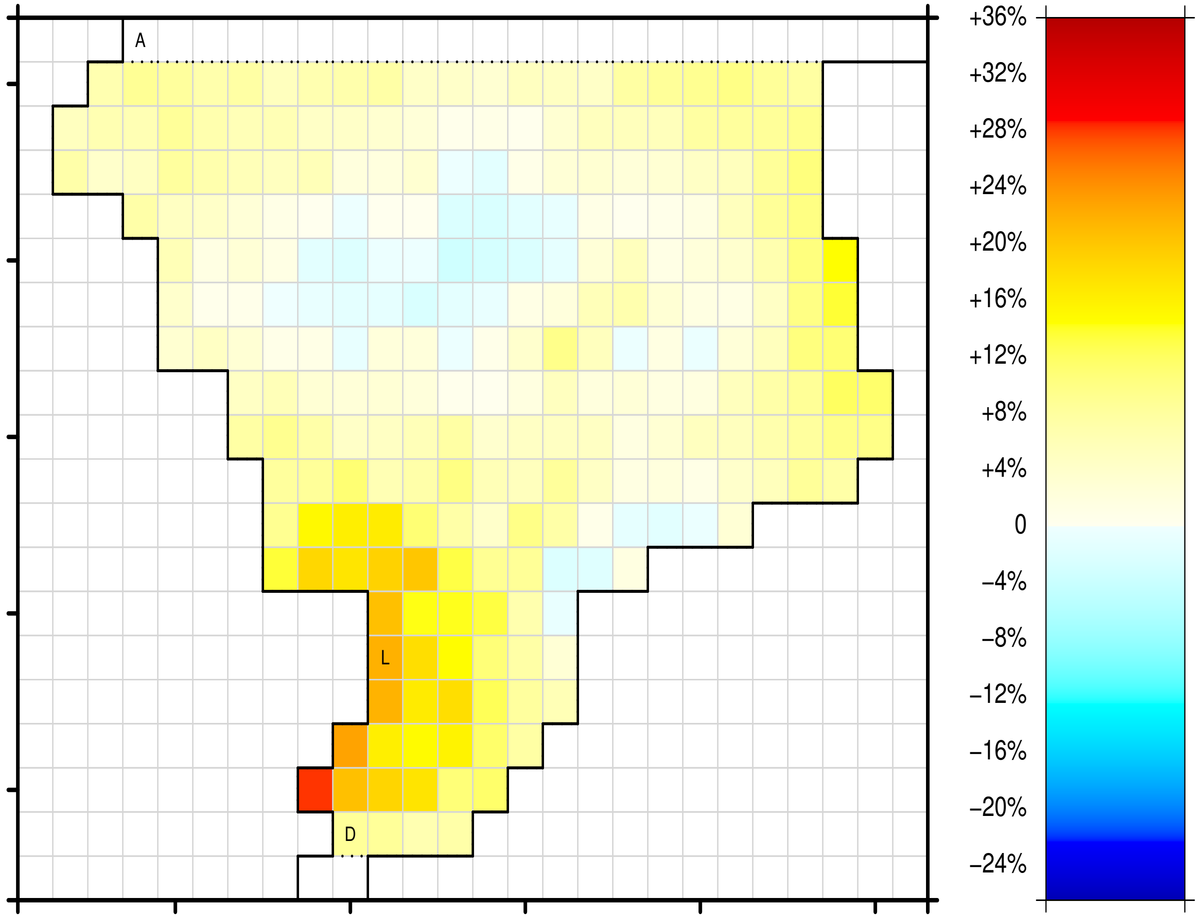
Deviations in 50y return levels from long-term means: 1975



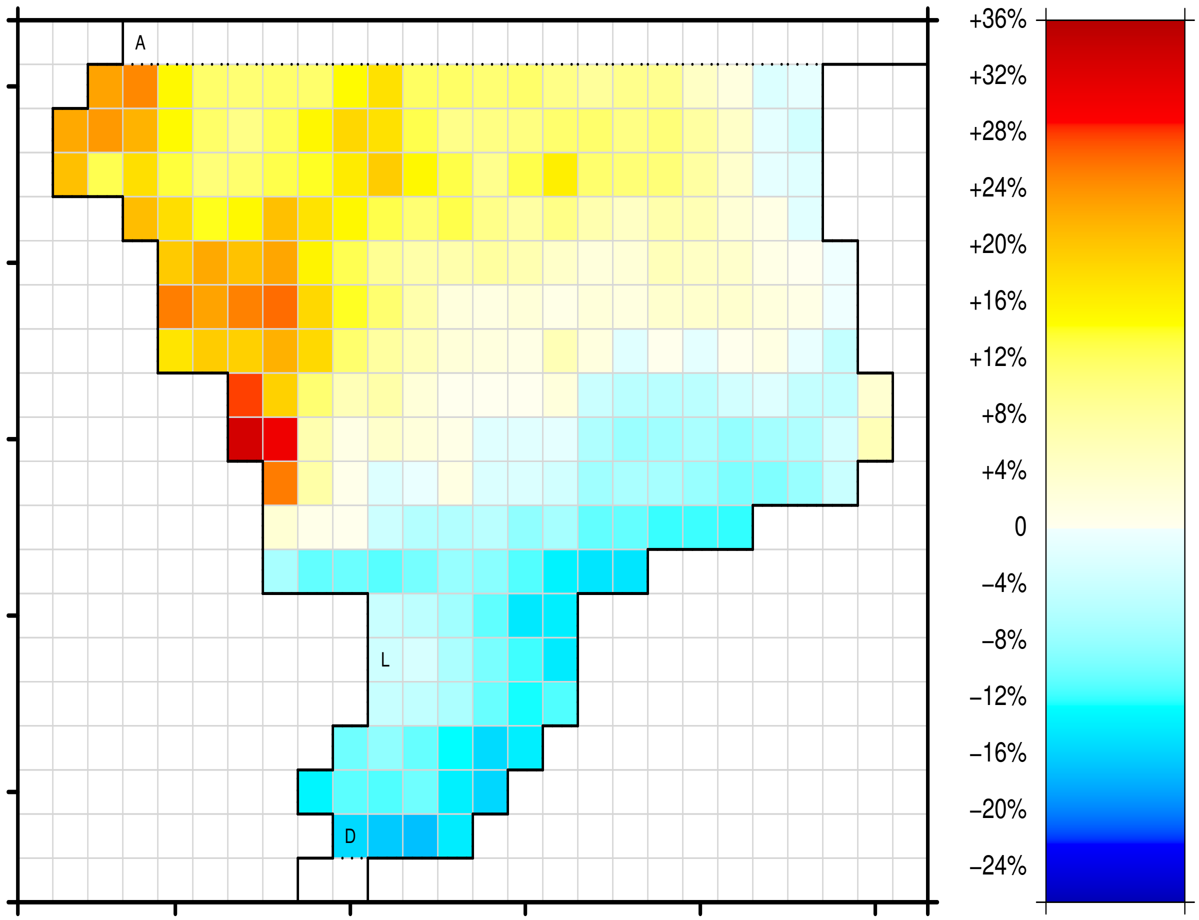
Deviations in 50y return levels from long-term means: 1982

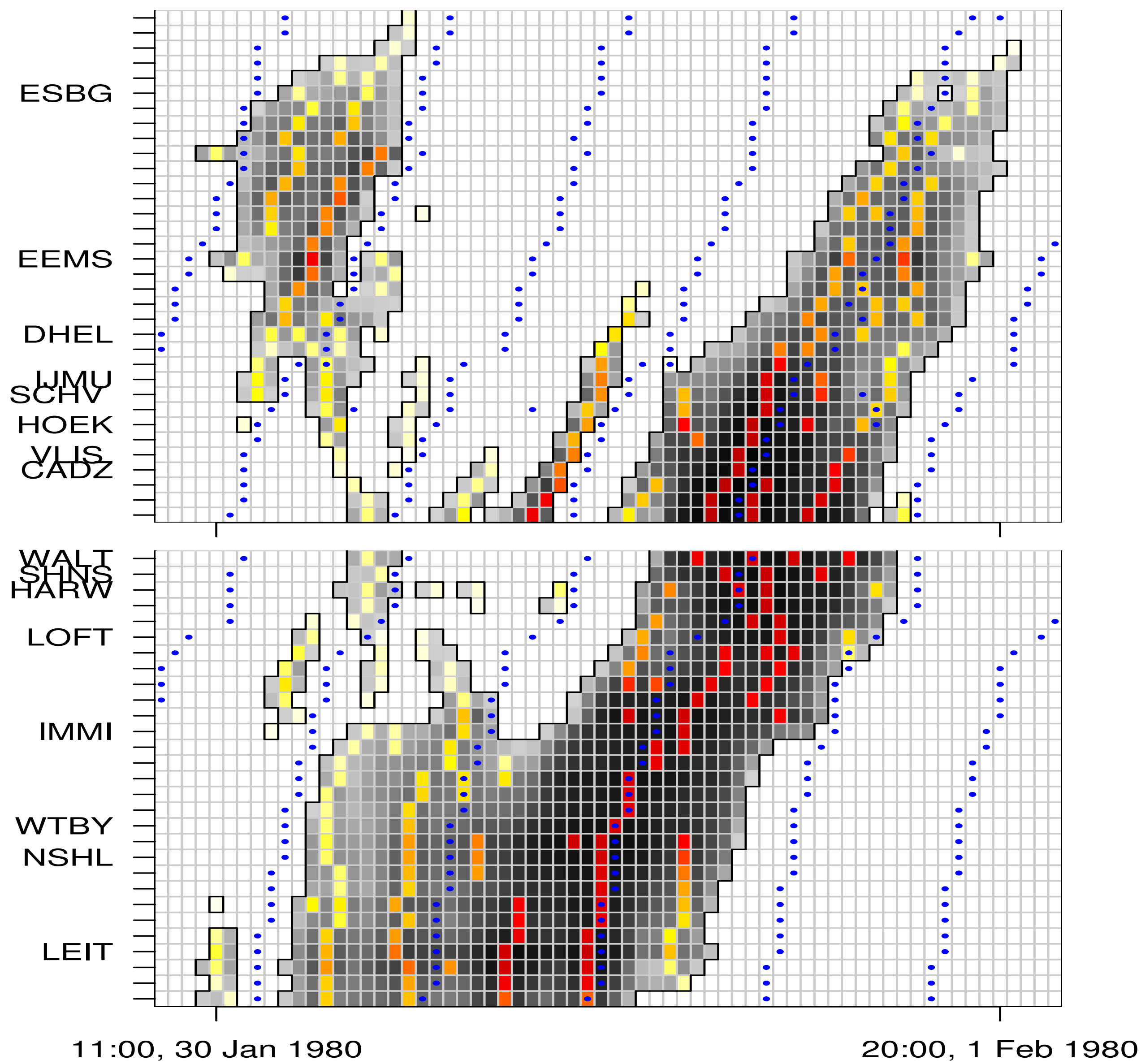


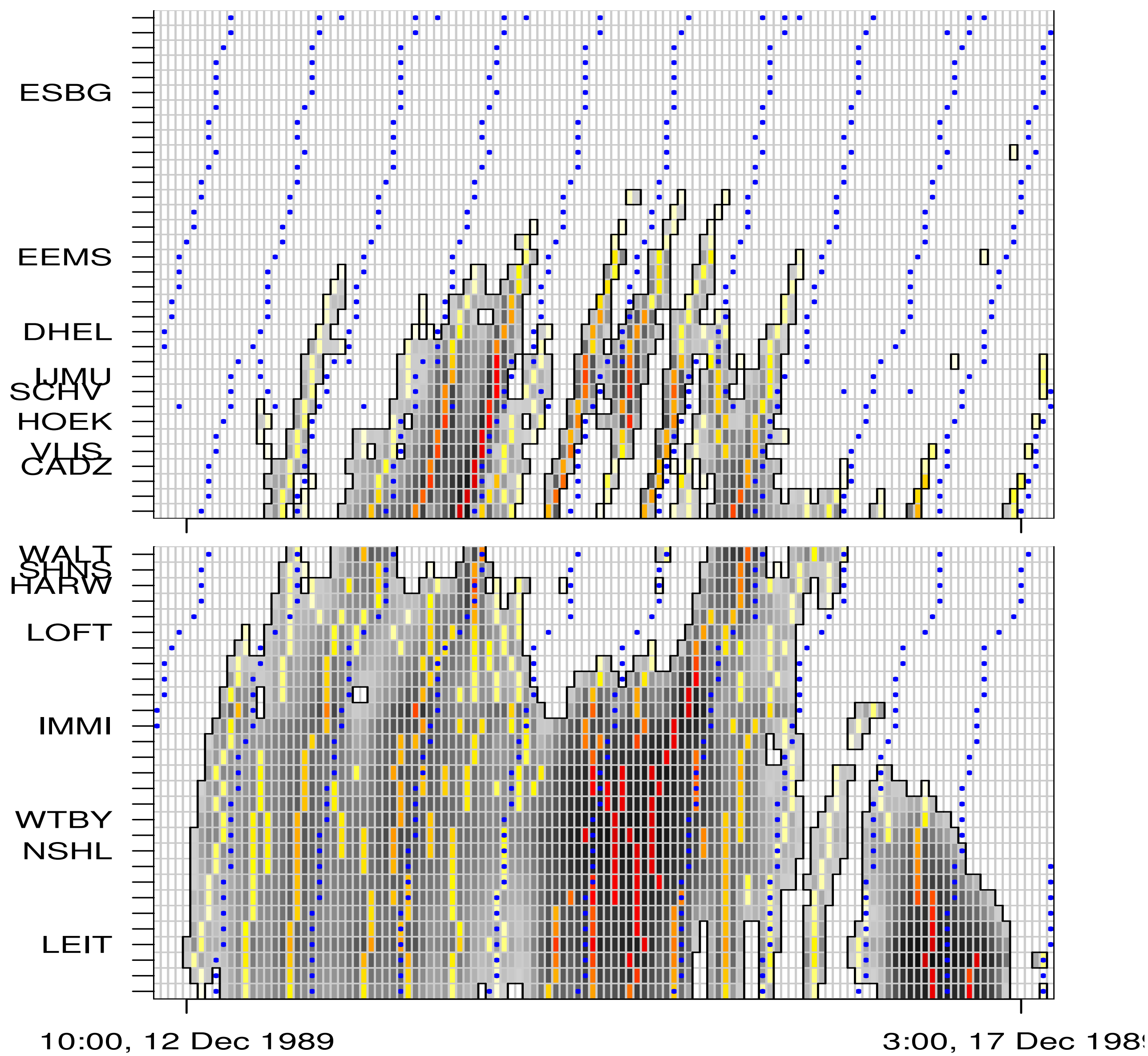
Deviations in 50y return levels from long-term means: 1990

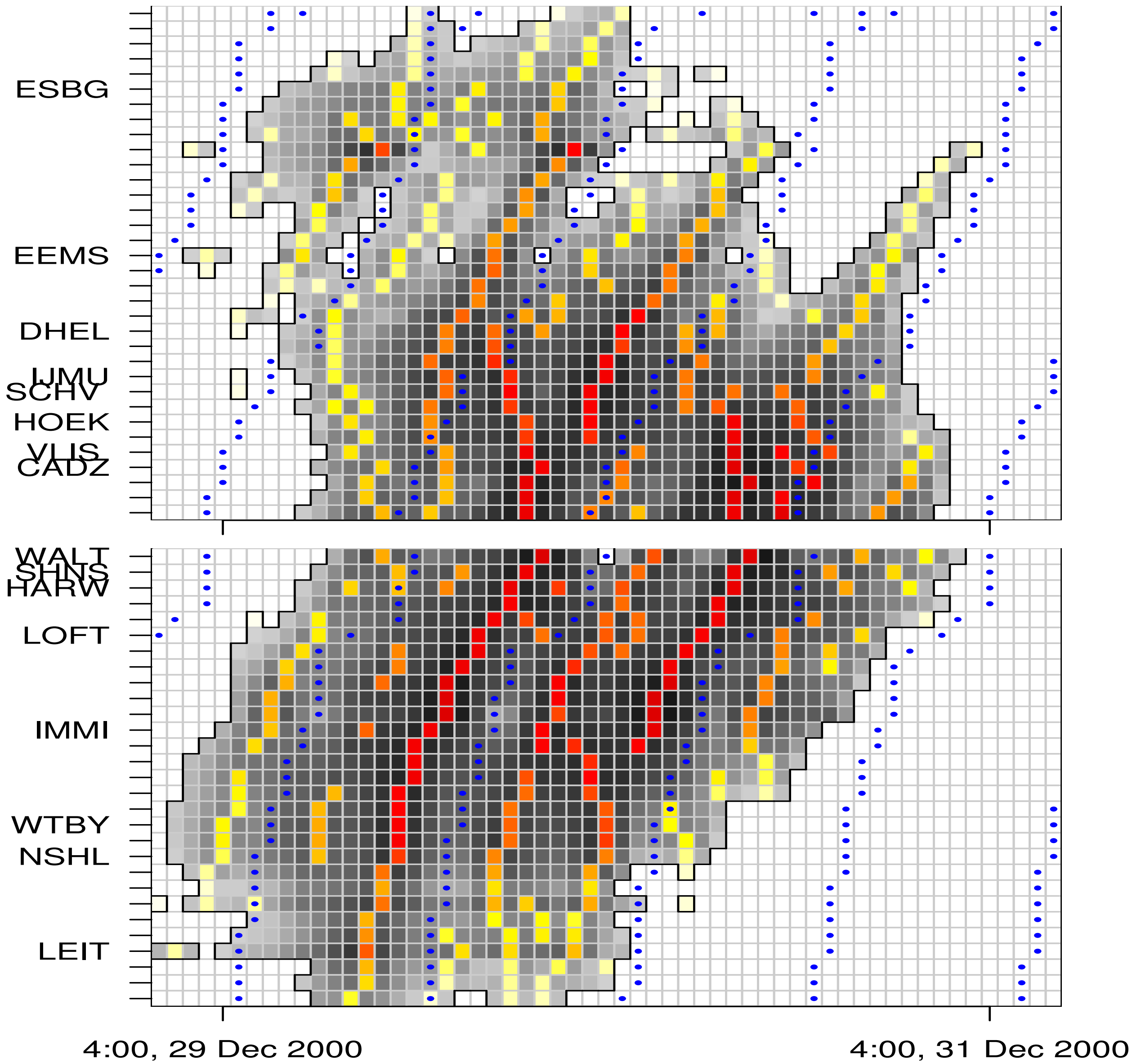


Deviations in 50y return levels from long-term means: 2000




ESBG
EEMS
DHEL
IJMU
SCHV
HOEK
VLIS
CADZ
WALT
SHNS
HARW
LOFT
IMMI
WTBY
NSHL
LEIT
11:00, 30 Jan 1980
20:00, 1 Feb 1980


ESBG
EEMS
DHEL
IJMU
SCHV
HOEK
VLIS
CADZ
WALT
SHNS
HARW
LOFT
IMMI
WTBY
NSHL
LEIT
10:00, 12 Dec 1989
3:00, 17 Dec 198


ESBG
EEMS
DHEL
IJMU
SCHV
HOEK
VLIS
CADZ
WALT
SHNS
HARW
LOFT
IMMI
WTBY
NSHL
LEIT
4:00, 29 Dec 2000
4:00, 31 Dec 2000

References

Chavez-Demoulin, V. (1999) *Two Problems in Environmental Statistics: Capture-Recapture Analysis and Smooth Extremal models*. Ph.D. thesis, Swiss Federal Institute of Technology (Laussanne).

Chavez-Demoulin, V., Jarvis, S. A., Perera, R., Roehrl, A. S. A., Schmiedl, S. W. and Sondergaard, M. P. (2003) Extreme datamining. In *Between Data Science and Applied Data Analysis; [P]roceedings of the 26th Annual Conference of the Gesellschaft Fr Klassifikation E. V*, 387–394. Springer-Verlag.

Davison, A. C. and Ramesh, N. I. (2000) Smoothing sample extremes. *J. Roy. Statist. Soc. Ser. B*, **92**, 191–208.

Gaetan, C. and Grigoletto, M. (2004) Smoothing sample extremes with dynamic models. *Extremes*, in press.

Hall, P. and Tajvidi, N. (2000) Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statist. Sci.*, **15**, 153–167.

Pauli, F. and Coles, S. G. (2001) Penalized likelihood inference in extreme value analyses. *J. Appl. Statist.*, **28**, 547–560.

Ramesh, N. I. (2001) Statistical analysis of extreme forest fires. In *Proceedings of the IUFRO 4.11 conference on Forest Biometry, Modelling and Information Science, University of Greenwich, June 2001*. International Union of Forest Research Organizations (Vienna, Austria). <`http://cms1.gre.ac.uk/conferences/iufro/proceedings/`> (accessed 17 November 2004).