Estimating Extremes

Clive Anderson, Univ. of Sheffield

SEAMOCS, Palmse, Oct 2007

Extremes of wave characteristics from data?   How uncertain?

1. Data, classical extremal estimation,
   uncertainties...
2. Refinements and Complications
3. Developments

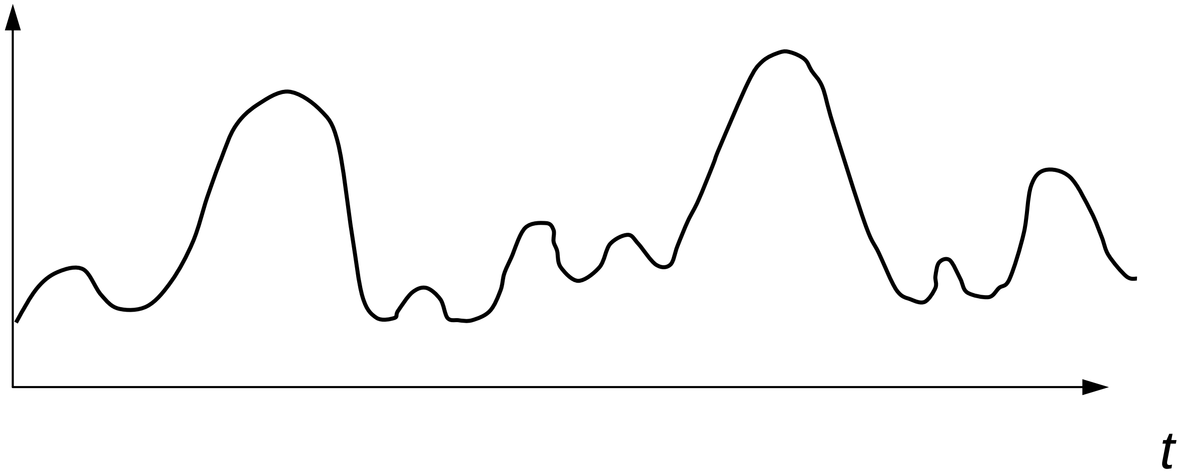
Specific focus: Estimation of high quantiles, return levels

1 Data, Classical Methods

Data:

- buoy, platform
- model
- satellite
- ship

} usually regular, frequent



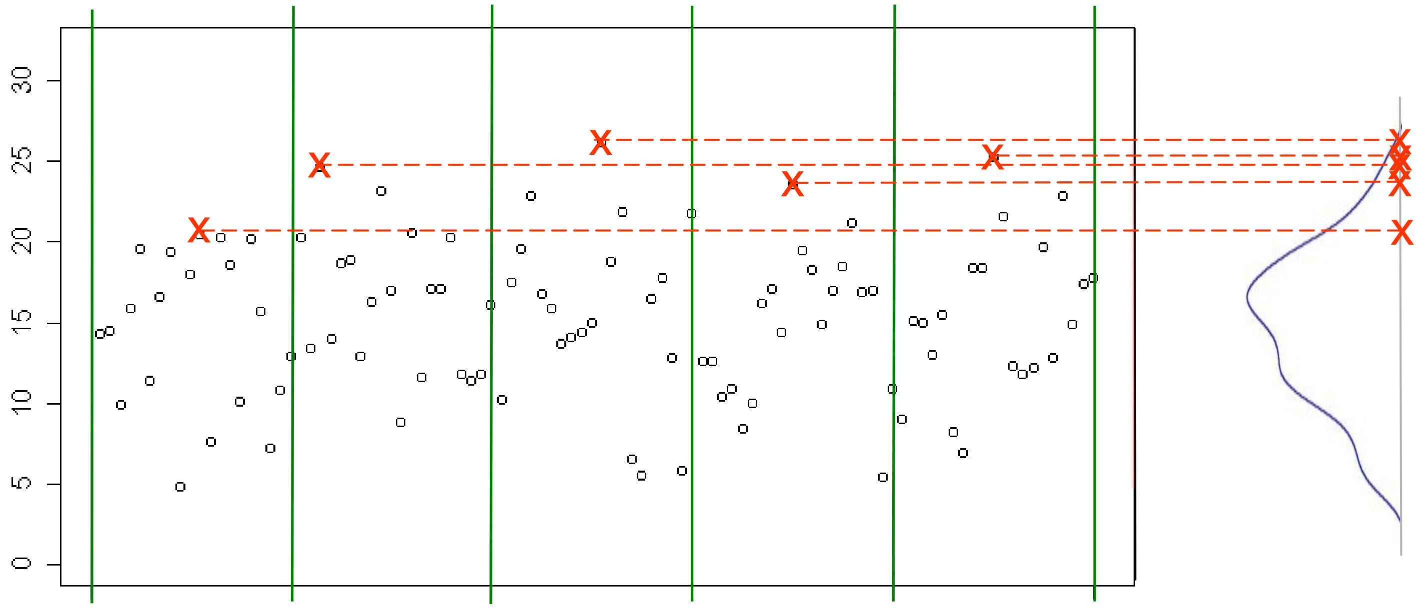
General Strategy: allow extreme observations to drive estimation model

'Nobody believes in a ~~theory~~ except the person who invented it; everyone believes in an observation except the person who made it.'

Albert Einstein

(a) Classical Estimation: Block Maxima

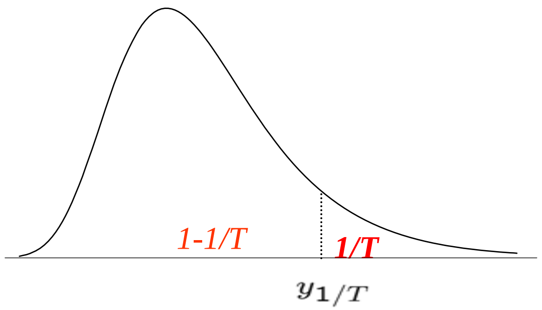
usually annual maxima



Fit a Generalized Extreme Value distribution to observed maxima

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

giving the $1 - 1/T$ quantile

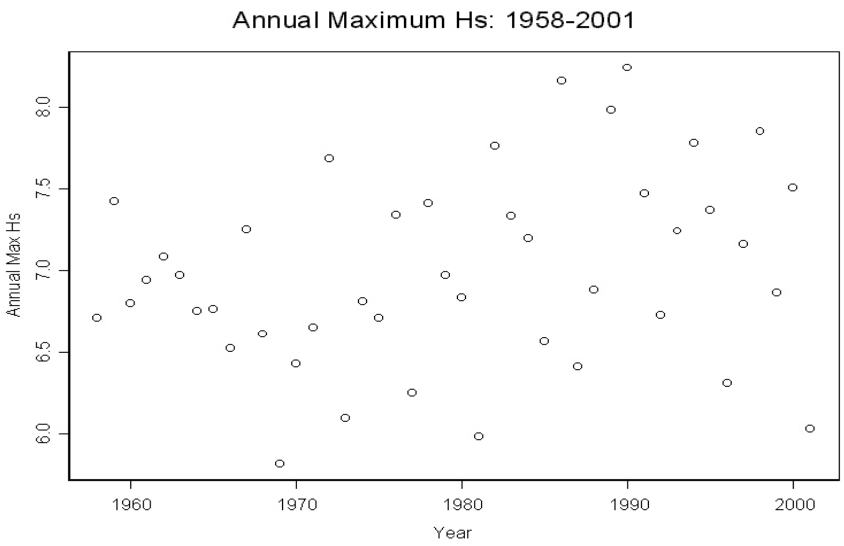


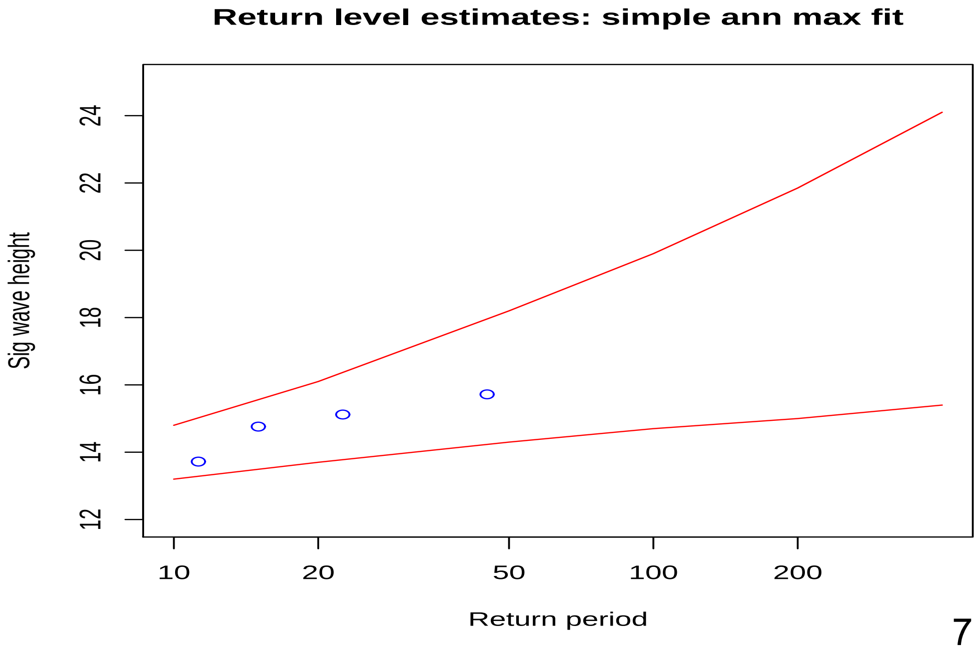
T-year return level

$$r_T = \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1 - 1/T)\}^{-\xi}\right]$$

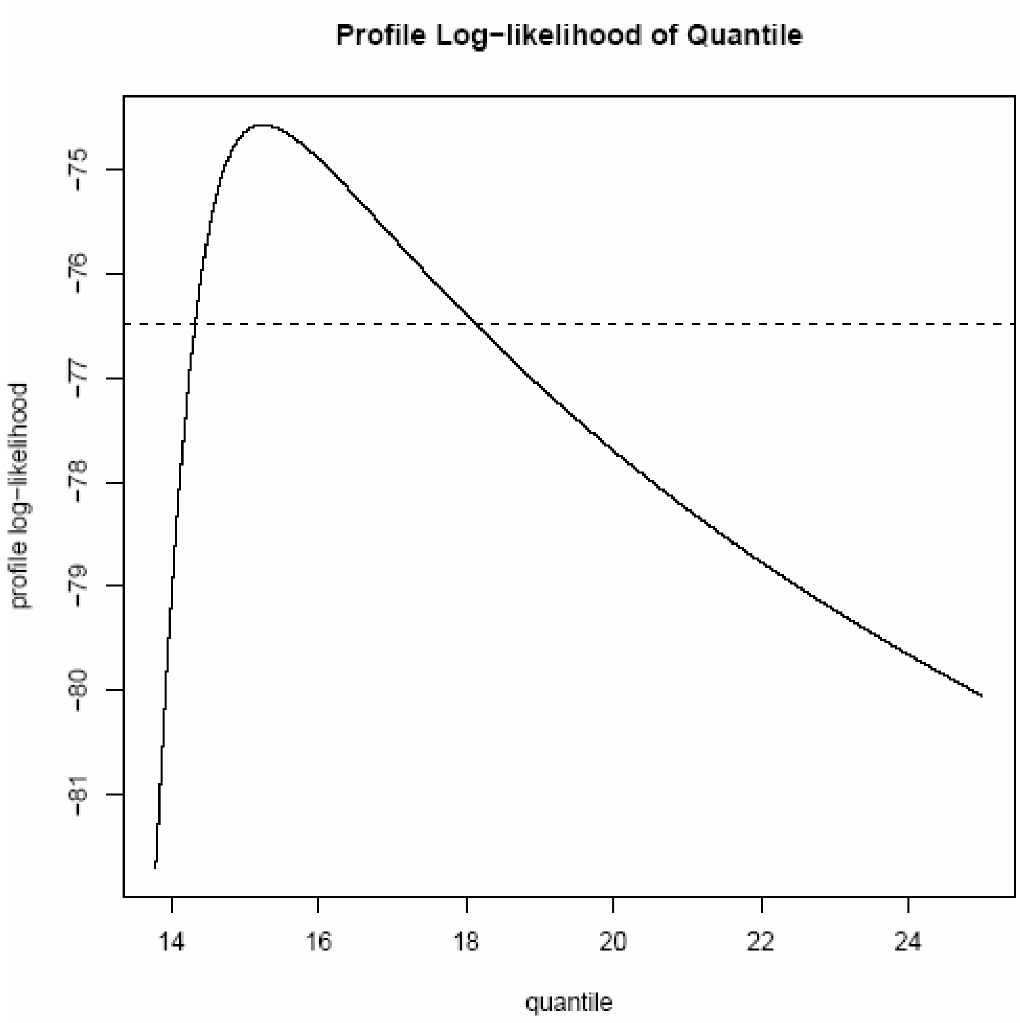
Maximum likelihood estimation of μ, σ, ξ leads to confidence intervals for r_T

ERA40 model data, 21°W 54°N

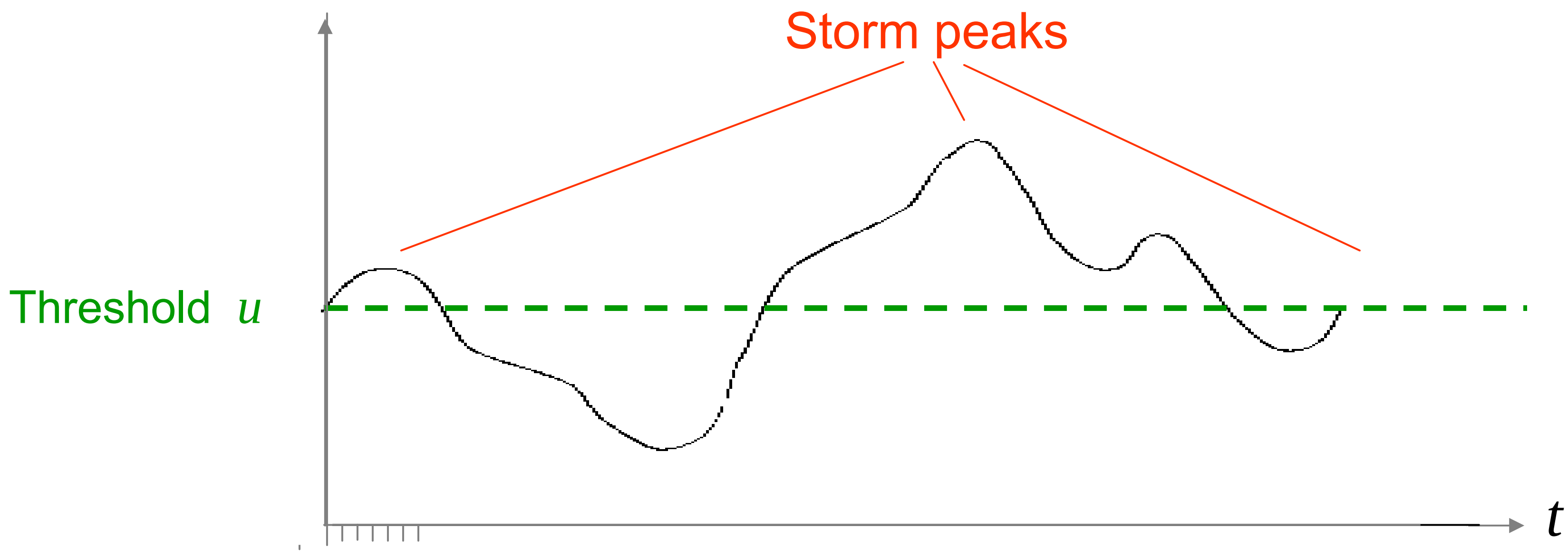




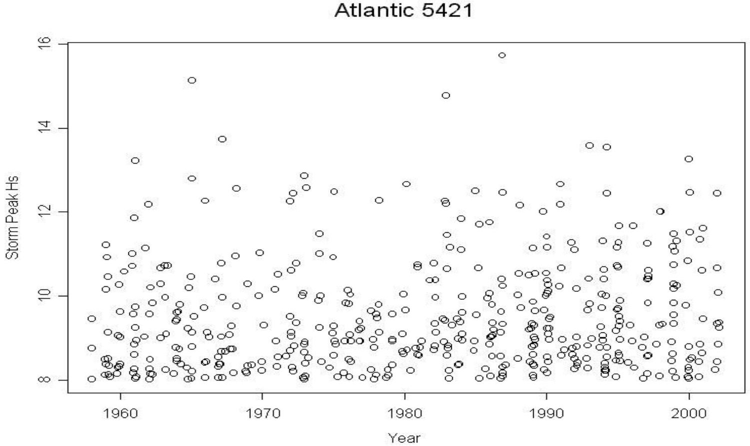
Example: Profile log-likelihood of 50-year return level



(b) Classical Estimation: Threshold Methods

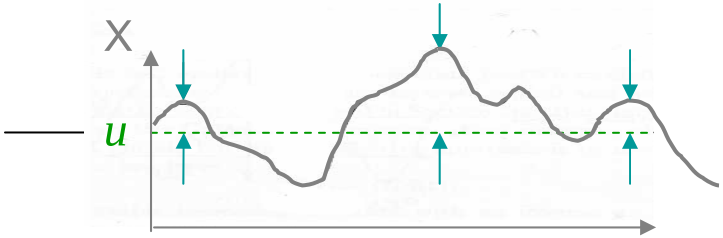


Concentrate attention on all storm peaks over a high threshold



Generalized Pareto distribution for

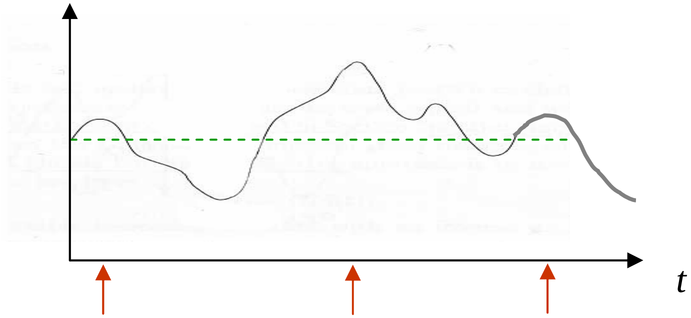
excesses of threshold u



$$Pr(X > u + y \,|\, X > u) = \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \quad y > 0$$

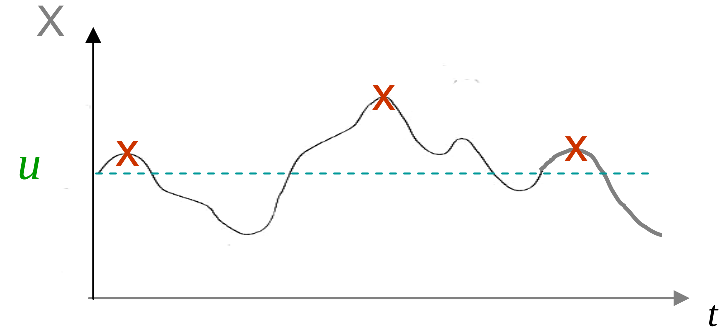
Poisson process of occurrence of storms

rate $\lambda(t)$



Point-process modelling:

a re-formulation, often more convenient for handling covariate dependence



Model $(t, x) =$ (times, sizes) of storm peaks by an inhomogeneous Poisson process with intensity density

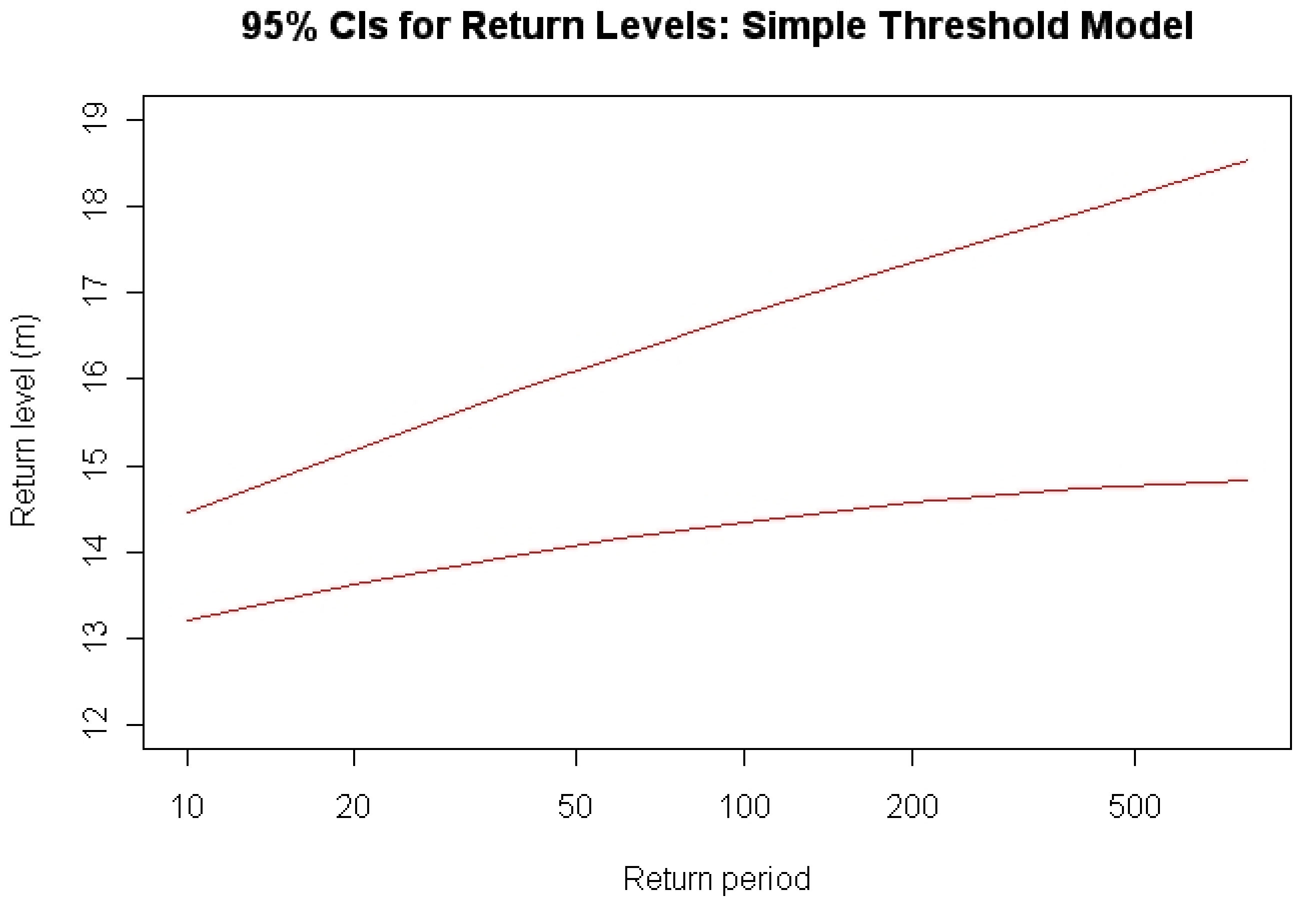
$$\lambda(t,x) = \frac{1}{\sigma_p(t)} \left\{ 1 + \xi_p(t) \frac{x - \mu_p(t)}{\sigma_p(t)} \right\}^{(-1/\xi_p(t))-1} \qquad ((t,x) \in C)$$

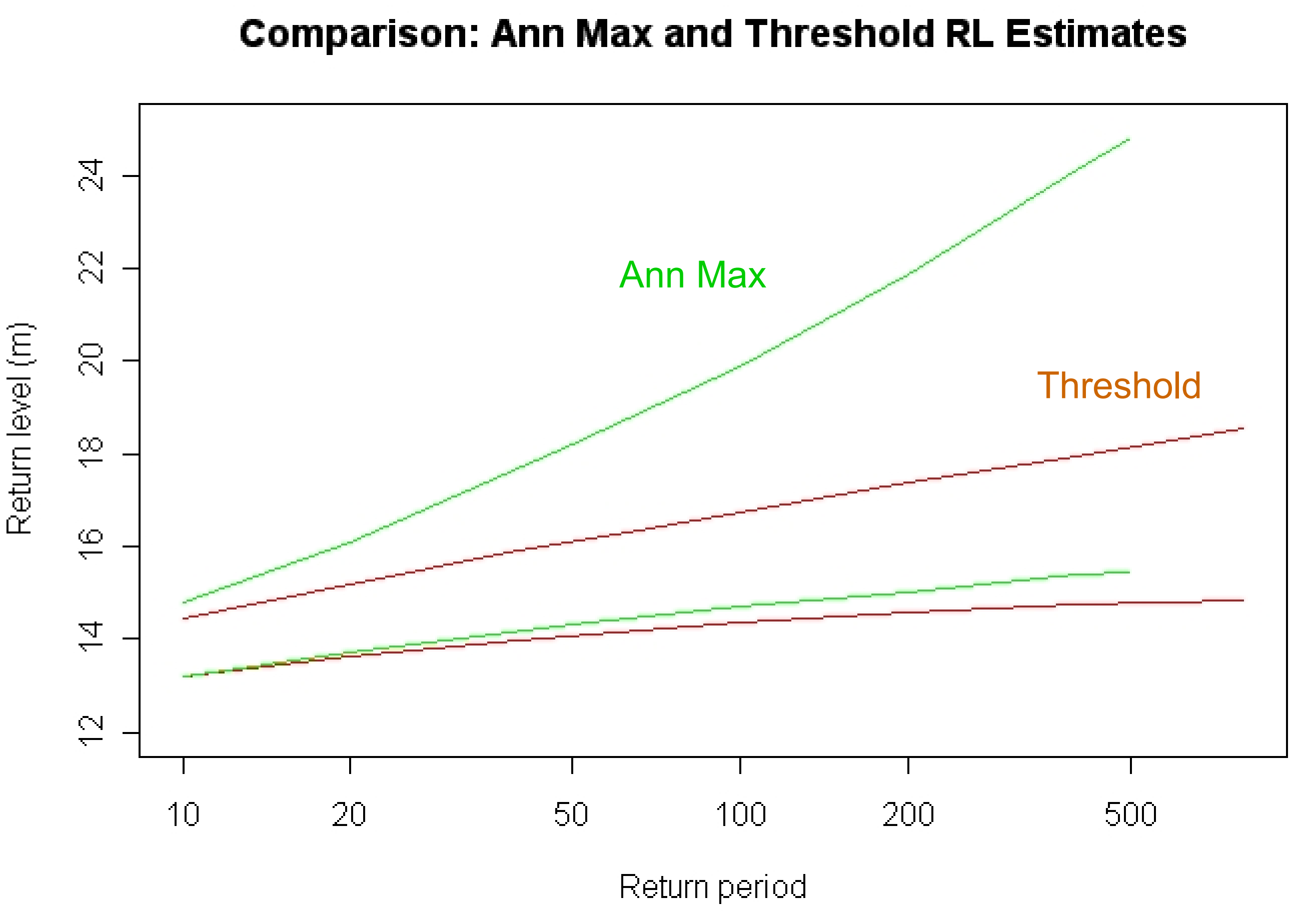
for a suitable region C, and possibly time-varying parameters σ_p, μ_p and ξ_p

Convenient for time-varying thresholds $u(t)$

Inference via likelihood methods as before.

ERA40 model data, 21°W 54°N



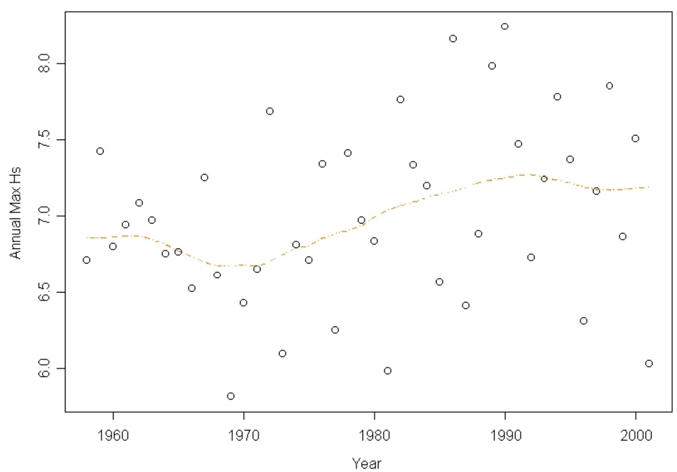


Comparison: Ann Max and Threshold RL Estimates
Ann Max
Threshold
Return level (m)
12
14
16
18
20
22
24
10
20
50
100
200
500
Return period

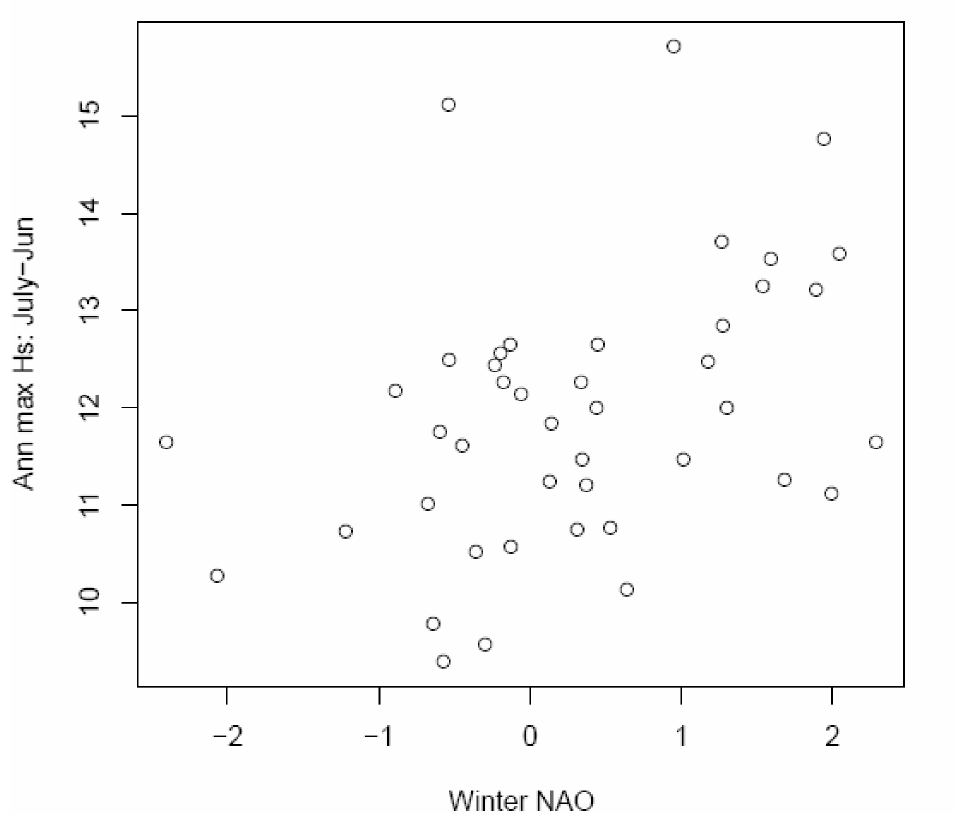
2 Refinements and Complications

(a) Non-stationarity, Covariate Dependence

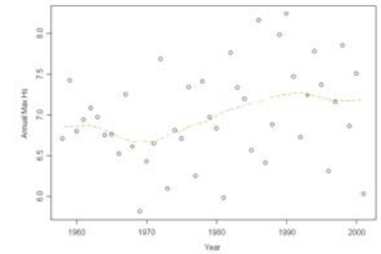
eg annual maximum possibly changing with time, with NAO, ...



Atlantic 5421 Annual Max (July-June) vs Winter NAO
1958/9 – 2001/2



Accommodate changes with time or in response to other variables by allowing GEV/GPD/point process parameters to vary.



eg $$\mu(t) = \beta_0 + \beta_1 z_1(t) + \ldots + \beta_p z_p(t)$$

known functions of t

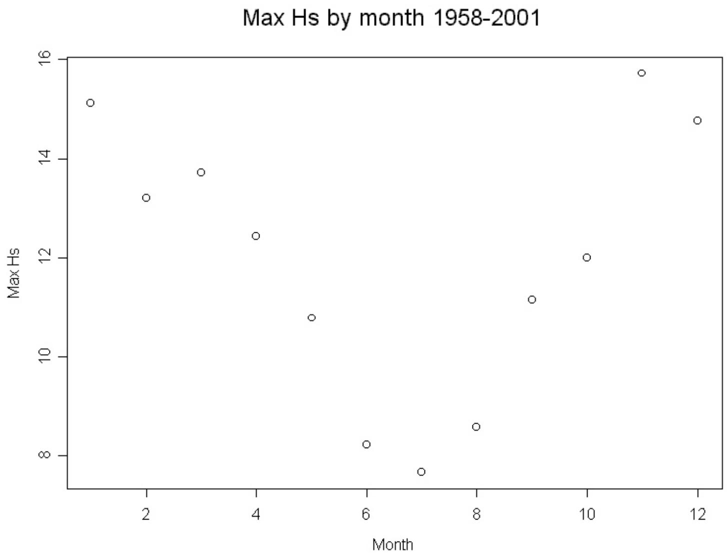
for new parameters $\beta_0, \beta_1, \ldots$ which can be estimated and tested.

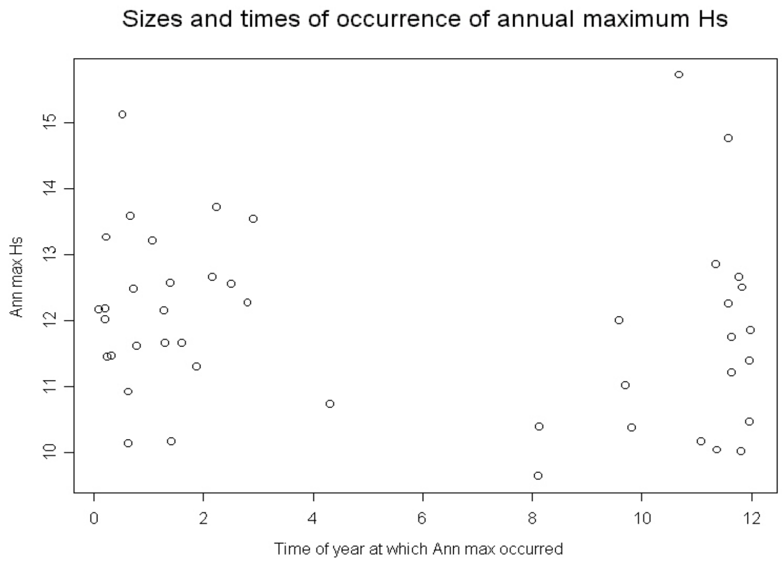
eg $$\beta_0 + \beta_1 t + \beta_2 \cos(2\pi t) + \beta_3 \sin(2\pi t) + \beta_4 NAO(t)$$

and similarly for ξ and $\log \sigma$

Atlantic 5421 data: strong evidence of NAO association,
no evidence of linear time-trend

(b) Seasonality



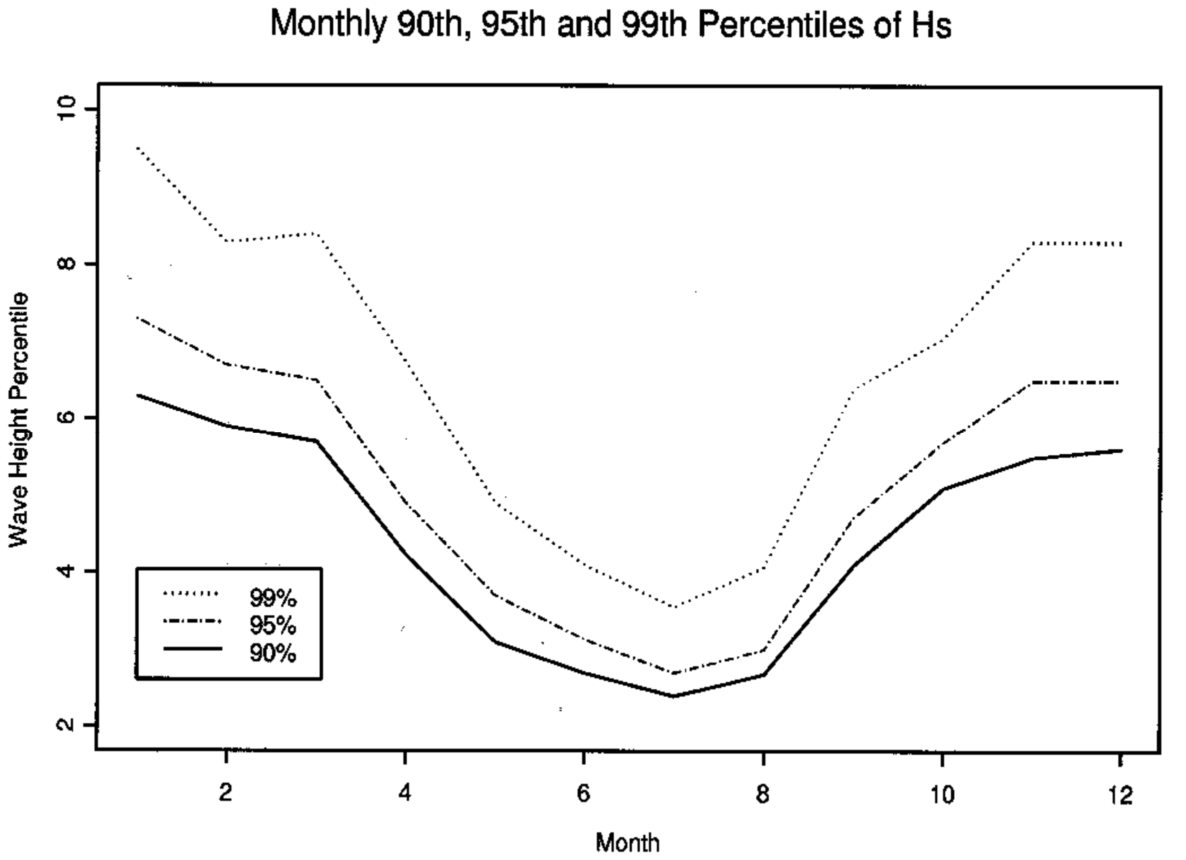


Take account of seasonality in estimation

– by using seasonally-varying parameters in either block maxima or threshold modelling

Wave heights over 5m, Northern N. Sea 1979-1999

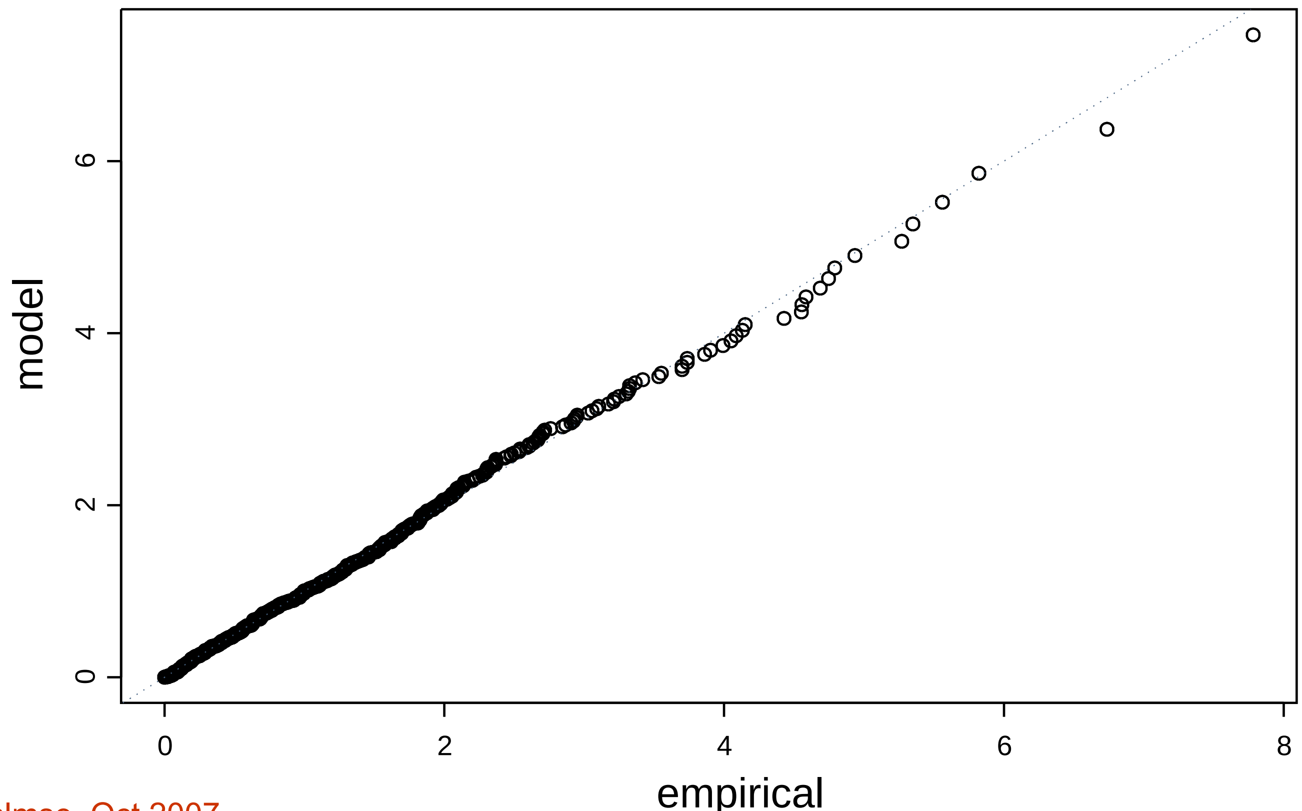
Seasonality



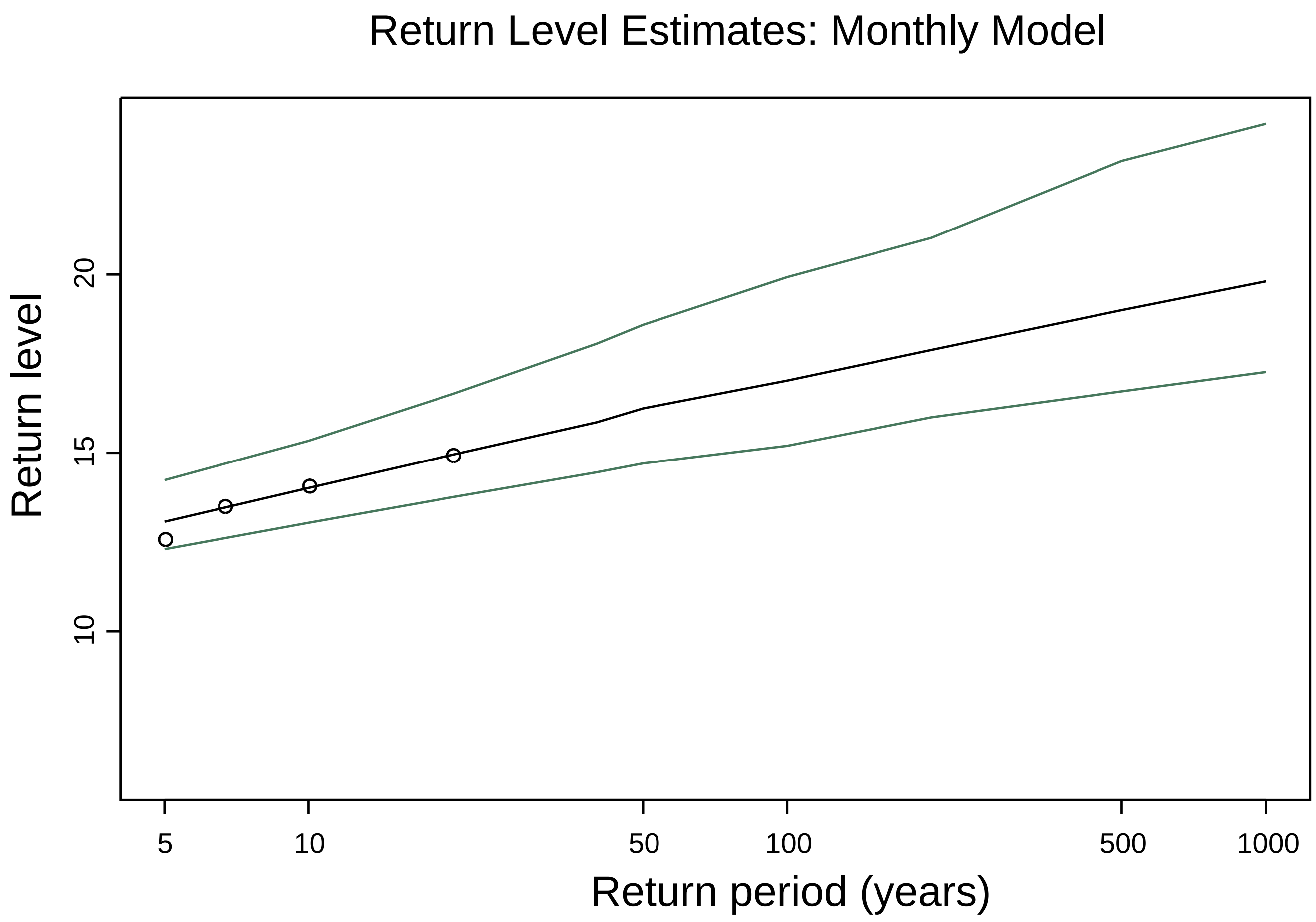
Separate-Months Point Process Model, N North Sea

$$
\begin{aligned}
\mu_p(t) &= \alpha_i \quad \text{for} \quad t \in \quad \text{Month}_i \\
\log(\sigma_p(t)) &= \beta_i \quad \text{for} \quad t \in \quad \text{Month}_i \\
\xi_p(t) &= \xi, \quad \text{for all } t,
\end{aligned}
$$

Standardized QQ Plot: Monthly Model



95% Confidence Intervals for Return Levels, NN Sea



The Effect of Neglecting Seasonality

(Carter & Challenor (1981))

100 year return level, Northern N Sea

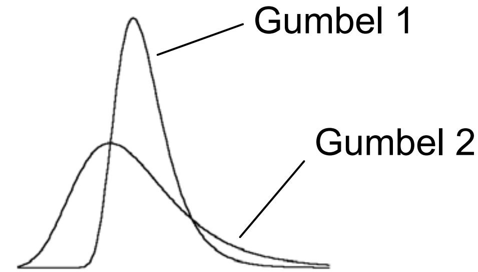
Model	ML Estimate	95% CI
Stationary	15.2	(13.8, 16.6)
Seasonal: separate months	17.0	(15.2, 19.9)

... under-estimation, over-confidence in this case

Explanation - 2-season case

Annual max = max of Season 1 and Season 2

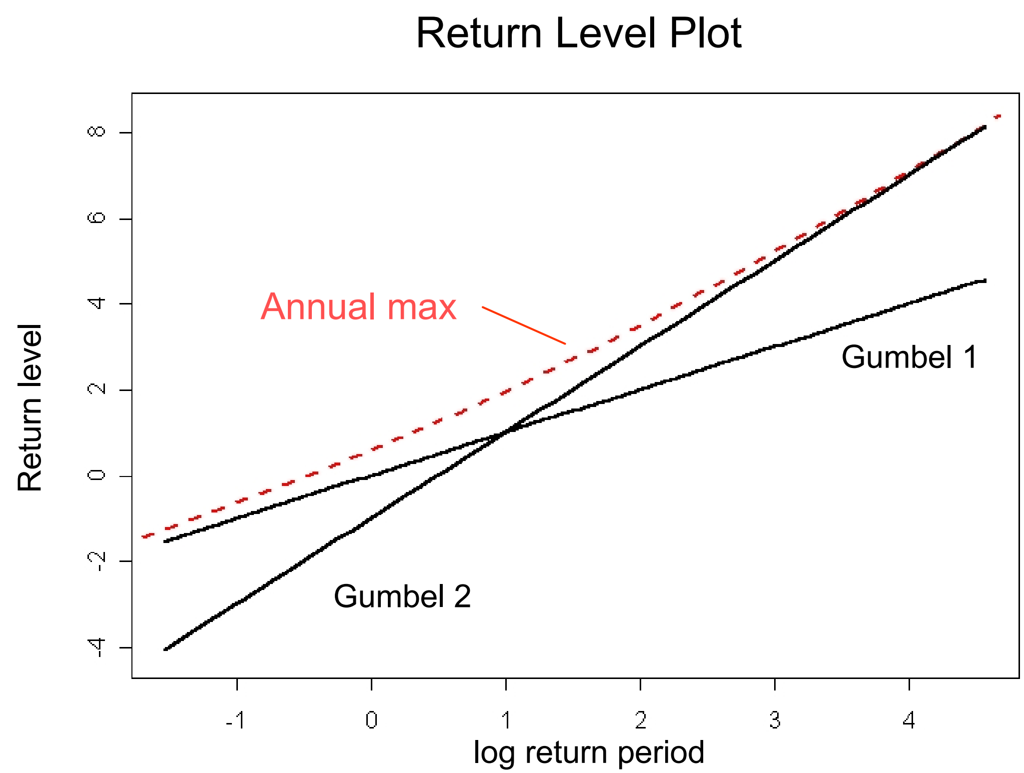
suppose independent Gumbel distributed



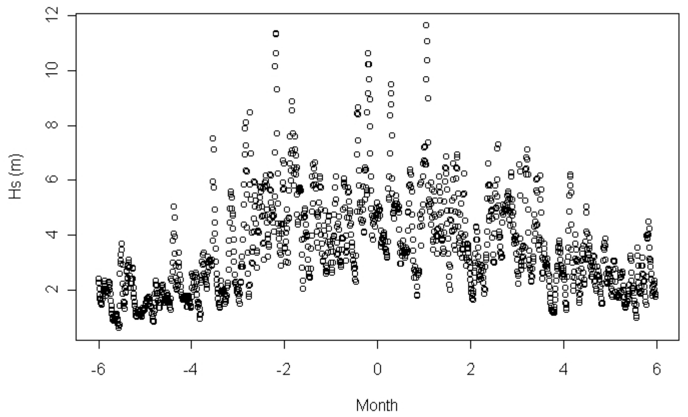
A single Gumbel fitted
to Annual Max
≡ linear approximation to

so underestimates true return
levels at high return periods

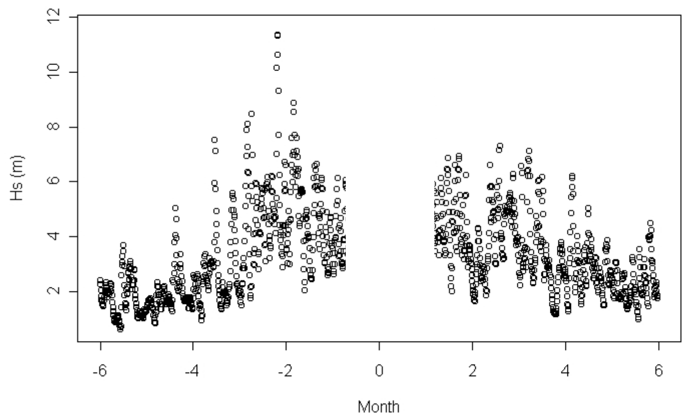
Moreover se of ret level est
is $\propto$ estimated slope, so
underestimates true uncertainty
too



(c) Effect of gaps in data



Annual maxima data: analysis based on censoring possible, but...

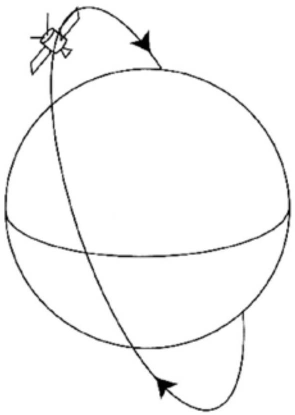


Threshold method: effect easier to accommodate, likely to have smaller influence

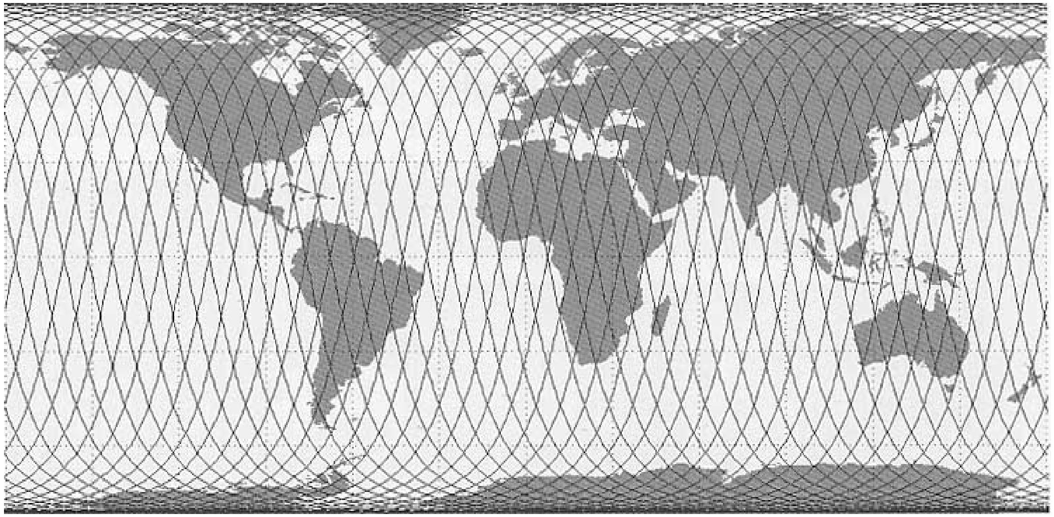
In general:

- loss of precision
- if data missing because of size, more detailed statistical modelling needed

(d) Estimation from Satellite Data



Ground track: ERS-1, 3 day repeat

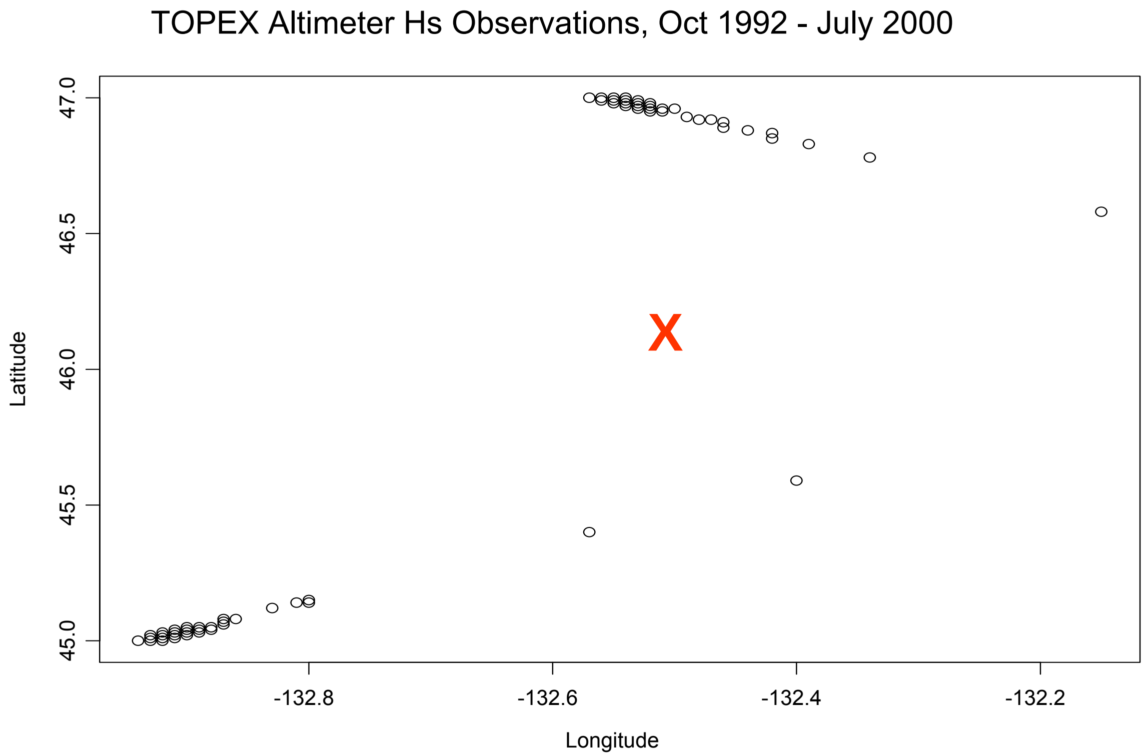


Problems:

- Spatial — no data from location of interest
- Temporal — miss storm peaks

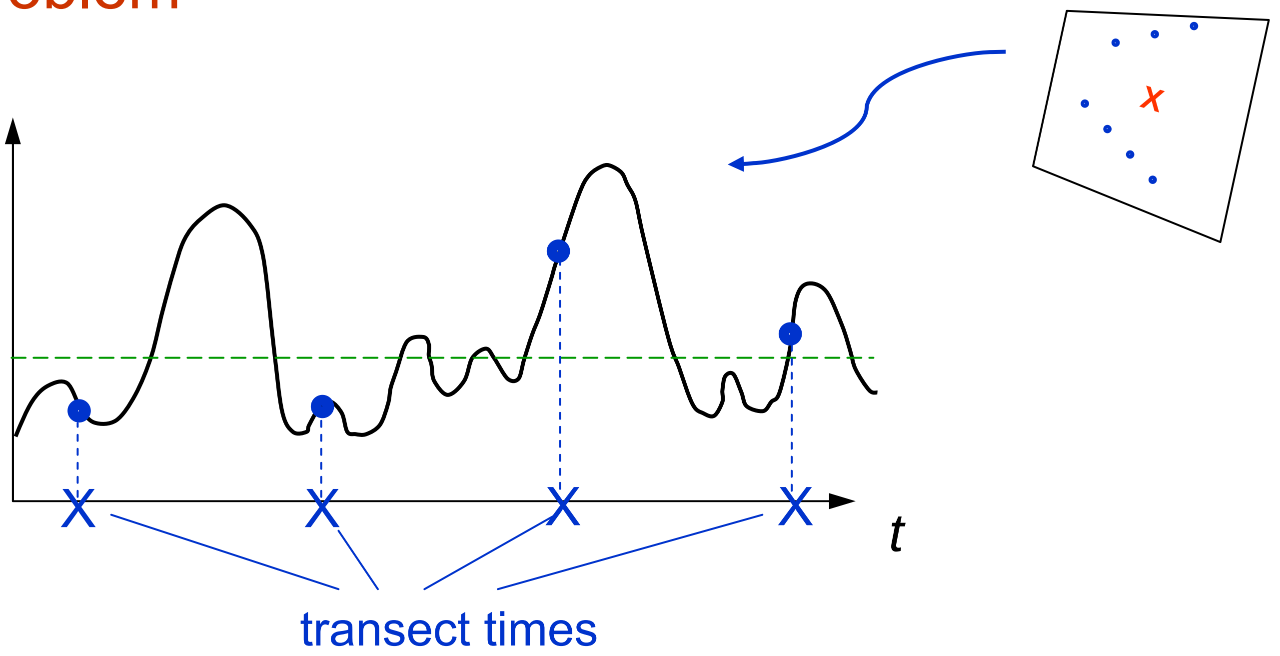
Spatial problem

Example: Wave heights off Vancouver
TOPEX observations



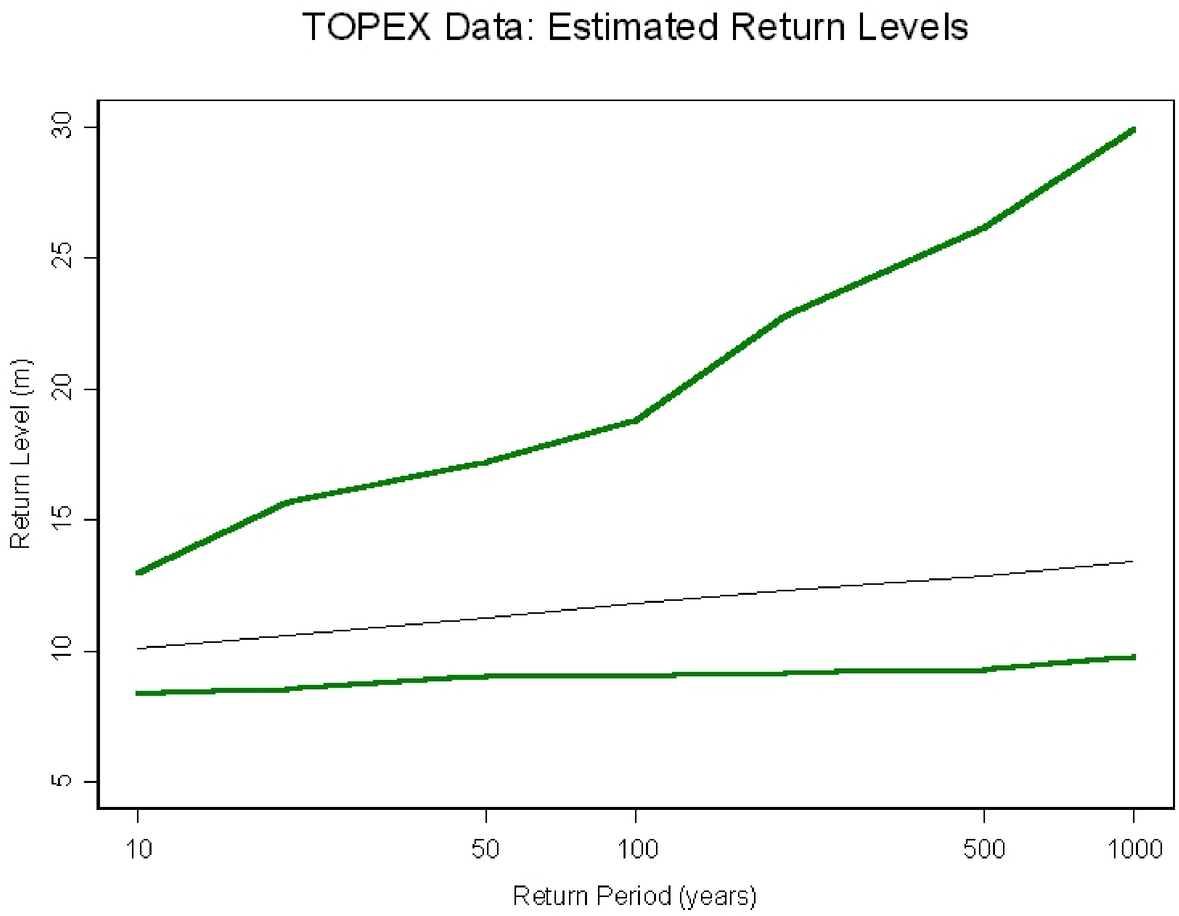
Take nearby observations to be representative of those at the target

Temporal problem



a) over-threshold observations unlikely to be storm peaks

b) many storms likely to be missed



eg 95% conf interval for 100-year return level for H_s = 9.1 – 18.8 m. No evidence of a trend in extremes

(e) Estimation from Multiple Sources of Data

eg satellite + buoy; data $\boldsymbol{x}_s$ and $\boldsymbol{x}_b$ say on extremes

Strategies:

i. Consistency checks over range of data

ii. For extremes: suppose true storm peak wave heights $\boldsymbol{x}$ are governed by a distribution $[\boldsymbol{x} \mid \boldsymbol{\theta}]$

and conditionally on $\boldsymbol{x}$ suppose observational models for data:

$$[\boldsymbol{x}_s \mid \boldsymbol{x}, \boldsymbol{\alpha}_s]$$

$$[\boldsymbol{x}_b \mid \boldsymbol{x}, \boldsymbol{\alpha}_b]$$

where $\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_b$ are parameters specific to the data sources and $\boldsymbol{\theta}$ describes the underlying wave characteristics of interest

(eg $\boldsymbol{x}_s = \alpha_{0s} + \alpha_{1s}\boldsymbol{x} + \epsilon_s$

observation with bias and instrument error)

Then, if observation processes conditionally independent given $\boldsymbol{x}$,

likelihood is

$$L = \int_{\boldsymbol{x}} [\boldsymbol{x}_s \mid \boldsymbol{x}, \boldsymbol{\alpha}_s][\boldsymbol{x}_b \mid \boldsymbol{x}, \boldsymbol{\alpha}_b][\boldsymbol{x} \mid \boldsymbol{\theta}]$$

whence estimation and uncertainty estimates

If observation processes are dependent, then L is a pseudo-likelihood and inference still feasible but more challenging.

3 Developments

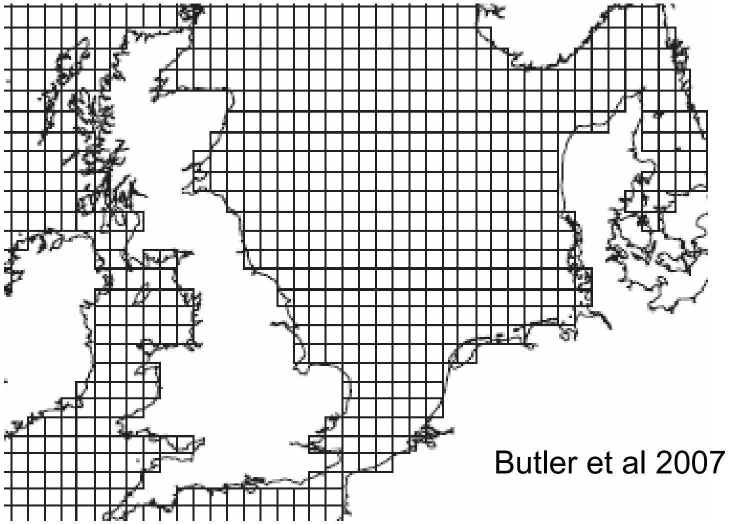
(a) Spatial Estimation

- neighbouring locations contain useful information about a site
- spatial structure is of interest in its own right

Represent spatial characteristics of extremes by allowing the parameters of the extremal distributions of §1 to depend smoothly (and non-parametrically) on location. Fit by local likelihood. Extends to spatio-temporal estimation

Chavez-Demoulin & Davison (2005), Butler et al (2007)

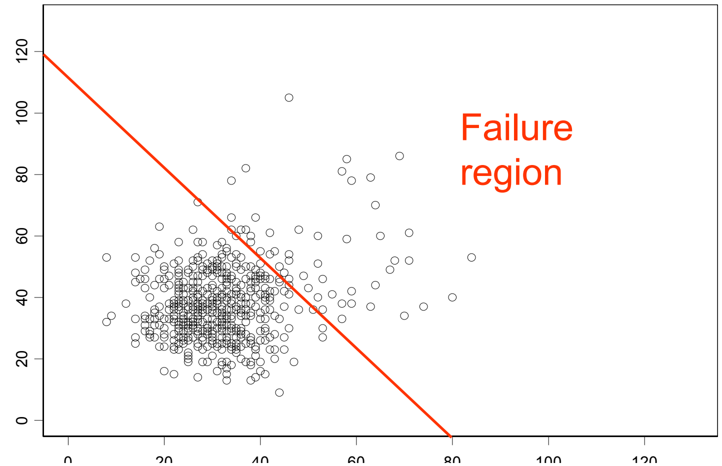
eg



Butler et al 2007

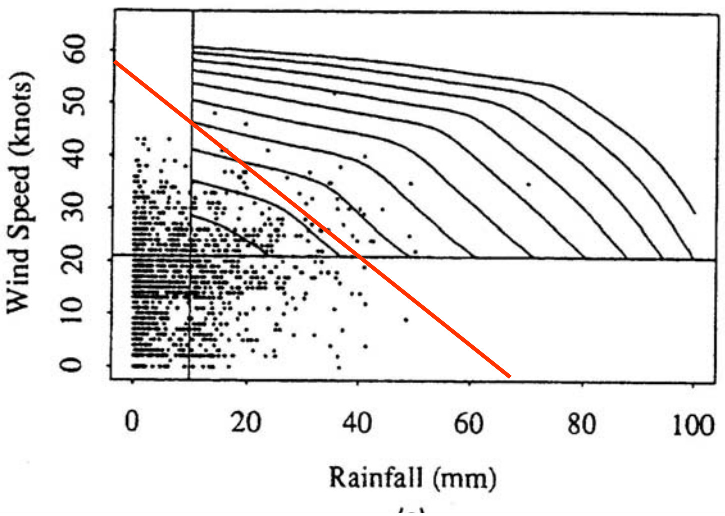
(b) Estimation of Multivariate Extremes

Relevant if structure threatened by a combination of large values of different variables



Ledford & Tawn (1997, 1998, 2003): estimation methods based on tail dependence models.

Example: from study for reservoir flood safety (DoE, Inst of Civ Eng)



Tail dependence model

Heffernan & Tawn (2003): powerful methods based on *conditional distribution* of components of a random vector given that at least one component is large.

(c) Model-Data Fusion for Extremes

Observational evidence

Model evidence

Comparison: model predictions vs observations helps us learn about model uncertainty and predictive reliability.

Systematize and quantify model uncertainty and reliability by treating model inadequacy as a random entity

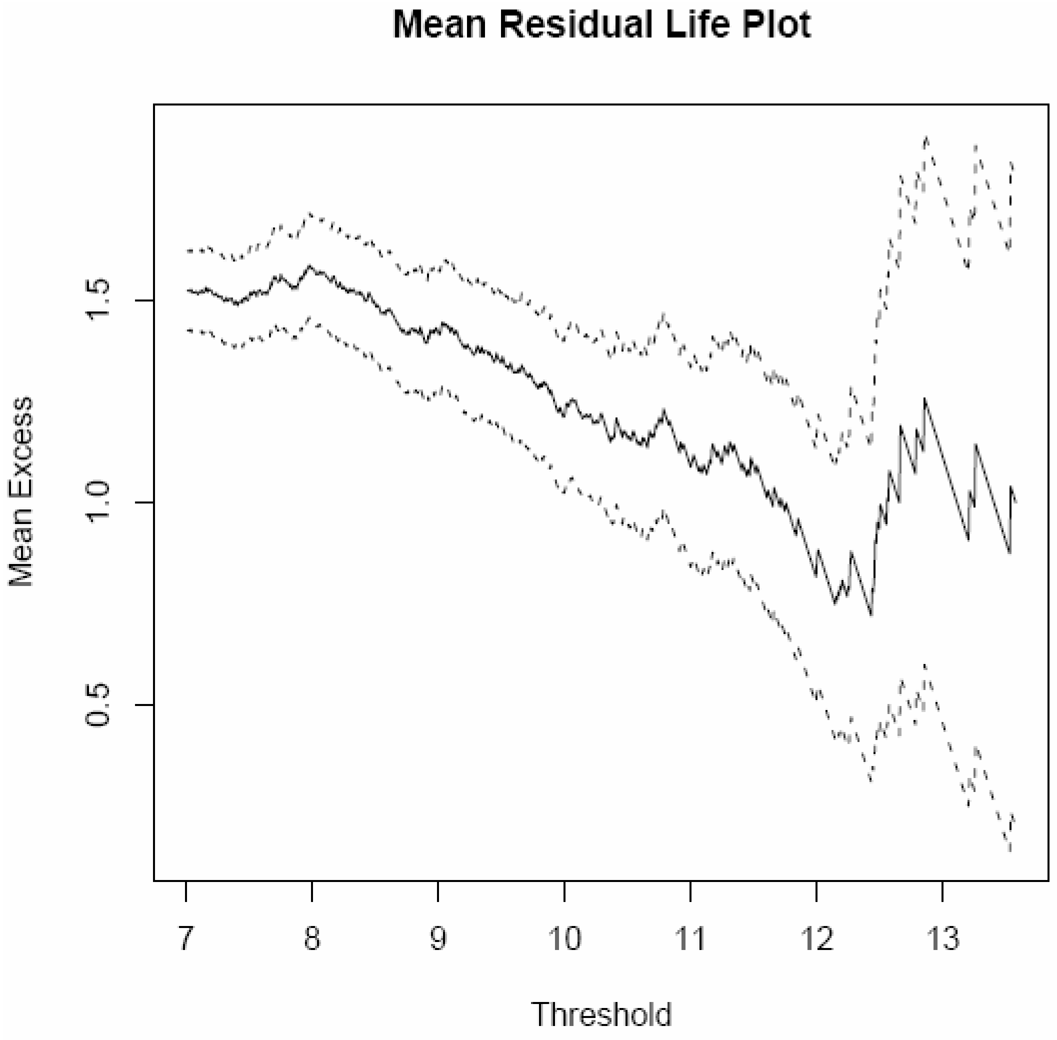
$$\text{data} = \text{model prediction} + \text{error}$$

model inadequacy
Gaussian process?

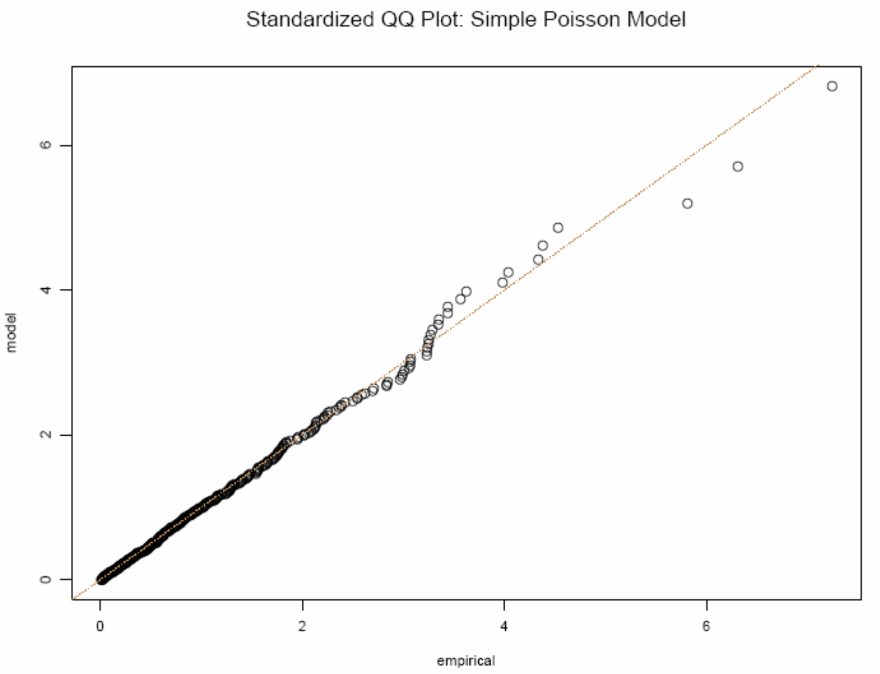
about which we update our knowledge in the light of model-data comparisons. Bayesian methods give a natural approach.

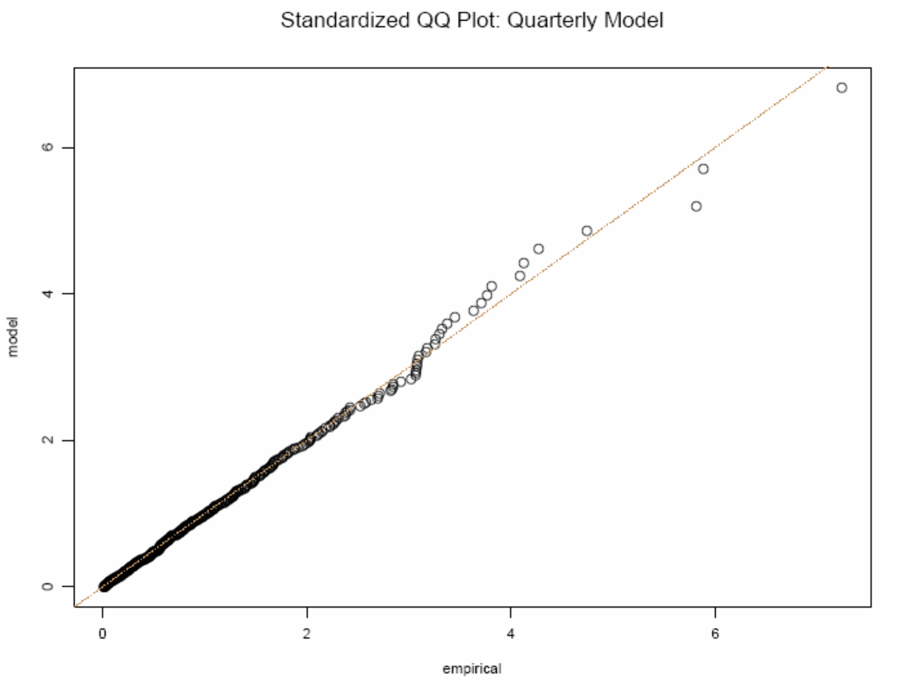
Being developed in programme *Managing Uncertainty in Complex Models (MUCM)* for non-extreme observations.

Atlantic 5421: Hs Mean Excess Plot



Atlantic 5421: QQ Plots for Threshold Models - Stationary and Seasonal





Atlantic 5421: RL CIs for a Seasonal and a Non-Seasonal model