A new approach to threshold selection in extreme value analysis

Paul Northrop and Claire Coleman
University College London
p.northrop@ucl.ac.uk

RMS
22nd January 2014

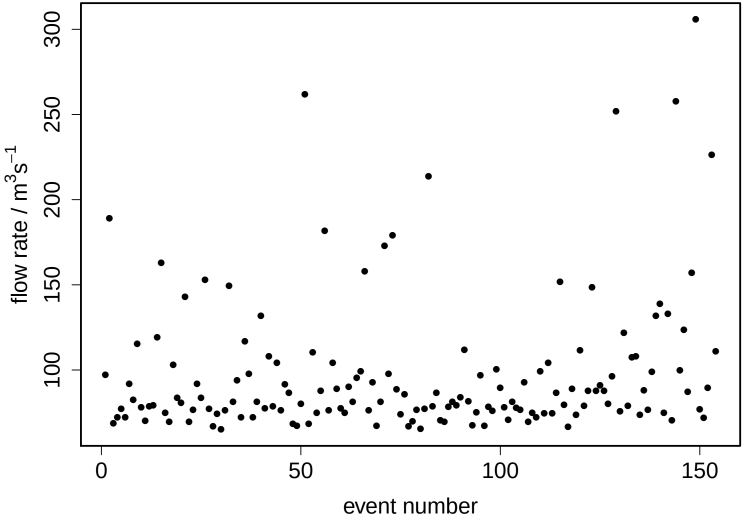
Outline

- Simple example dataset
- Generalized Pareto (GP) modelling of threshold excesses
- Threshold selection methods
  - parameter stability plots
  - others
- Improved parameter stability plots
- (Automatic?) threshold selection on many datasets

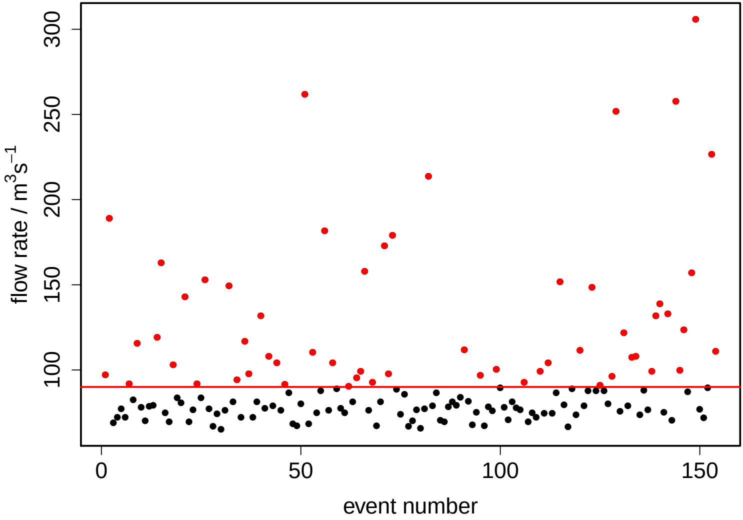
Motivating (classic) example:

- 154 flow rates from the River Nidd (Yorkshire), 1934–1969
- peaks flows: $\approx$ independent, pre-processing has already extracted extreme values

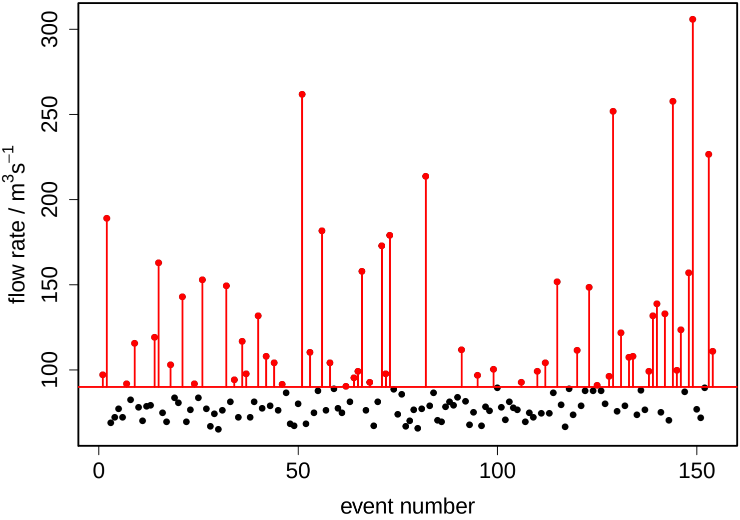
River Nidd 1934-1969: 154 flow rates



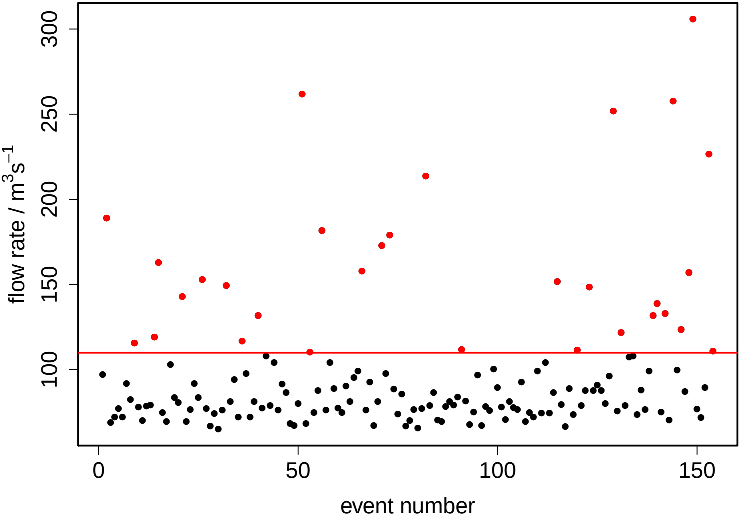
Threshold $u = 90$



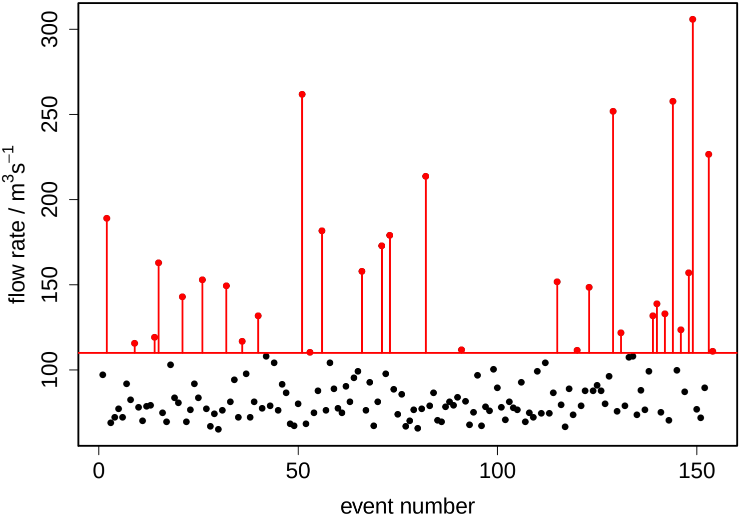
Excesses of $u = 90$



Threshold $u = 110$



Excesses of $u = 110$



Generalized Pareto model

- Set up: $X_1, X_2, \ldots X_n$ are i.i.d.
- $Y = (X - u) \mid X > u$: excess of threshold u
- Extreme value theory suggests the GP distribution as a model for Y.

GP distribution function:

$$P(Y \leqslant y) = G(y) = \begin{cases} 1 - (1 + \xi y/\sigma_u)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y/\sigma_u), & \xi = 0, \end{cases} \tag{1}$$

where $y > 0$, $x_+ = \max(x, 0)$.

- For $\xi \geqslant 0$ we have $y > 0$;
- For $\xi < 0$ we have $0 < y < -\sigma_u/\xi$.

(…how often the threshold is exceeded also matters)

Threshold selection methods

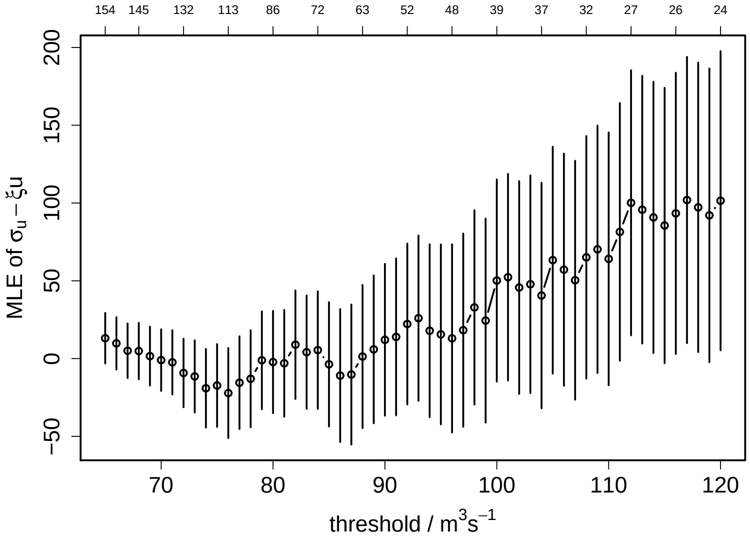
Bias-variance trade-off:

- u too low: GP model inappropriate $\to$ bias.
- u too high: fewer excesses $\to$ unnecessary imprecision.

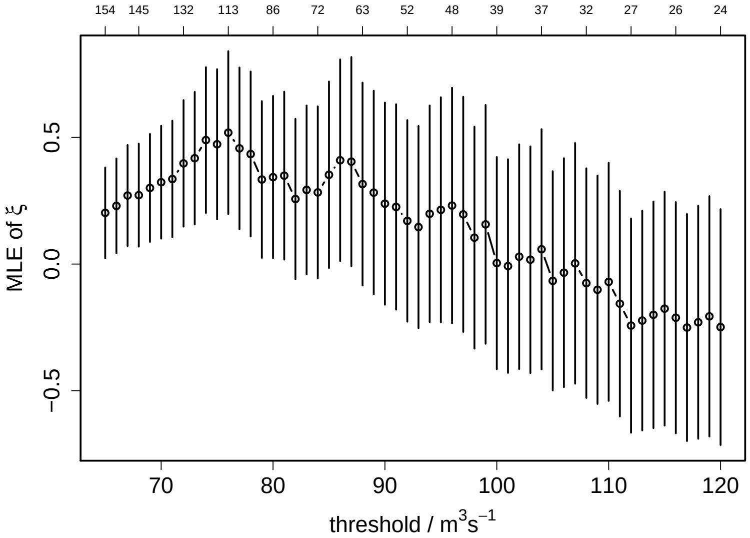
Review paper: Scarrott and MacDonald (2012).

- Parameter stability plot : $\widehat{\sigma}_u - \widehat{\xi} u$ vs. u and $\widehat{\xi}$ vs. u. Estimates stable above u^*?
- Mean residual life plot : sample mean excess vs. u. Linear above u^*?
- Goodness-of-fit test: AD or KS p-value vs. u. For which u don't we reject GP model?
- Extend model below u, make u a model parameter, make inferences about u, e.g. Wadsworth and Tawn (2012).

Parameter stability : modified scale



Parameter stability : shape



Problems with parameter stability plots

- $\widehat{\xi}$ and $\widehat{\sigma}_u - \widehat{\xi}u$ very strongly negatively associated across u: only one plot needed.
- Estimates of ξ based on thresholds u_1 and u_2 are depdendent (one datasets a subset of the other).
- viewer compares many pairs of thresholds (multiple-testing).
- viewer invited to ask whether CIs overlap: not the appropriate assessment.
- threshold choice rather subjective.

Aim: make assessment more formally, by testing (a discrete version of)

$$H_0 : \xi(x) = \xi(u), \quad \text{for } x > u.$$

Two-threshold GP model (cf. WT2012)

- Penultimate theory: GP model valid for lowish thresholds, shape parameter ξ varies slowly with threshold.
- Notation: ξ_i is the GP shape local to threshold u_i.
- Model ξ as piecewise constant in variable x:

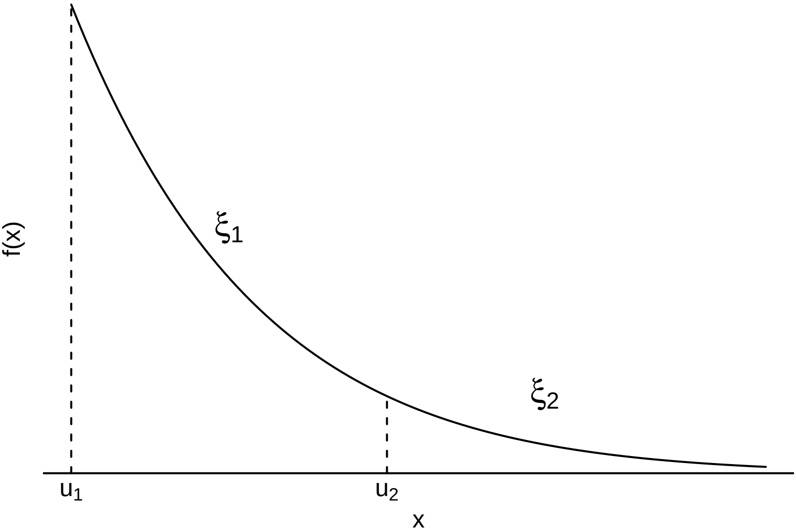
$$\xi(x) = \begin{cases} \xi_1, & u_1 < x < u_2, \\ \xi_2, & x > u_2. \end{cases} \tag{2}$$

Test $H_0 : \xi_i = \xi_j$ for all possible pairs (u_i, u_j) from a set of thresholds $(u_1, \ldots, u_m)$.

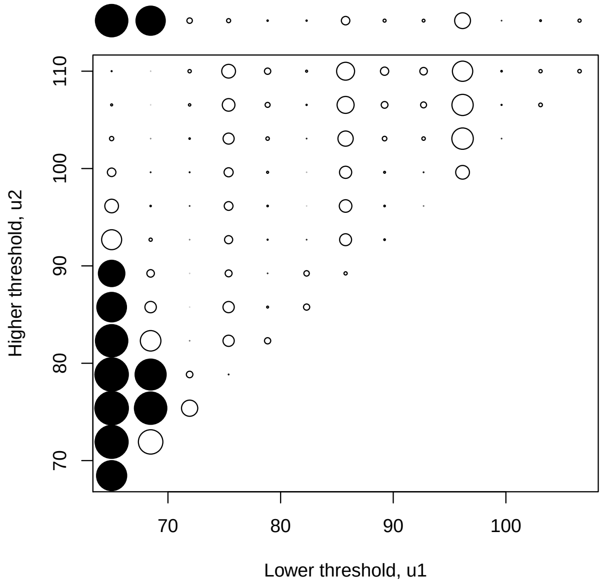
Drawbacks:

- Simulation required to test $H_0 : \xi_1 = \cdots = \xi_m$ by combining pairwise tests.
- Very computationally-intensive : prohibitively so.

Two-threshold model



Plot of pairwise test results



Multiple-threshold GP model (NC2014)

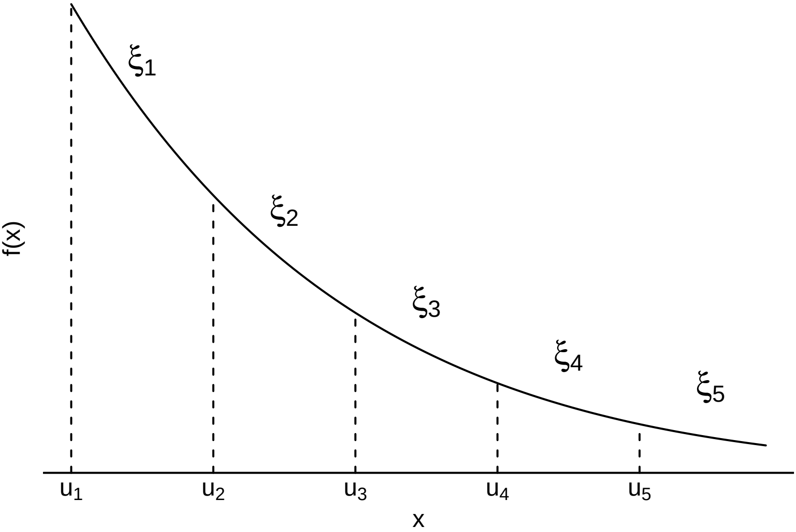
$$\xi(x) = \begin{cases} \xi_i, & u_i < x < u_{i+1}, \quad \text{for } i = 1, \ldots, m-1, \\ \xi_m, & x > u_m. \end{cases} \tag{3}$$

$$H_0 : \xi_1 = \cdots = \xi_m \qquad \text{vs} \qquad H_A : H_0 \text{ not true .}$$

- Scale parameters set to achieve a continuous p.d.f.
- Parameter vector: $\theta = (\sigma_1, \xi_1, \ldots, \xi_m)$
- $\widehat{\theta}_0$: MLE under H_0.
- $\widehat{\theta}$: MLE under H_A.

General idea: do the data suggest a departure from H_0 (standard GP model) in the direction of m-threshold GP model.

m-threshold model ($m = 5$)



Likelihood ratio and score tests

LR stat. : compare maximized log-likelihoods under H_0 and H_A

$$W = 2\left\{l(\widehat{\theta}) - l(\widehat{\theta}_0)\right\},$$

Score stat. : how far is the log-likelihood from being flat at $\widehat{\theta}_0$?

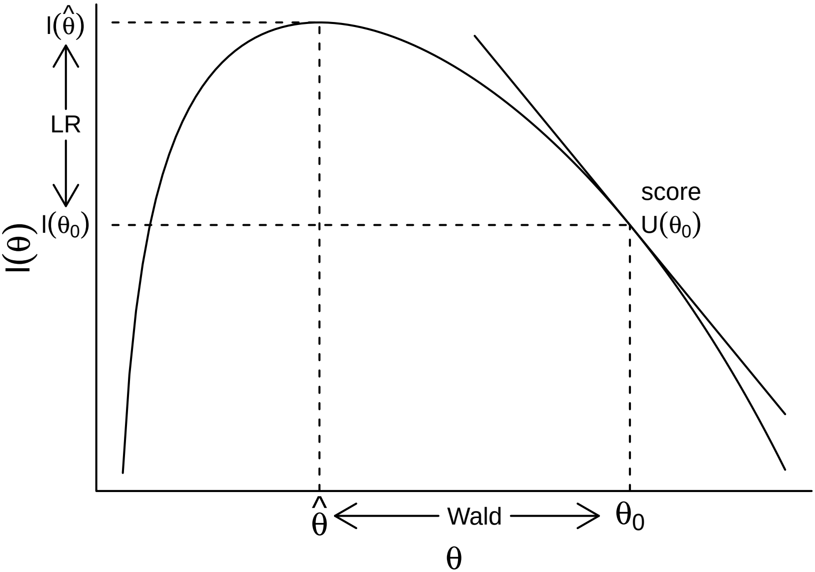
$$S = U(\widehat{\theta}_0)^T\, i^{-1}(\widehat{\theta}_0)\, U(\widehat{\theta}_0),$$

- $U(\theta)$ is the score function: $U_i = \partial l(\theta)/\partial\theta_i$
- $i(\theta)$ is the expected Fisher information matrix: $i(\theta)_{ij} = \mathrm{E}\left[-\partial^2 l(\theta)/\partial\theta_i\partial\theta_j\right]$.
- Derivation of $U(\theta)$ and $i(\theta)$ nasty: it made my head hurt!

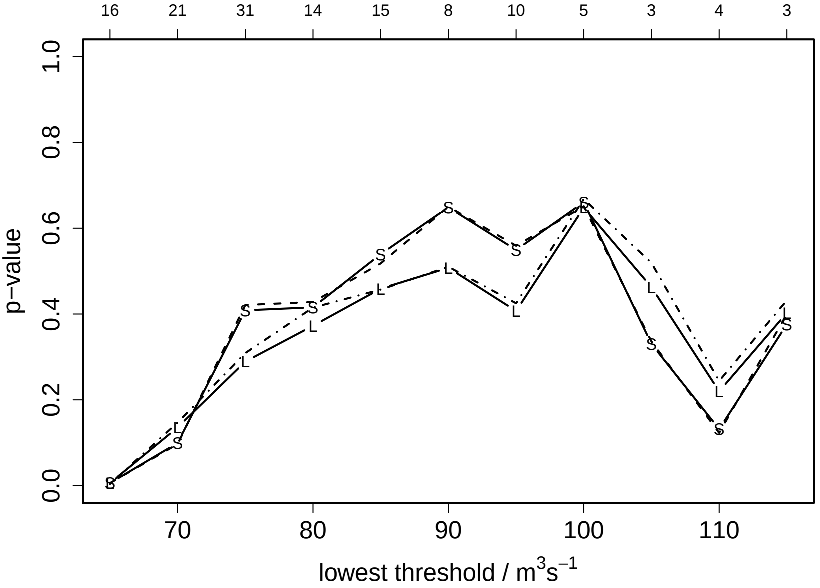
If $\xi_m > -1/2$ then W and S are approx. χ^2_{m-1} under H_0.

LLR test. Need to fit the full multiple-threshold GP model.
Score test. Only fit the null model, i.e. a single GP(σ, ξ) fit.

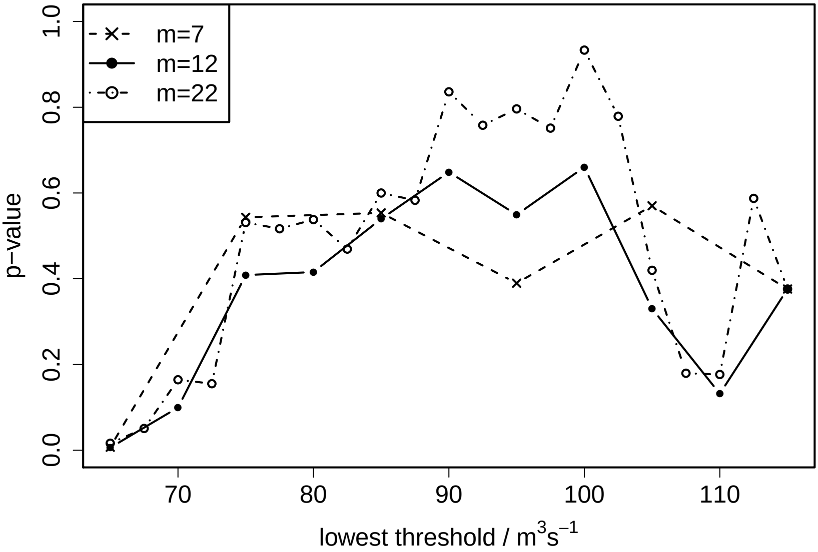
LR, score and Wald test statistics



Multiple threshold diagnostic plot



Effect of number of thresholds



Automatic threshold selection

How to make use of the *p*-values?
Consider $H_0 : \xi_i = \cdots = \xi_m$.

- Small *p*-value suggests that H_0 isn't true, i.e. u_i isn't high enough.
- Large *p*-value : perhaps u_i is high enough.

Possibilities

- Formal: set size of test beforehand, e.g. 5%. Reject u_i as too low if *p*-value is < 0.05.
- Informal: view the *p*-values as a measure of the disagreement between the data and the null hypothesis when inspecting plot.

Some multiple-testing remains: we perform tests with lowest thresholds $u_1, u_2, \ldots, u_{m-1}$.

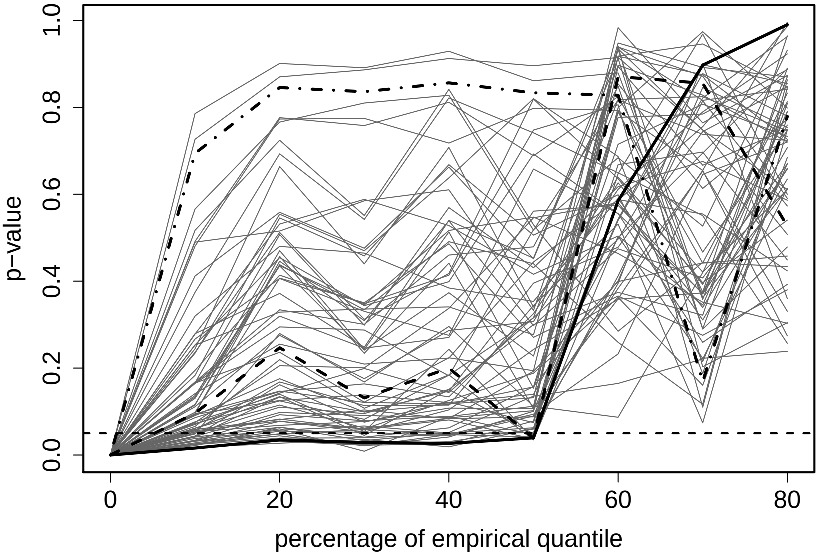
Wave heights from the Gulf of Mexico

- Hindcasts of H_s storm peak significant wave height (in metres) in the Gulf of Mexico.
  - Data from Northrop and Jonathan (2011).
  - wave height : trough to the crest of the wave.
  - significant wave height : the average of the largest 1/3 wave heights. A measure of sea surface roughness.
  - storm peak: largest value from each (hurricane-induced) storm.
- a 6×12 grid of 72 sites ($\approx$ 14 km apart).
- Sep 1900 to Sep 2005 : 315 storms .
- average of 3 observations (storms) per year, at each site.

Storm damage

- Interested in extremal behaviour of H_s at centre of data grid
- Pool (spatially-dependent) data over space.
- What level (quantile) of threshold is appropriate at site 1, site 2, . . . , site 72?

Gulf of Mexico wave heights



Automatic threshold selection

Suppose that we wish to automate the selection of a threshold for each of the datasets, based on tests of size 5%, say.

Two strategies

1. Select the lowest threshold for which H_0 is not rejected.

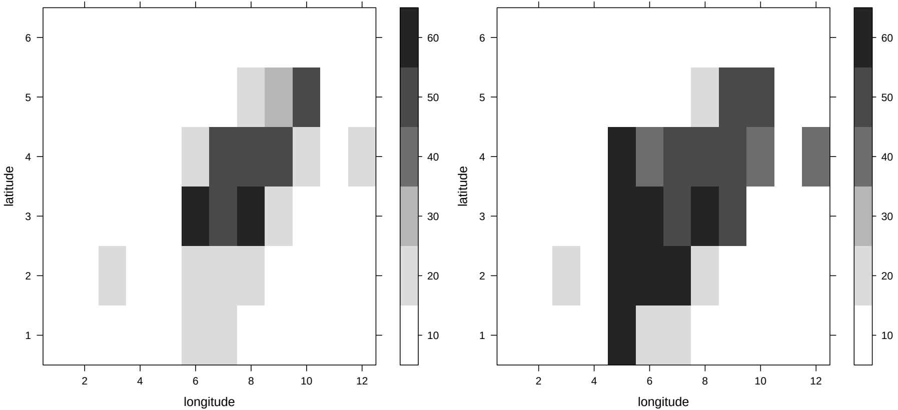
   [However, the p-values are not constrained to be non-decreasing in the lowest threshold.]

2. Select the lowest threshold with the property that H_0 is not rejected at it and at all the higher thresholds considered.

   [We need to bear in mind that large variability is expected at the very highest thresholds.]

Automatic threshold selection

Quantile at which threshold chosen



Strategy 1 Strategy 2

Concluding remarks

- Makes the (frequentist) fixed threshold selection part of Wadsworth and Tawn (2012) quick.
- Parameter stability is only part of the story (model checking).
- Threshold sensitivity vs. threshold uncertainty:
  - u a tuning parameter vs. u is a model parameter
  - sensitivity/uncertainty shouldn't be ignored.
- Motivated by slide 28 of Jo's talk:
  - Chavez-Demoulin *et al*'s discussion of Northrop and Jonathan (2011);
  - r-largest order statistics model for in a multi-site analysis, with (only) latitude and longitude as covariates;
  - ...no need to set explicitly a threshold u;
  - ...but what if there is a continuous covariate, such as "distance to nearest gate".
- Adjustment for serial dependence?

References

Northrop, P. J. and Coleman, C. L. Improved threshold diagnostic plots for extreme value analyses. *Extremes*. To appear.

Northrop PJ, Jonathan P (2011) Threshold modelling of spatially dependent nonstationary extremes with application to hurricane-induced wave heights. *Environmetrics* **22(7)**, 799809.

Scarrott C, MacDonald A (2012) A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT* - Statistical Journal **10(1)**, 3360.

Wadsworth, J. and J. Tawn (2012). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *Journal of the Royal Statistical Society - Series B: Statistical Methodology* **74 (3)**, 543–567.

Thank you for your attention.