Using Quantile Regression to Set Thresholds for Extreme Value Analyses

Paul Northrop
University College London
paul@stats.ucl.ac.uk

Royal Statistical Society
29th May 2013

With thanks to Philip Jonathan (Shell Research) and Nicolas Attalides (UCL)

Outline

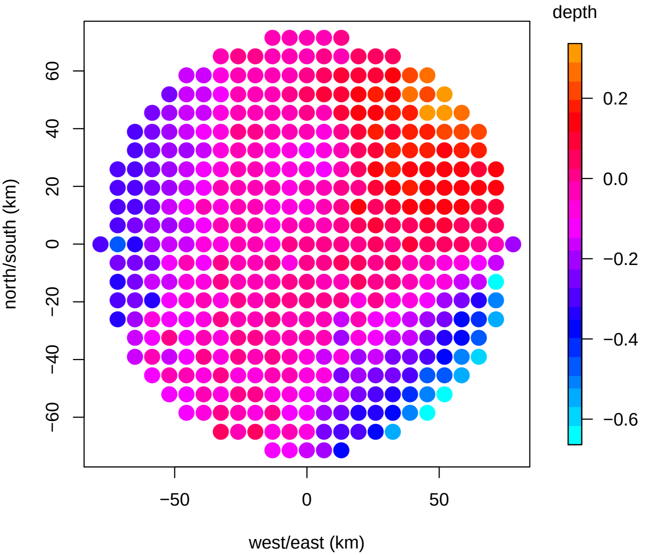
1. Wave height data $\rightarrow$ design of safe marine structures.
2. Threshold-based extreme value modelling.
3. Quantile regression $\rightarrow$ thresholds for extreme value regression models.
4. Wave height data.

Wave heights (at an unnamed location)

- Hindcasts of storm peak significant wave height (Y).
  - wave height : trough to the crest of the wave;
  - significant wave height : the average of the largest third of wave heights. A measure of sea surface roughness;
  - storm peak: largest value from each 'storm' identified;
  - assume storms are approximately independent.
- 427 sites : within $\approx$ 80km of site of interest.
- 1970 – 2007 : 76 storms .
- Storms occur between November and May.
- $\approx$ 2 storms per year, at each site, on average.
- Potential covariates: water depth, longitude, latitude.

For confidentiality Y has been scaled to $[0, 100]$.

(Scaled) water depth and location



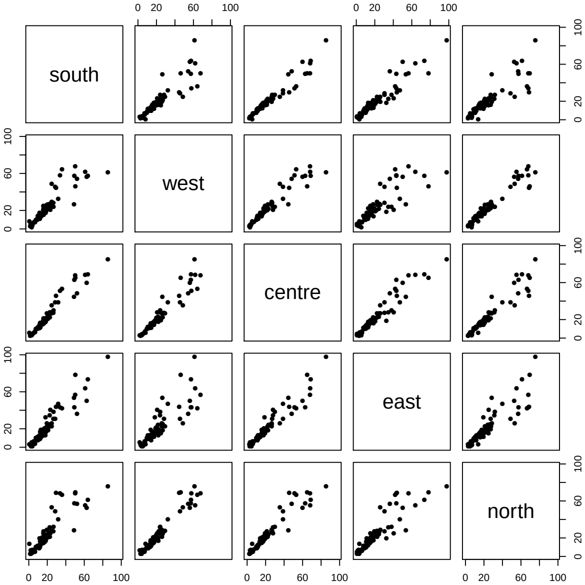
Storm damage

Storm damage

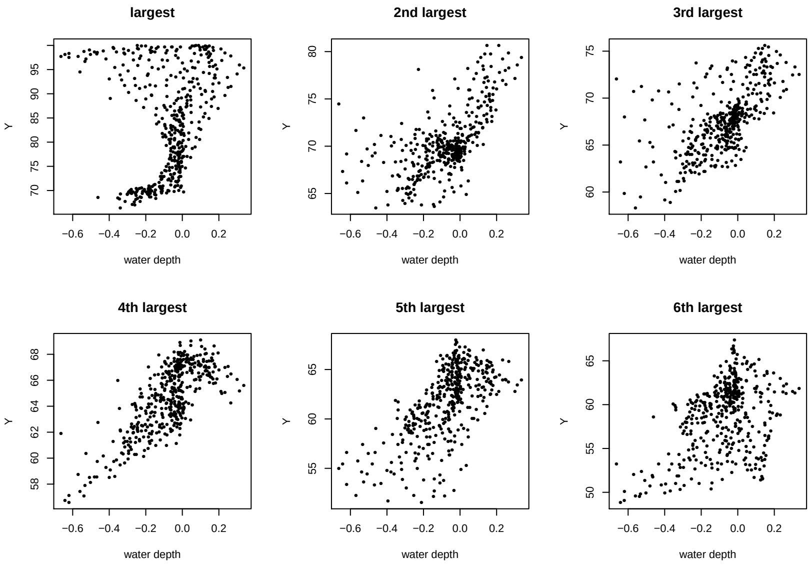
At the centre of the grid of data (where scaled water depth = 0)

- How large will significant wave heights be in the next 100 years? . . . or the next 1000 years?
- Estimate extreme quantiles (upper tail).
- Issues: pooling of spatially-dependent data over space; effect of water depth; extrapolation.

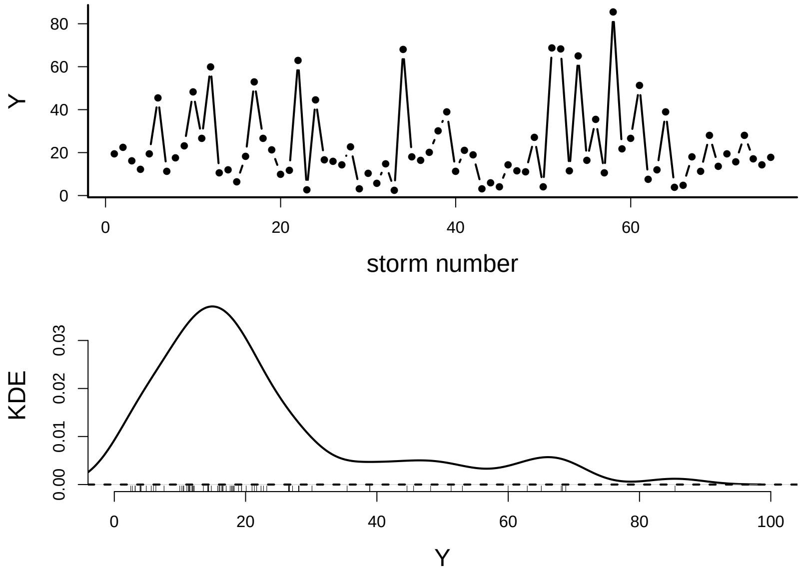
Spatial dependence in sig. wave height



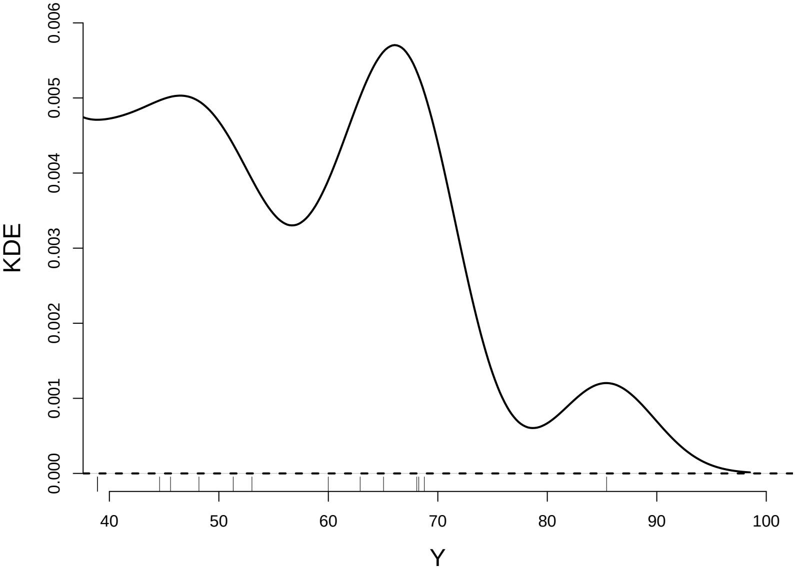
Effect of water depth



Storm peaks at centre of grid



Zoom in on upper tail



Possibilities:

- Fit a model to all the data. Extrapolate from this model.

  ...but (unless the link between typical and atypical behaviour is well-understood) inferences about extremes could be influenced adversely by the modelling of non-extreme data.

Possibilities:

- Fit a model to all the data. Extrapolate from this model.

  ...but (unless the link between typical and atypical behaviour is well-understood) inferences about extremes could be influenced adversely by the modelling of non-extreme data.

- Base inferences about future extreme behaviour on extreme data:
  - block maxima;
  - r-largest order statistics in a block;
  - exceedances of a high threshold, u.

Block maxima

Assume that $Y_1, Y_2, \ldots$ are i.i.d..

Let $M_N = \max(Y_1, \ldots, Y_N)$.

- Any possible (non-degenerate) distribution of $Z_N = (M_N - b_N)/a_N$ as $N \to \infty$ is in the GEV (Generalised Extreme Value) family, with c.d.f.

$$P(Z_N \leqslant z) = \exp\left\{ - \left[1 + \xi \left(\frac{z - \mu_N}{\sigma_N} \right) \right]_+^{-1/\xi} \right\},$$

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

Assume that $Y_1, Y_2, \ldots$ are i.i.d..

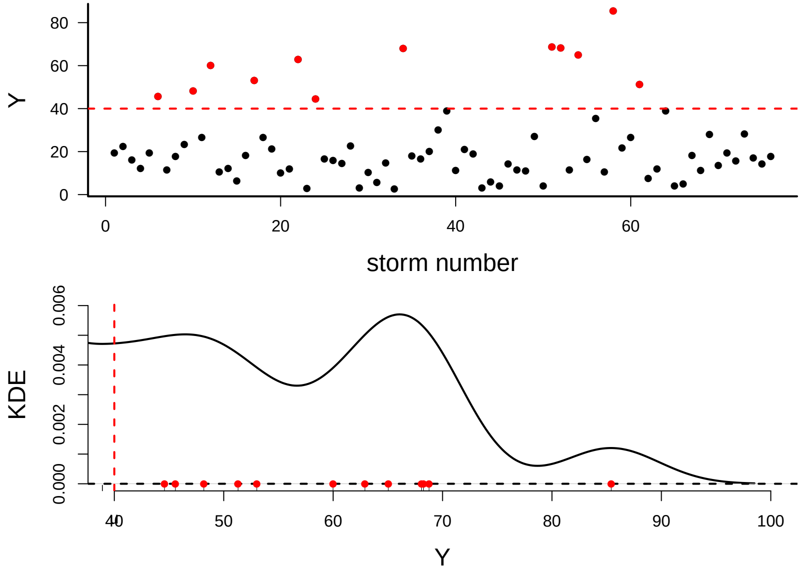
Let $M_N = \max(Y_1, \ldots, Y_N)$.

- Any possible (non-degenerate) distribution of $Z_N = (M_N - b_N)/a_N$ as $N \to \infty$ is in the GEV (Generalised Extreme Value) family, with c.d.f.

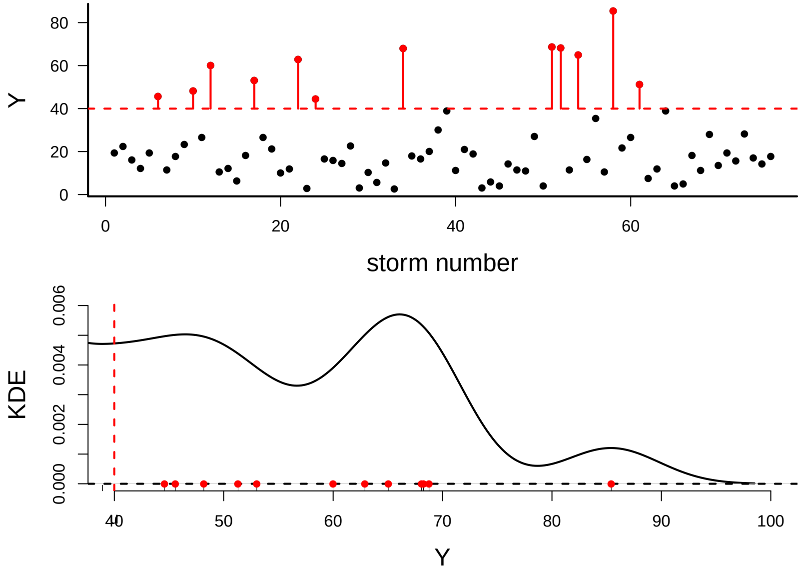
$$P(Z_N \leqslant z) = \exp\left\{-\left[1 + \xi\left(\frac{z - \mu_N}{\sigma_N}\right)\right]_+^{-1/\xi}\right\},$$

  where $x_+ = \max(x, 0)$ and $\sigma > 0$.
- Suggests GEV(μ_N, σ_N, ξ) as a model for M_N for large N.
- Upper end point is finite for $\xi < 0$ and infinite for $\xi \geqslant 0$.
- Related asymptotic model for r-largest order statistics.

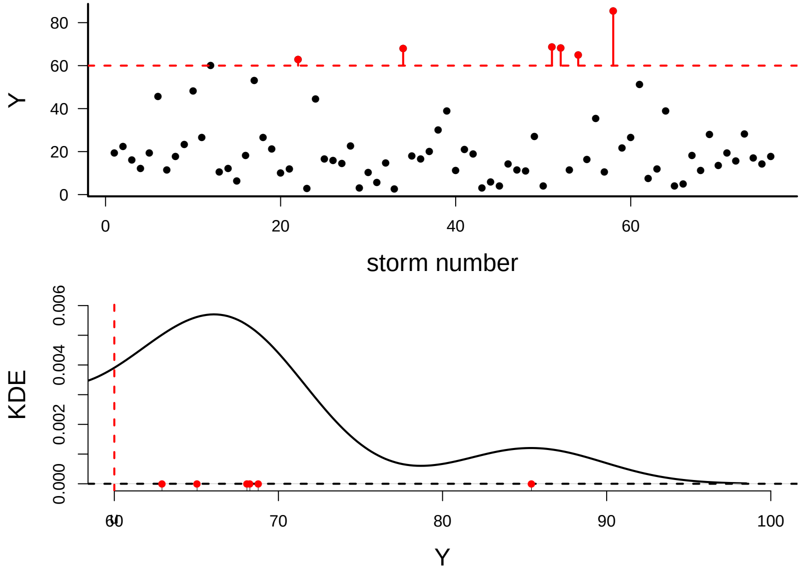
Threshold exceedances, u=40



Threshold excesses, u=40



Threshold excesses, u=60



Threshold exceedances I : binomial-GP

- How often is a (high) threshold u exceeded?

  Let $p_u = P(Y > u)$.

Threshold exceedances I : binomial-GP

- How often is a (high) threshold u exceeded?

  Let $p_u = P(Y > u)$.

- Given that u is exceeded, by how much is it exceeded?

  Any possible distribution of $(Y - u) \mid Y > u$ as $u \to \infty$ is in the Generalised Pareto (GP) family, with conditional c.d.f.

$$P(Y \leqslant u \mid Y > u) = 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)_+^{-1/\xi},$$

  where $\sigma_u > 0$.

Threshold exceedances I : binomial-GP

- How often is a (high) threshold u exceeded?

  Let $p_u = P(Y > u)$.

- Given that u is exceeded, by how much is it exceeded?

  Any possible distribution of $(Y - u) \mid Y > u$ as $u \to \infty$ is in the Generalised Pareto (GP) family, with conditional c.d.f.

$$P(Y \leqslant u \mid Y > u) = 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)_+^{-1/\xi},$$

  where $\sigma_u > 0$.

If $Y_1, Y_2, \ldots, Y_n$ are independent then this suggests a binomial-GP model, for sufficiently high u.

Threshold exceedances II : NHPP

- We rescale time (storm number) to (0,1) and let $M(t_1, t_2, u)$ be the number of observations in $[t_1, t_2] \times (u, \infty)$.
- Asymptotic arguments lead to a 2D non-homogeneous Poisson process NHPP(μ_N, σ_N, ξ) model, s.t.

$$M(t_1, t_2, u) \sim \text{Poisson}\left(\frac{n}{N}(t_2 - t_1)\left[1 + \xi\left(\frac{u - \mu_N}{\sigma_N}\right)\right]_+^{-1/\xi}\right).$$

- Here, we choose $N = n = 76$ so that (μ_N, σ_N, ξ) relate to maximum on dataset.

Informally ...

Threshold exceedances II : NHPP

- We rescale time (storm number) to (0,1) and let $M(t_1, t_2, u)$ be the number of observations in $[t_1, t_2] \times (u, \infty)$.
- Asymptotic arguments lead to a 2D non-homogeneous Poisson process NHPP(μ_N, σ_N, ξ) model, s.t.

$$M(t_1, t_2, u) \sim \text{Poisson}\left(\frac{n}{N}(t_2 - t_1)\left[1 + \xi\left(\frac{u - \mu_N}{\sigma_N}\right)\right]_+^{-1/\xi}\right).$$

- Here, we choose $N = n = 76$ so that (μ_N, σ_N, ξ) relate to maximum on dataset.

Informally ...

1. Reparameterise: $(p_u, \sigma_u, \xi) \to (\mu_N, \sigma_N, \xi)$, using

$$\sigma_u = \sigma_N + \xi(u - \mu_N);$$
$$p_u \approx \frac{1}{N}\left[1 + \xi\left(\frac{u - \mu_N}{\sigma_N}\right)\right]^{-1/\xi}.$$

- We rescale time (storm number) to (0,1) and let $M(t_1, t_2, u)$ be the number of observations in $[t_1, t_2] \times (u, \infty)$.
- Asymptotic arguments lead to a 2D non-homogeneous Poisson process NHPP(μ_N, σ_N, ξ) model, s.t.

$$M(t_1, t_2, u) \sim \text{Poisson}\left(\frac{n}{N}(t_2 - t_1)\left[1 + \xi\left(\frac{u - \mu_N}{\sigma_N}\right)\right]_+^{-1/\xi}\right).$$

- Here, we choose $N = n = 76$ so that (μ_N, σ_N, ξ) relate to maximum on dataset.

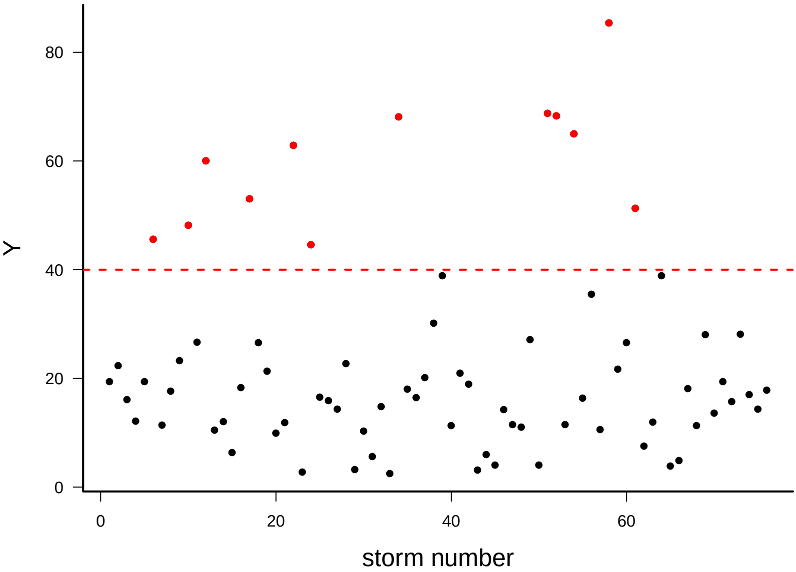
Informally ...

1. Reparameterise: $(p_u, \sigma_u, \xi) \to (\mu_N, \sigma_N, \xi)$, using

$$\sigma_u = \sigma_N + \xi(u - \mu_N);$$
$$p_u \approx \frac{1}{N}\left[1 + \xi\left(\frac{u - \mu_N}{\sigma_N}\right)\right]^{-1/\xi}.$$

2. Poisson $\approx$ binomial, for large n, small p_u.

NHPP on $(60, \infty)$



Extreme value regression modelling

What if we have covariate effects?

- Appeal to standard theory conditional on the covariates.
- Specify that extreme value parameters, e.g. μ_N, σ_N, ξ are functions of the value of a covariate x, e.g.

$$\text{NHPP}(\mu_N(x), \sigma_N(x), \xi(x)).$$

Extreme value regression modelling

What if we have covariate effects?

- Appeal to standard theory conditional on the covariates.
- Specify that extreme value parameters, e.g. μ_N, σ_N, ξ are functions of the value of a covariate x, e.g.

$$\text{NHPP}(\mu_N(x), \sigma_N(x), \xi(x)).$$

- The PP model has the advantage (over the bin-GP model) that its parameters are invariant to u.

…but does a constant threshold still make sense?

Covariate-dependent thresholds

Arguments for:

- Asymptotic justification : the threshold $u(x)$ needs to be high for each x.

Covariate-dependent thresholds

Arguments for:

- Asymptotic justification : the threshold $u(x)$ needs to be high for each x.
- Design : spread exceedances across a wide range of covariate values.

Covariate-dependent thresholds

Arguments for:

- Asymptotic justification : the threshold $u(x)$ needs to be high for each x.
- Design : spread exceedances across a wide range of covariate values.
- Parsimony : simpler model than with a constant threshold (Eastoe and Tawn, 2009).

Covariate-dependent thresholds

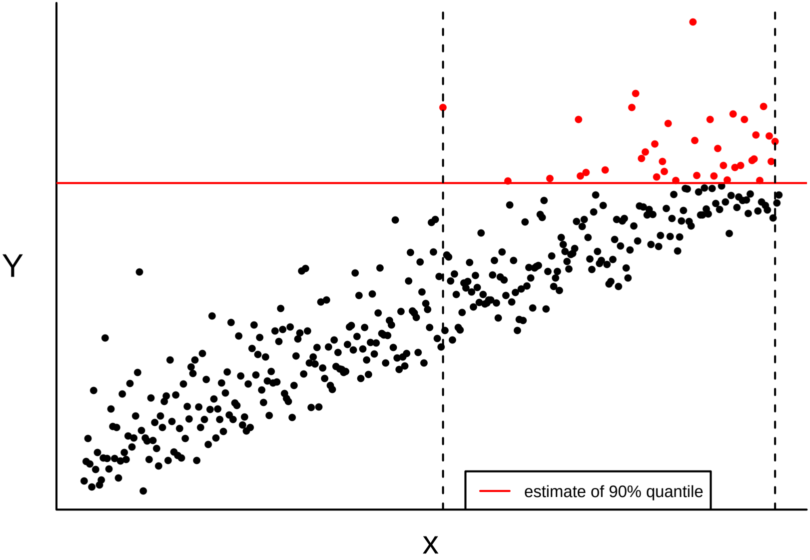
Arguments for:

- Asymptotic justification : the threshold $u(x)$ needs to be high for each x.
- Design : spread exceedances across a wide range of covariate values.
- Parsimony : simpler model than with a constant threshold (Eastoe and Tawn, 2009).

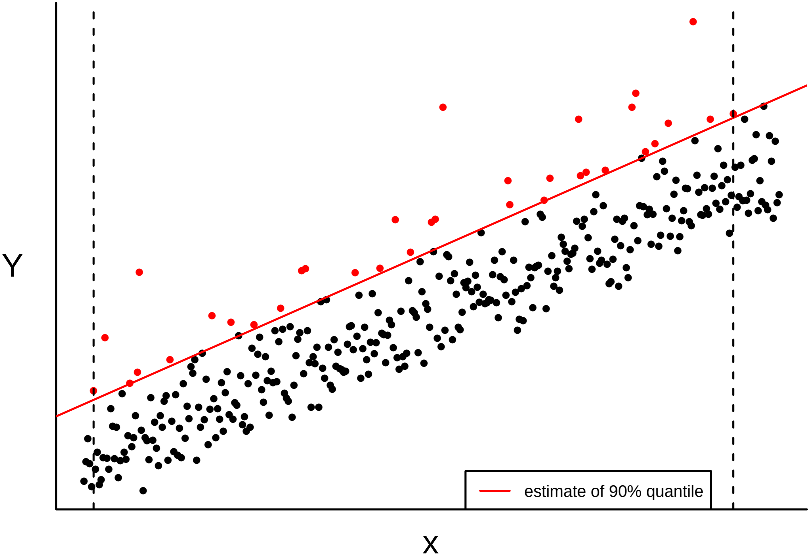
Set $u(x)$ so that $p_u(x) = P(Y > u(x))$, is approx. constant for all x.

- Set $u(x)$ by trial-and-error or by discretising x, e.g. different threshold for different locations, months etc.
- Quantile regression (QR) : model quantiles of Y as a function of covariates.

Constant threshold



Quantile regression



QR and NHPP model parameterisation

Let $p_u(x) = P(Y > u(x))$. Then, if $\xi(x) = \xi$ is constant,

$$p_u(x) \approx \frac{1}{N}\left[1 + \xi\left(\frac{u(x) - \mu(x)}{\sigma(x)}\right)\right]^{-1/\xi}.$$

If $p_u(x) = p_u$ is constant then

$$u(x) = \mu(x) + c\,\sigma(x).$$

QR and NHPP model parameterisation

Let $p_u(x) = P(Y > u(x))$. Then, if $\xi(x) = \xi$ is constant,

$$p_u(x) \approx \frac{1}{N}\left[1 + \xi\left(\frac{u(x) - \mu(x)}{\sigma(x)}\right)\right]^{-1/\xi}.$$

If $p_u(x) = p_u$ is constant then

$$u(x) = \mu(x) + c\,\sigma(x).$$

The form of $u(x)$ is determined by the extreme value model:

- if $\mu(x)$ and/or $\sigma(x)$ are linear in x: linear QR ;
- if $\log \mu(x)$ and/or $\log \sigma(x)$ is linear in x: non-linear QR .

Return to significant wave height data

- Fit NHPP regression model using maximum likelihood: model effects on EV parameters as simple functions of (scaled) water depth x , e.g.

$$\mu(x) = \mu_0 + \mu_1 x.$$

- Assume $\xi > -1/2$ for regularity.
- Threshold : use quantile regression to achieve approx. constant probability p_u of threshold exceedance over space:

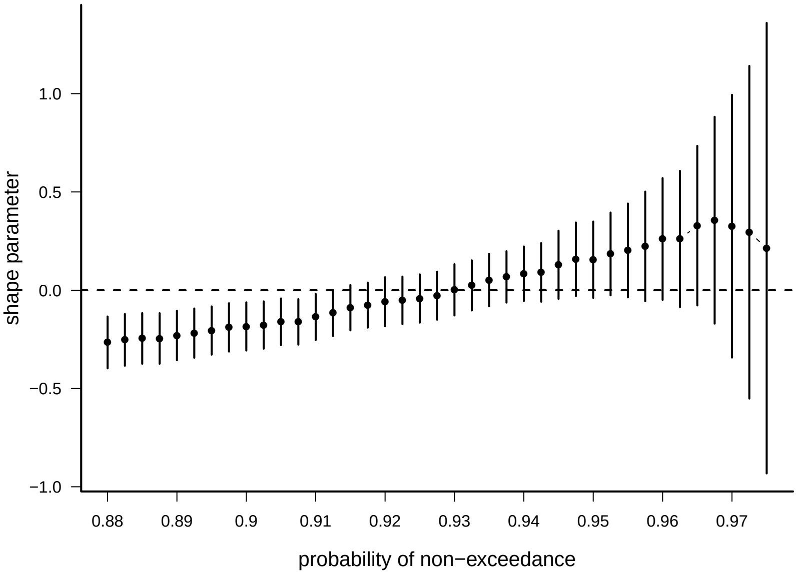
  Model $100(1 - p_u)\%$ quantile as a function of covariates.

- Spatial dependence .
  - Not modelled explicitly: interest in one location only.
  - Initially, assume conditional independence of responses given covariate values.
  - Adjust standard errors etc. for spatial dependence.
  - ...Chandler and Bate (2007): scale log-likelihood so that Hessian at MLE matches “sandwich” estimator of var(MLE).

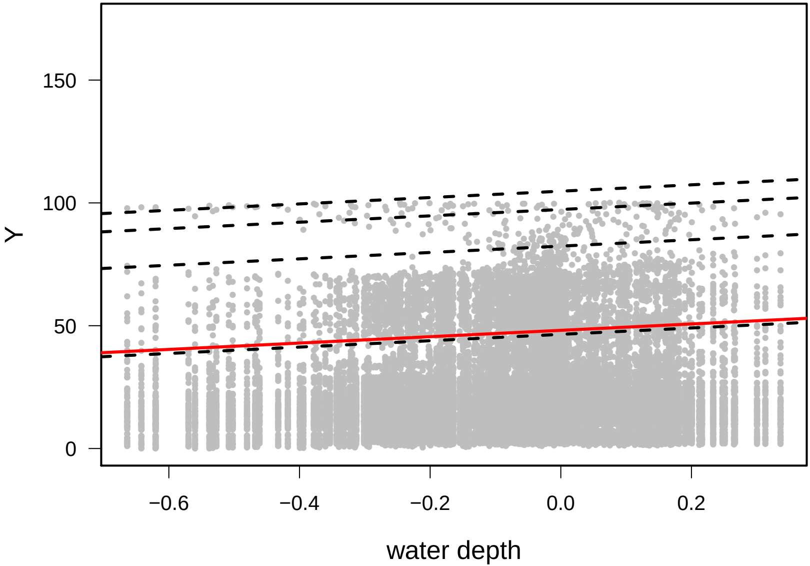
Threshold selection

- Threshold level (determined by p_u): bias-variance trade-off.
- Iterative: form of threshold depends on NHPP covariate model.
- For given EV model set threshold using appropriate QR model.
- Treat QR threshold as fixed: simulation study (Northrop and Jonathan, 2011) suggests effect of ignoring uncertainty is minimal.
- Choice of exceedance probability p_u: look for stability in parameter estimates.
- Final model: μ linear in water depth, σ and ξ constant.

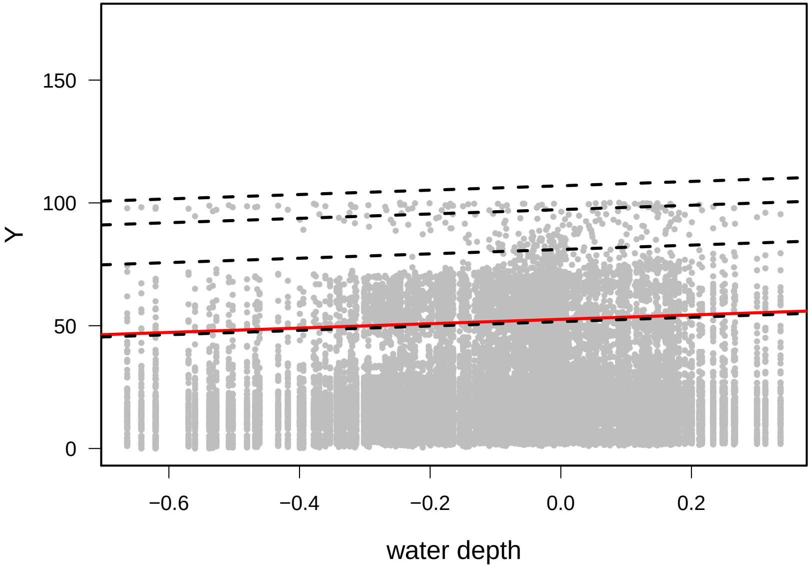
MLEs of shape parameter ξ vs. $1 - p$.



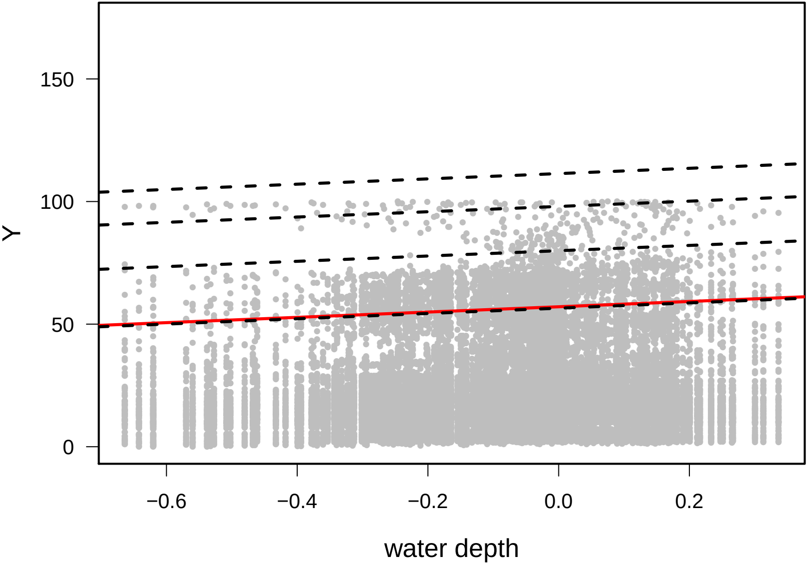
87% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



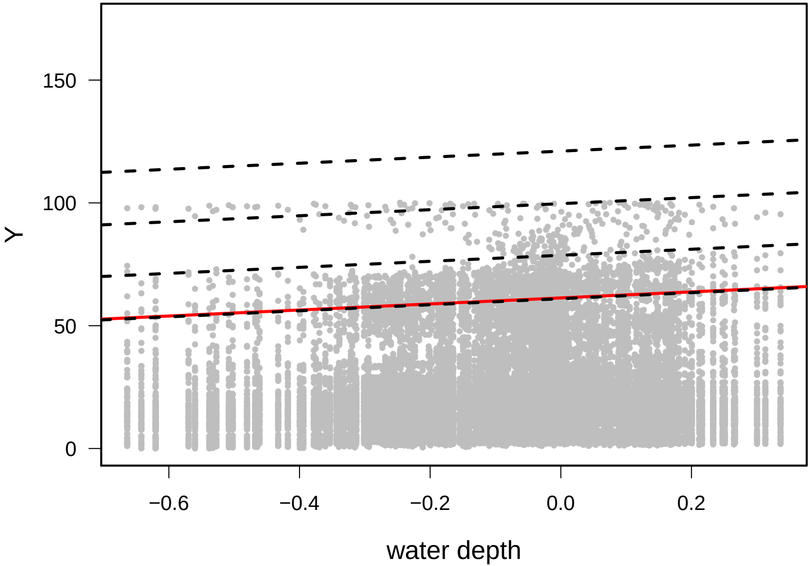
89% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



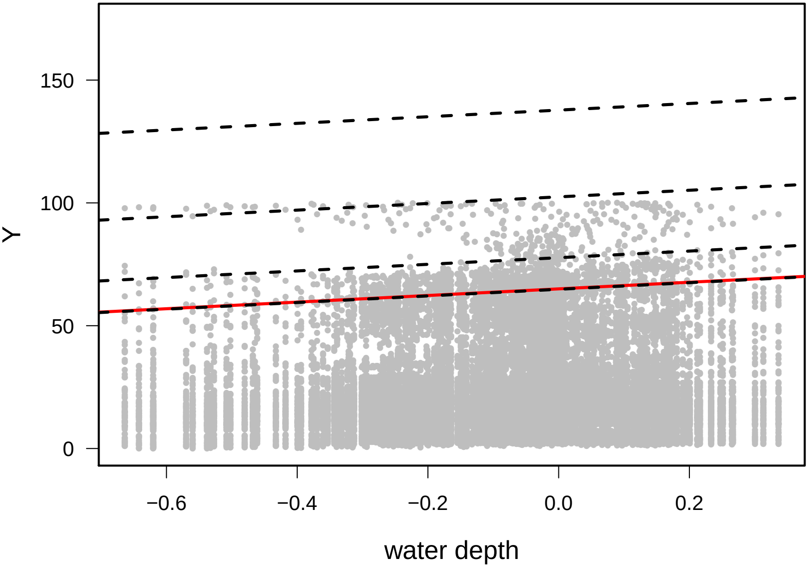
91% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



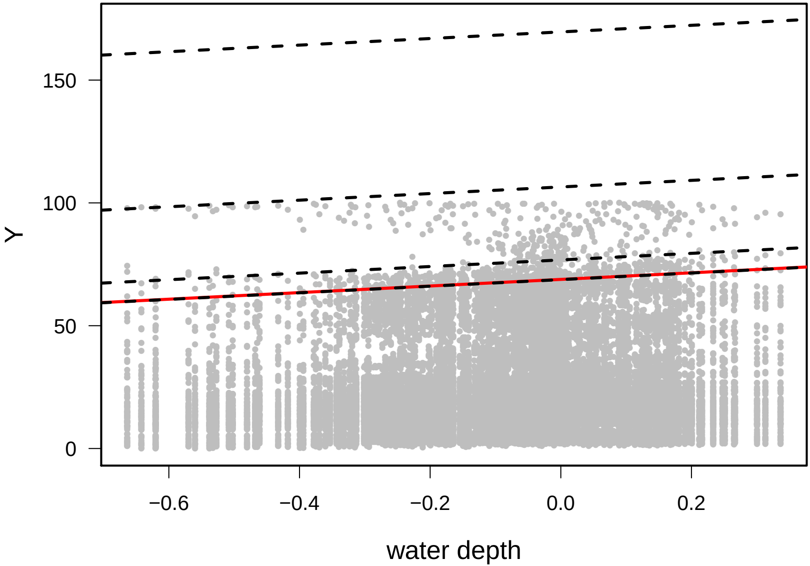
93% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



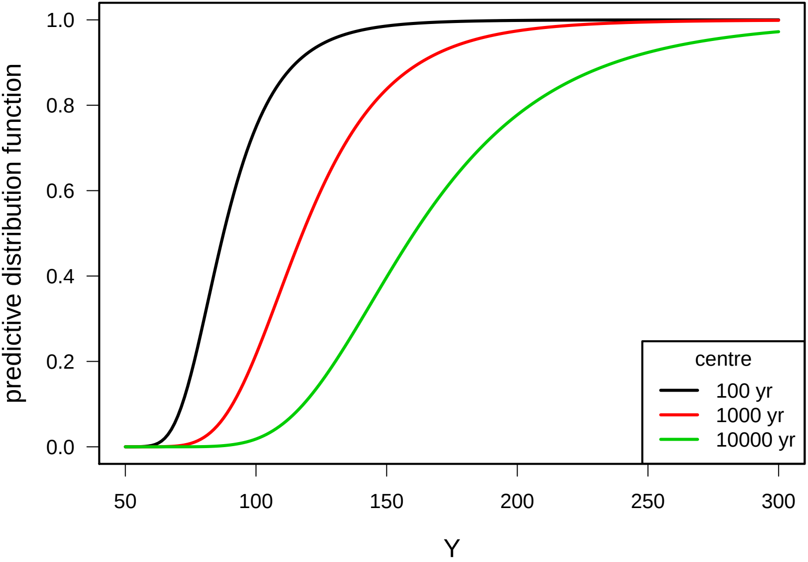
95% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



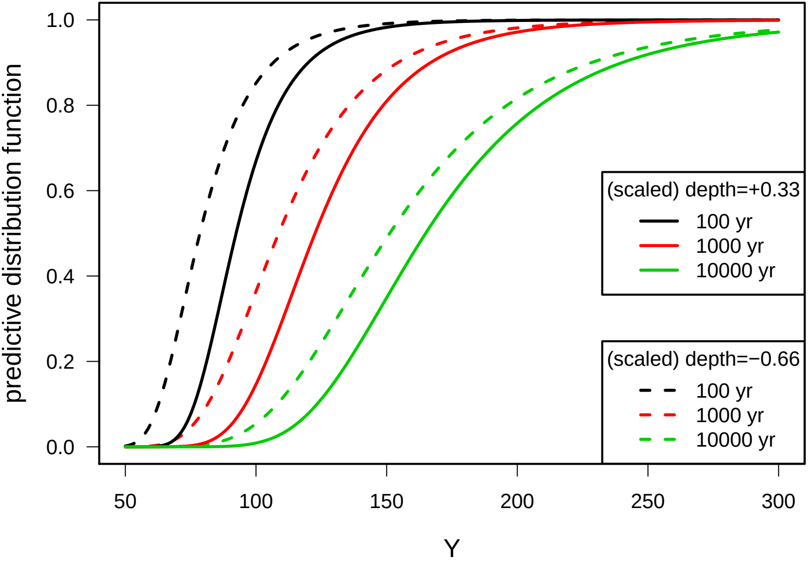
97% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



Approx. predictive DF of *m*-year maximum



Approx. predictive DF of *m*-year maximum



Quantile regression

- A simple and effective strategy to set thresholds for non-stationary EV models.
- Theoretical work (Nicolas Attalides):
  - Suppose that $\mu(x_1, \ldots, x_q) = \mu_0 + \sum_{i=1}^{q} \mu_i x_i$.
  - If each of the q covariates are distributed symmetically then a QR-threshold minimizes the generalised asymptotic variance of $(\widehat{\mu}_1, \ldots, \widehat{\mu}_q)$.
  - (…but this doesn't address bias).

Concluding remarks

Quantile regression

- A simple and effective strategy to set thresholds for non-stationary EV models.
- Theoretical work (Nicolas Attalides):
  - Suppose that $\mu(x_1, \ldots, x_q) = \mu_0 + \sum_{i=1}^{q} \mu_i x_i$.
  - If each of the q covariates are distributed symmetically then a QR-threshold minimizes the generalised asymptotic variance of $(\widehat{\mu}_1, \ldots, \widehat{\mu}_q)$.
  - (…but this doesn't address bias).

Ongoing work

- Address bias-variance tradeoff.
- Threshold sensitivity/uncertainty. Extend Wadsworth and Tawn (2012) to regression siutation.

References

Chandler, R. E. and Bate, S. B. (2007) Inference for clustered data using the independence loglikelihood. *Biometrika* **94 (1)**, 167–183.

Eastoe, E. F. and Tawn J. A. 2009. Modelling non-stationary extremes with application to surface level ozone. *Applied Statistics* **58 (1)**, 2545.

Koenker R. 2009. *Quantreg: Quantile Regression. R Package Version 4.44.* http://CRAN.R-project.org/package=quantreg

Northrop, P. J. and Jonathan, P. Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights. Published online in *Environmetrics*. **22 (7)**, 799–809 (with discussion.)

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.