Marginal modelling of spatially-dependent non-stationary extremes

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

Workshop EVT2013 in honour of Ivette Gomes
10th September 2013

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

Outline

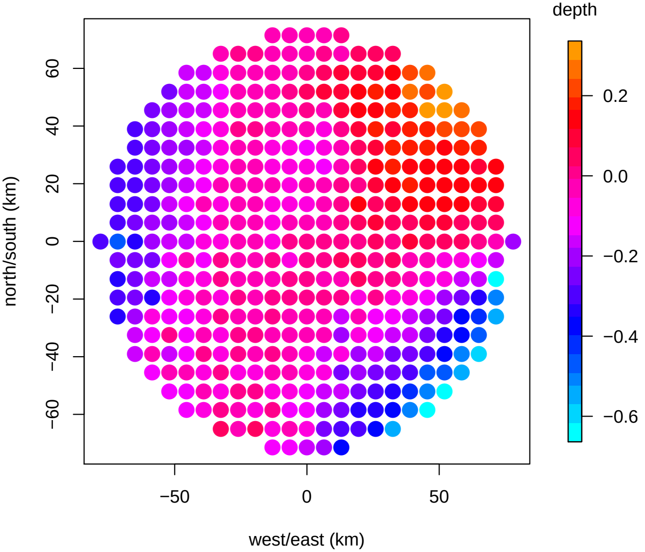
1. Wave height data $\rightarrow$ design of safe marine structures.
2. Threshold-based extreme value modelling.
3. Quantile regression $\rightarrow$ thresholds for extreme value regression models.
4. Adjustment for spatial dependence.
5. Wave height data.

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 (over 3 hour period).
  - storm peak: largest value from each 'storm' identified;
  - assume storms are approximately independent (cf. declustering).
- 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 1,000 years? . . . or 10,000 years?
- Estimate extreme quantiles (upper tail).
- Issues: pooling of spatially-dependent data over space; effect of water depth; extrapolation.

Comments

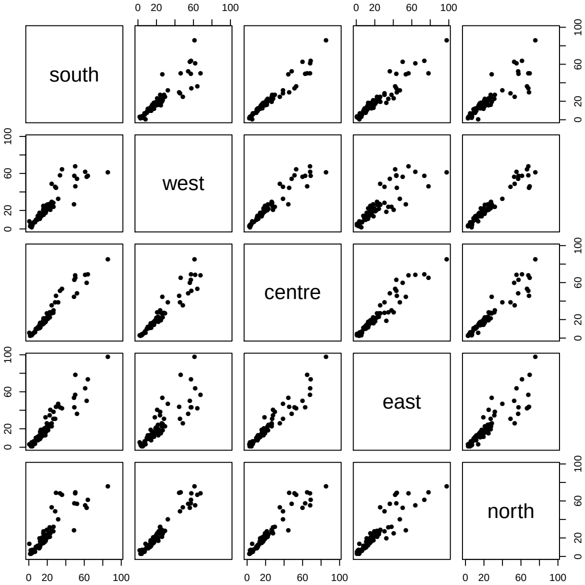
Measurement issues:

- Field measurement uncertainty greatest for extreme values; long, homogeneous records in places of interest are rare;
- Hindcast data are simulations based on pragmatic physics, calibrated to historical observation.

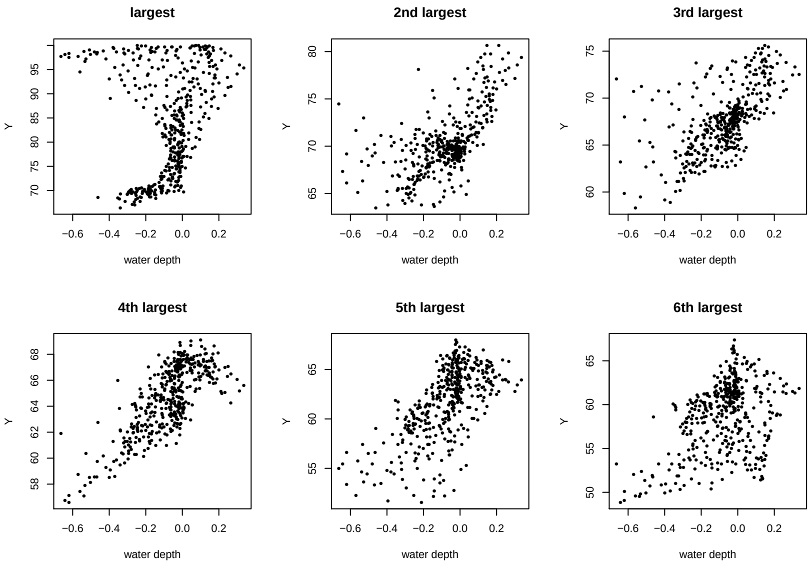
In-house analyses:

- desire to pool data over space (precise track of 'wave-causing events' is somewhat random);
- ...but don't account for strong spatial dependence;
- compromise: pool over a small number (5, say) of non-contiguous sites ;
- potential covariate effects ignored.

Spatial dependence in sig. wave height



Effect of water depth



NHPP model

- Consider responses $Y_1, \dots, Y_n$ from a single site.
- Rescale time (storm number) to (0,1) and let $M(t_1, t_2, u)$ be the number of observations in $[t_1, t_2] \times (u, \infty)$.
- 2D non-homogeneous Poisson process NHPP(μ_N, σ_N, ξ) model, s.t.

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

- μ_N, σ_N and ξ are GEV parameters of $\max(Y_1, \dots, Y_N)$.
- Here, we choose $N = n = 76$ so that (μ_N, σ_N, ξ) relate to maximum on dataset.

Extreme value regression modelling

What if anticipate have covariate effects, e.g. water depth?

- 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 anticipate have covariate effects, e.g. water depth?

- 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

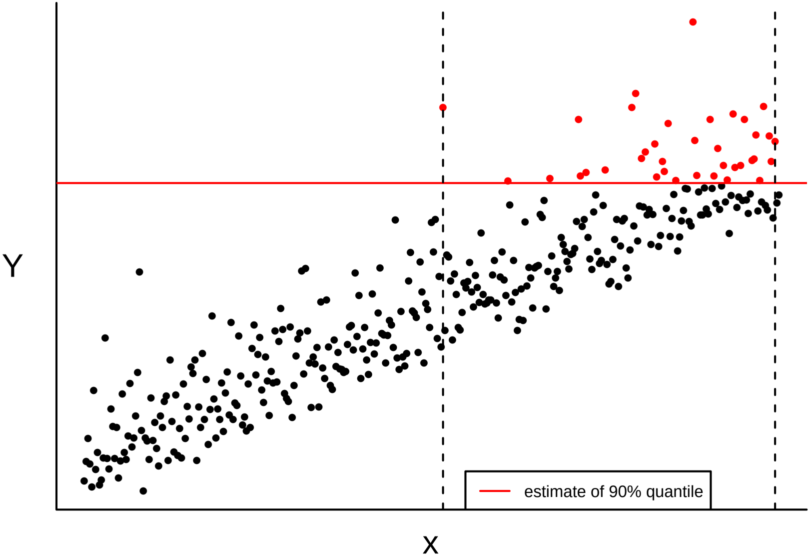
Arguments for:

- 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).

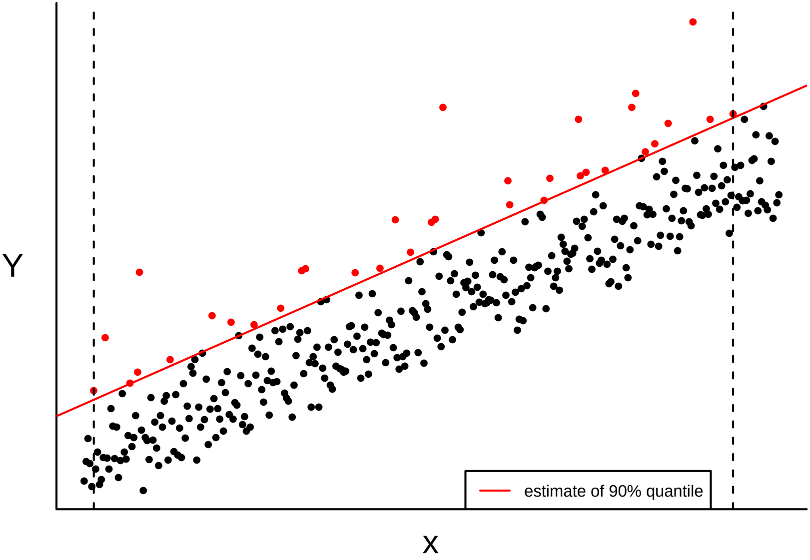
Set $u(x)$ so that $p_u(x) = P(Y > u(x))$, is approx. constant for all x.
[Model $u(x)$ for fixed p, rather than $p_u(x)$ for fixed u.]

- 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 .

Adjustment for spatial dependence

Independence log-likelihood:

$$l_{IND}(\phi) = \sum_{j=1}^{k} \sum_{i=1}^{427} \log f_{ij}(y_{ij}; \phi, x_i) = \sum_{j=1}^{k} l_j(\phi).$$

(storms) (space)

In regular problems, as $k \to \infty$,

$$\widehat{\phi} \to N(\phi_0, H^{-1}\, V\, H^{-1}),$$

- $H =$ expected Hessian: $\mathrm{E}\left(\frac{\partial^2}{\partial \phi^2}\, l_{IND}(\phi_0)\right)$;
- $V = \mathrm{var}\left(\frac{\partial}{\partial \phi}\, l_{IND}(\phi)\right)$

Adjustment of $l_{IND}(\phi)$

- $\widehat{H}$ = observed Hessian, at $\widehat{\phi}$;
- $\widehat{V} = \sum_{j=1}^{k} U_j\left(\widehat{\phi}\right)^T U_j\left(\widehat{\phi}\right), \qquad U_j(\phi) = \frac{\partial l_j(\phi)}{\partial \phi}.$

Let $\widehat{H}_A = \left(-\widehat{H}^{-1}\,\widehat{V}\,\widehat{H}^{-1}\right)^{-1}$ ("sandwich" estimator of var($\widehat{\phi}$)).

Chandler and Bate (2007):

$$l_{ADJ}(\phi) = l_{IND}(\widehat{\phi}) + \frac{(\phi-\widehat{\phi})'\,\widehat{H}_A\,(\phi-\widehat{\phi})}{(\phi-\widehat{\phi})'\,\widehat{H}\,(\phi-\widehat{\phi})}\left(l_{IND}(\phi) - l_{IND}(\widehat{\phi})\right),$$

- Adjust $l_{IND}(\phi)$ so that its Hessian is H_A at $\widehat{\phi}$ rather than H.
- Preserves the usual asymptotic distribution of the likelihood ratio statistic.

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:

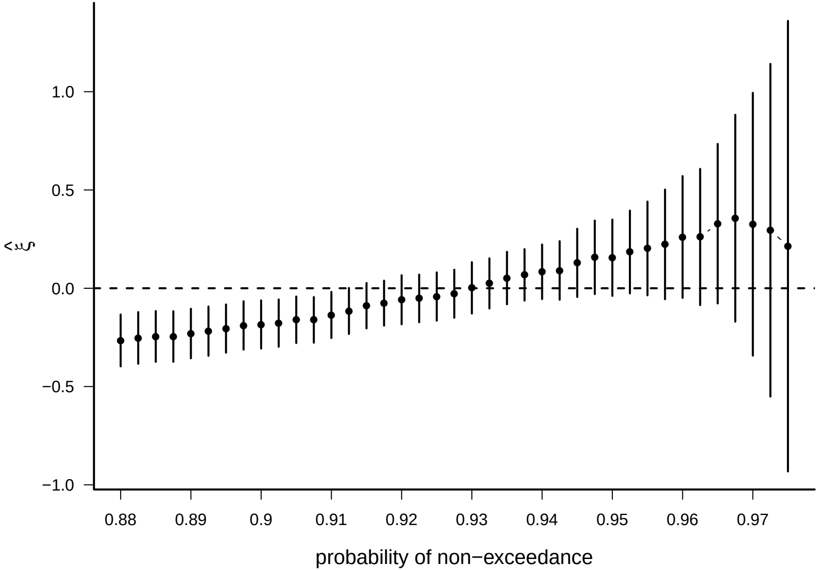
  Model $100(1 - p_u)\%$ quantile as a function of covariates.

- Spatial dependence .
  - Not modelled explicitly: interest in one location only.
  - Use Chandler-Bate adjustment of independence log-likelihood.

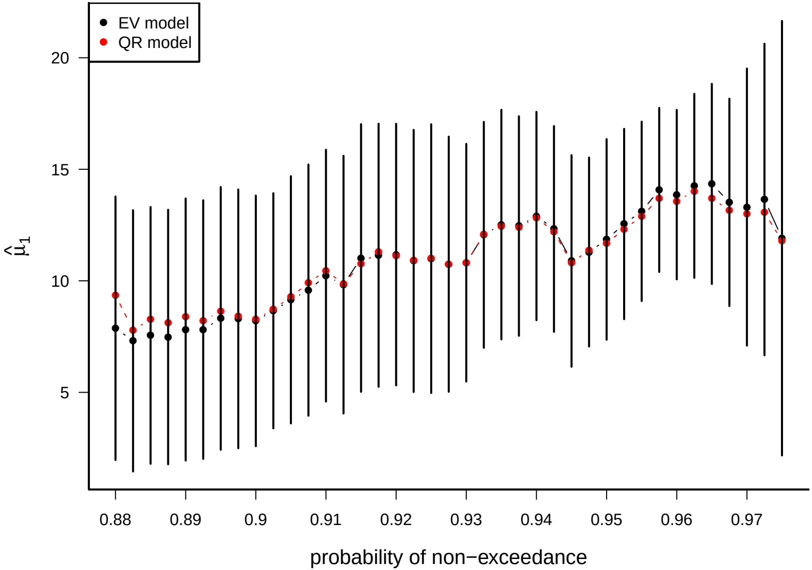
Threshold selection

- 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. [No significant spatial effects.]

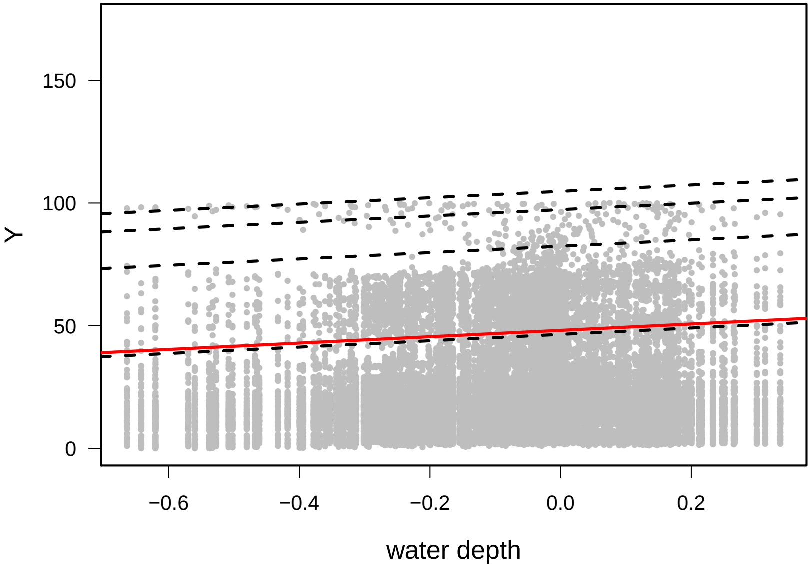
MLEs of shape parameter ξ vs. $1 - p$.



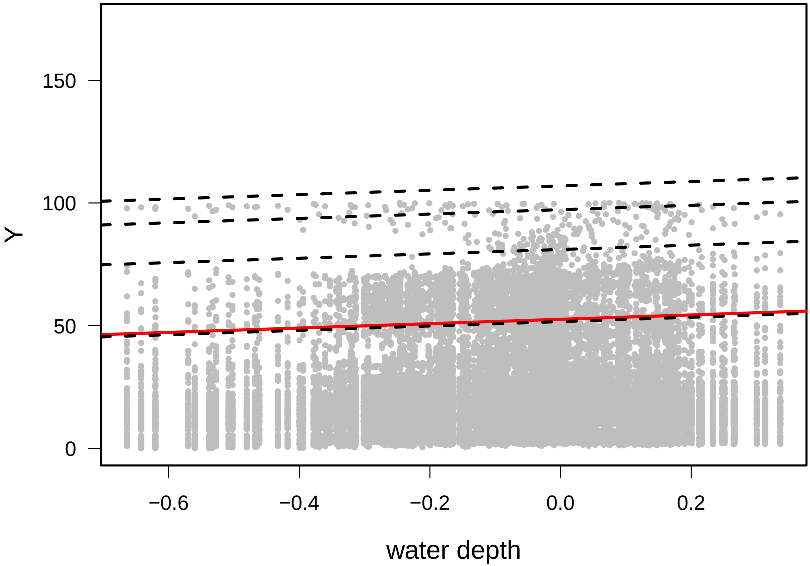
MLEs of coeff. of depth μ_1 vs. $1 - p$.



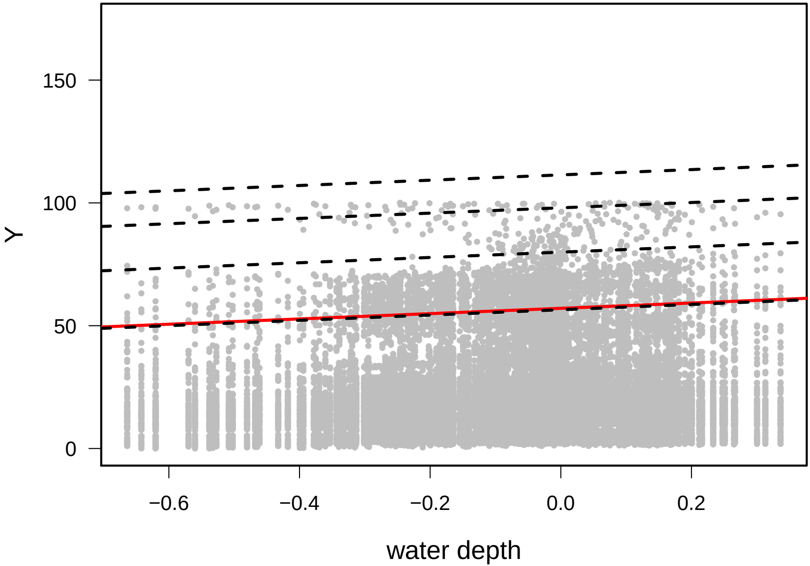
87% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



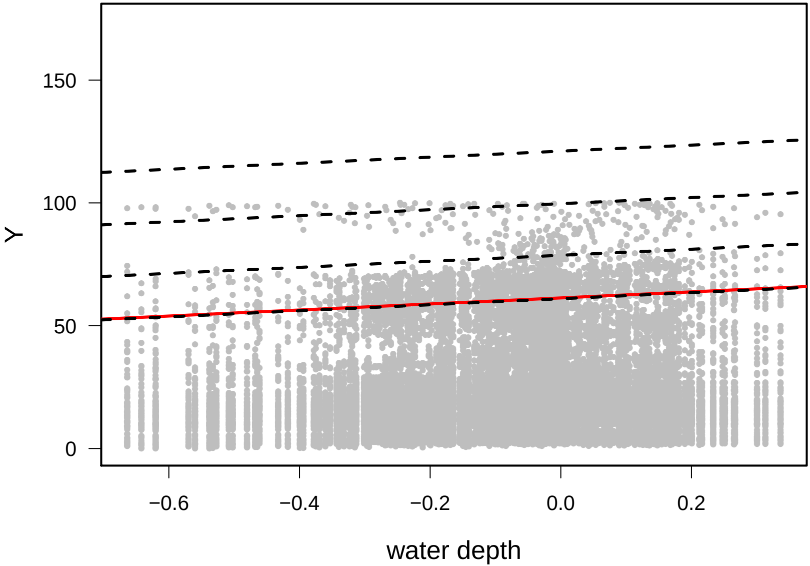
89% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



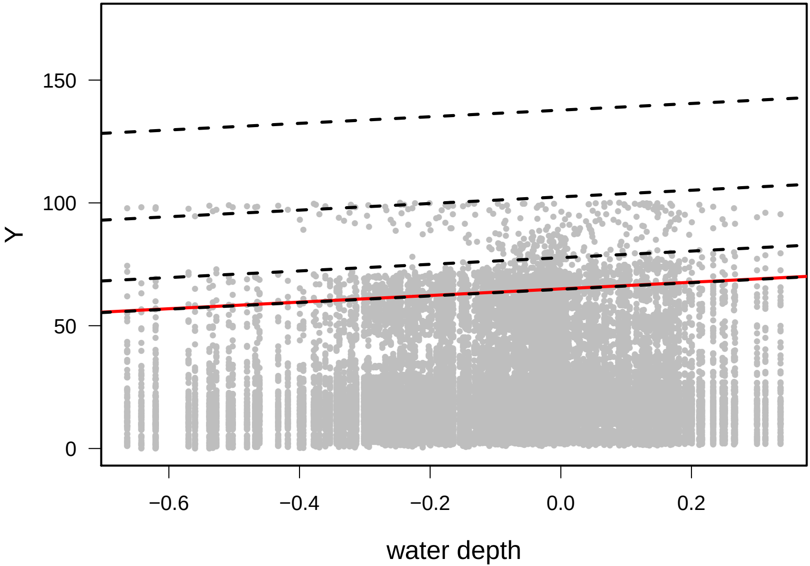
91% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



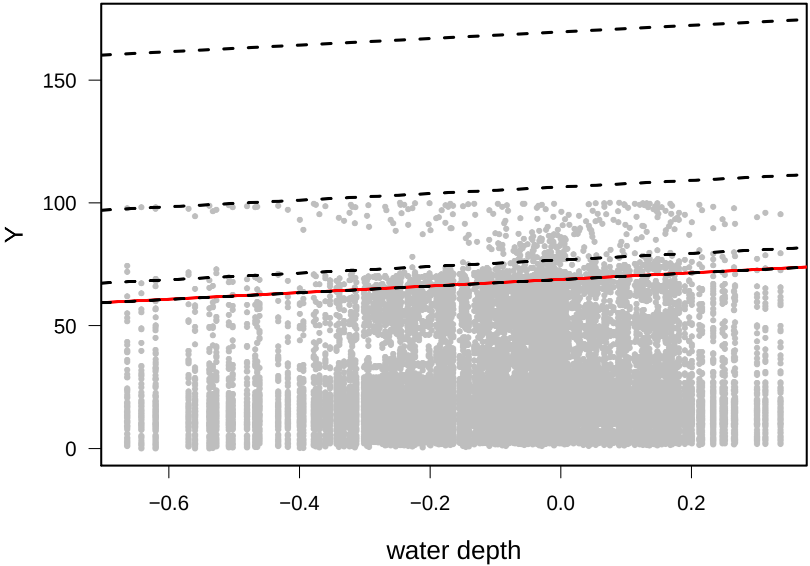
93% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



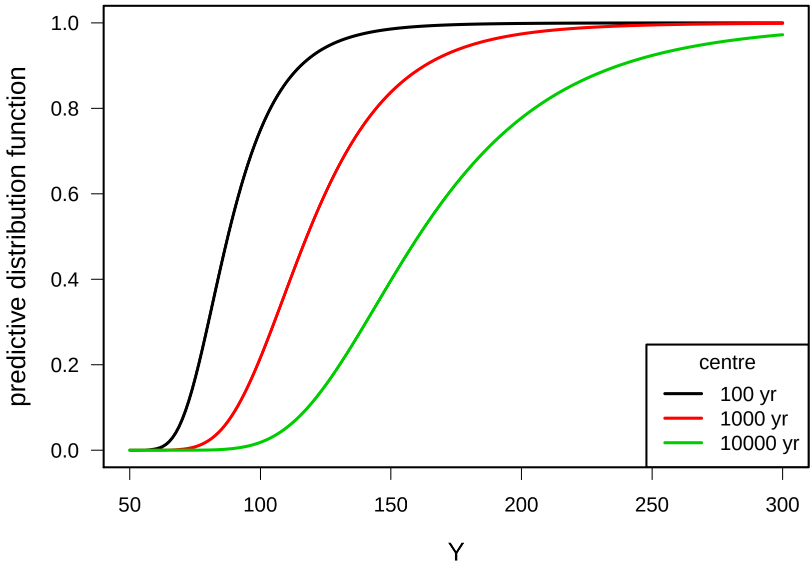
95% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



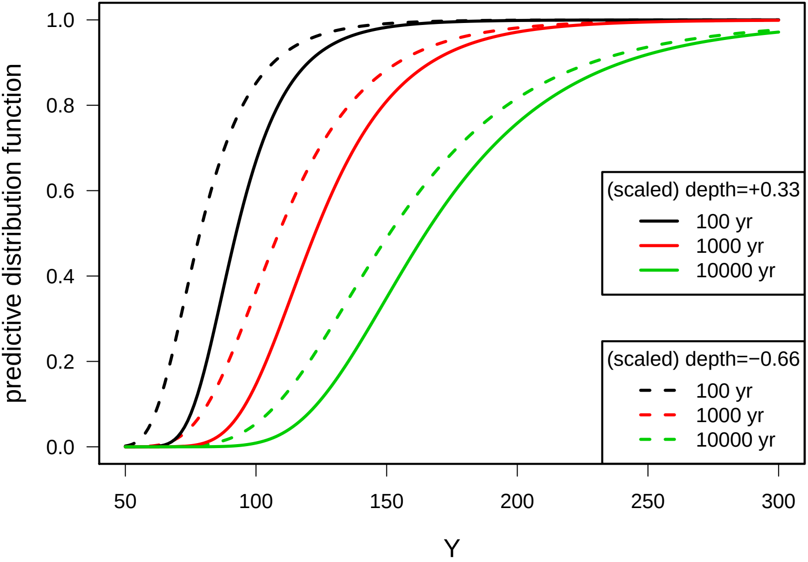
97% $u(x)$ and NHPP: 99%, 99.9%, 99.99%



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.
- Simpler $r-$largest order statistics approach (Chavez-Demoulin et al., 2011) for discrete covariates;
- 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.
- Simpler $r-$largest order statistics approach (Chavez-Demoulin et al., 2011) for discrete covariates;
- 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 situation.

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.)

Chavez-Demoulin, V., Davison, A. C. and Frossard, L. (2011). Discussion of NJ2011. Environmetrics 22 (7), 810–812.

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.