Marginal modelling of spatially-dependent non-stationary extremes using threshold modelling

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

Environmental Extremes
Royal Statistical Society
2nd December 2010

Outline

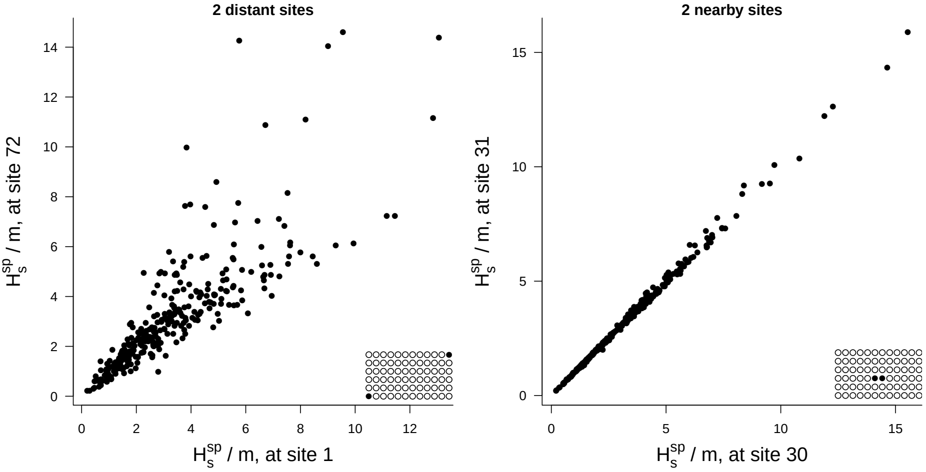
- Wave height data
- Spatial non-stationarity and dependence
- Thresholds for non-stationary extremes
- Model parameterisation
- Theoretical and simulation studies
- Wave height data

Wave height hindcasts from the Gulf of Mexico

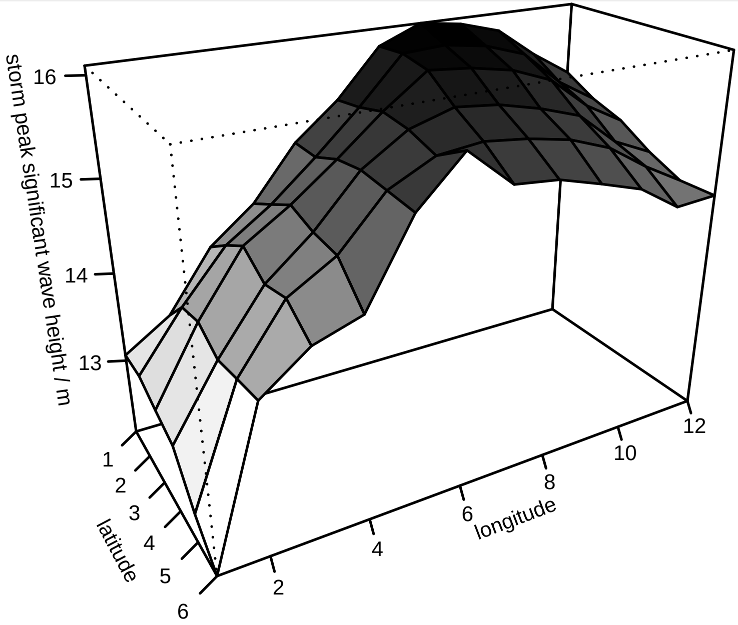
- Data supplied by Philip Jonathan at Shell Research UK.
- Hindcasts of Y storm peak significant wave height (in metres) in the Gulf of Mexico.
  - **wave height**: trough to the crest of the wave.
  - **significant wave height**: the average of the largest 1/3 wave heights. A measure of sea surface roughness.
  - **storm peak:** largest value from each storm (cf. declustering).
- a 6×12 grid of 72 sites ($\approx$ 14 km apart).
- Sep 1900 to Sep 2005 : 315 storms in total.
- average of 3 observations (storms) per year, at each site.

Aim: quantify the extremal behaviour of Y at each site, making appropriate adjustment for spatial dependence.

Spatial dependence



Spatial non-stationarity



Modelling issues

- Spatial non-stationarity.
    - Simple approach: model spatial effects on EV parameters as Legendre polynomials in longitude and latitude.
    - More flexible approaches are possible.
- Use a threshold that varies over space?
- Spatial dependence.
    - Estimate parameters assuming conditional independence of responses given covariate values.
    - Adjust standard errors etc. for spatial dependence.

Extreme value regression model

Conditional on covariates $\mathbf{x}_{ij}$ exceedances over a high threshold $u(\mathbf{x}_{ij})$ follow a 2-dimensional non-homogeneous Poisson process.

If responses $Y_{ij}, i = 1, \ldots, 72$ (space), $j = 1, \ldots, 315$ (storm) are conditionally independent:

$$
\begin{aligned}
L(\theta) \;=\; & \prod_{j=1}^{315} \prod_{i=1}^{72} \exp\left\{ -\frac{1}{\lambda} \left[1 + \xi(\mathbf{x}_{ij}) \left(\frac{u(\mathbf{x}_{ij}) - \mu(\mathbf{x}_{ij})}{\sigma(\mathbf{x}_{ij})} \right) \right]_+^{-1/\xi(\mathbf{x}_{ij})} \right\} \\
& \times \prod_{j=1}^{315} \prod_{i:y_{ij}>u(\mathbf{x}_{ij})} \frac{1}{\sigma(\mathbf{x}_{ij})} \left[1 + \xi(\mathbf{x}_{ij}) \left(\frac{y_{ij} - \mu(\mathbf{x}_{ij})}{\sigma(\mathbf{x}_{ij})} \right) \right]_+^{-1/\xi(\mathbf{x}_{ij})-1} .
\end{aligned}
$$

λ : mean number of observations per year;
$\mu(\mathbf{x}_{ij}), \sigma(\mathbf{x}_{ij}), \xi(\mathbf{x}_{ij})$: GEV parameters of annual maxima at $\mathbf{x}_{ij}$;
θ : vector of all model parameters:

Covariate-dependent thresholds

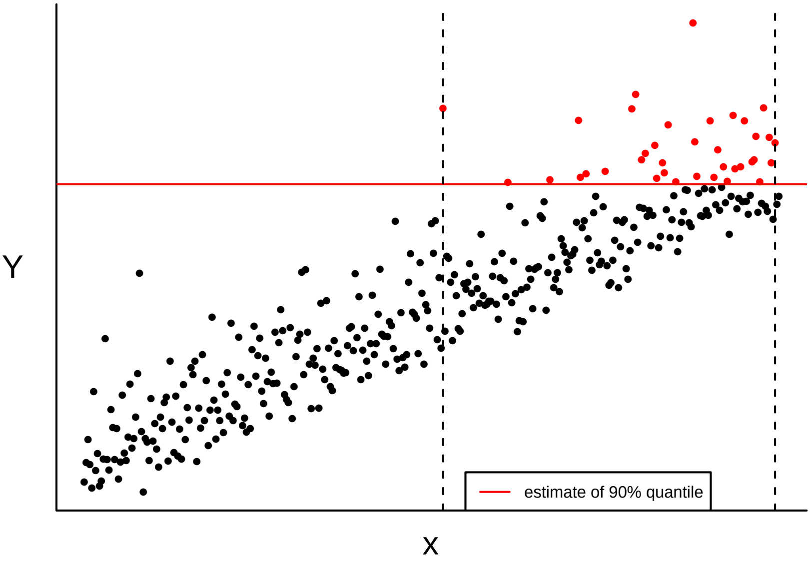
Arguments for:

- Asymptotic justification for EV regression model : the threshold $u(\mathbf{x}_{ij})$ needs to be high for each $\mathbf{x}_{ij}$.
- Design : spread exceedances across a wide range of covariate values.

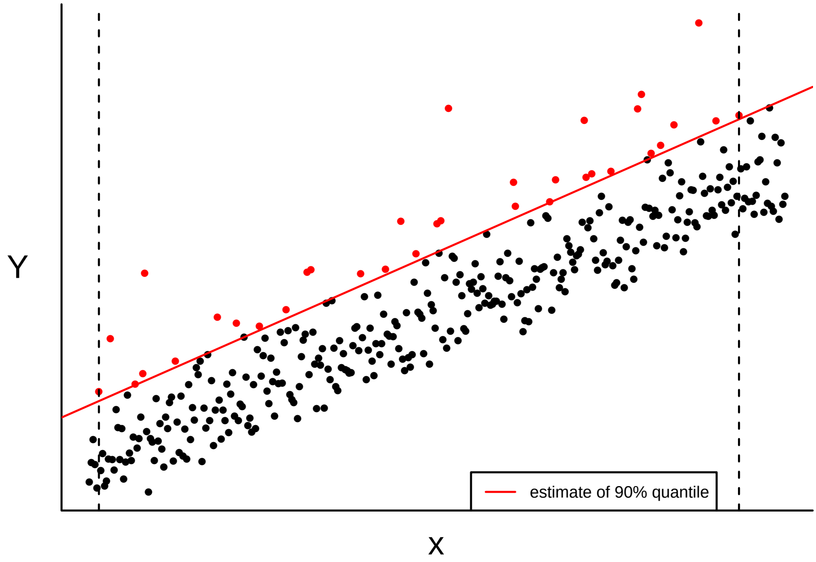
Set $u(\mathbf{x}_{ij})$ so that $P(Y > u(\mathbf{x}_{ij}))$, is approx. constant for all $\mathbf{x}_{ij}$.

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

Constant threshold



Quantile regression



Model parameterisation

Let $p(\mathbf{x}_{ij}) = P(Y_{ij} > u(\mathbf{x}_{ij}))$. Then, if $\xi(\mathbf{x}_{ij}) = \xi$ is constant,

$$p(\mathbf{x}_{ij}) \approx \frac{1}{\lambda}\left[1 + \xi\left(\frac{u(\mathbf{x}_{ij}) - \mu(\mathbf{x}_{ij})}{\sigma(\mathbf{x}_{ij})}\right)\right]^{-1/\xi}.$$

If $p(\mathbf{x}_{ij}) = p$ is constant then

$$u(\mathbf{x}_{ij}) = \mu(\mathbf{x}_{ij}) + c\,\sigma(\mathbf{x}_{ij}).$$

The form of $u(\mathbf{x}_{ij})$ is determined by the extreme value model:

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

Theoretical study (with Nicolas Attalides)

Data-generating process: for covariate values $x_1, \ldots, x_n$

$$Y_i \mid X = x_i \overset{\text{indep}}{\sim} GEV(\mu_0 + \mu_1 \, x_i, \sigma, \xi).$$

Set threshold

$$u(x) = u_0 + u_1 \, x.$$

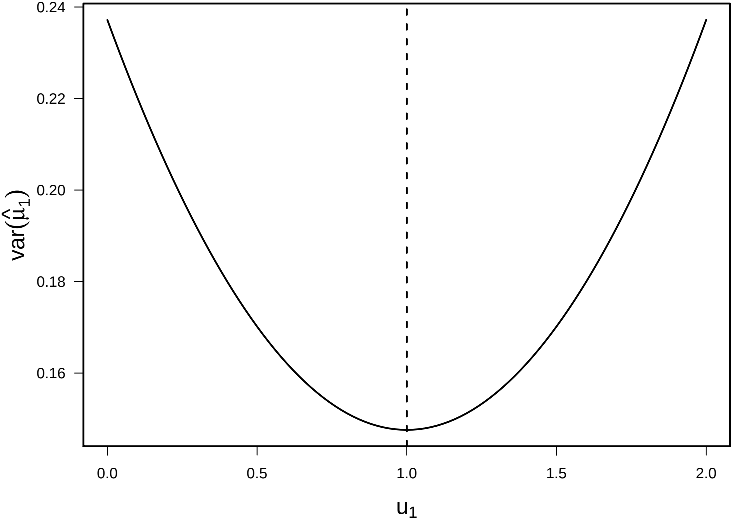
Vary u_1, set u_0 so that the expected proportion of exceedances is kept constant at p.

- Calculate Fisher expected information for $(\mu_0, \mu_1, \sigma, \xi)$.
- Invert to find asymptotic V-C of MLEs $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\sigma}, \widehat{\xi}$ and hence $\text{var}(\widehat{\mu}_1)$.
- Find the value of u_1 that minimises $\text{var}(\widehat{\mu}_1)$.

Let $\tilde{u}_1$ be the value of u_1 that minimises $\text{var}(\widehat{\mu}_1)$.

- If covariate values $x_1, \ldots, x_n$ are symmetrically distributed then $\tilde{u}_1 = \mu_1$ (quantile regression).

$\mu_1 = 1$: symmetric x

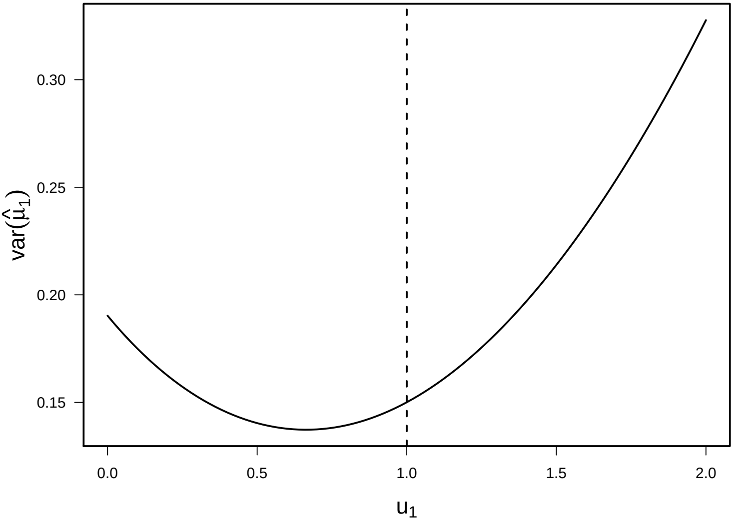


Preliminary findings

Let $\tilde{u}_1$ be the value of u_1 that minimises var$(\widehat{\mu}_1)$.

- If covariate values $x_1, \ldots, x_n$ are symmetrically distributed then $\tilde{u}_1 = \mu_1$ (quantile regression).
- If $x_1, \ldots, x_n$ are positive (negative) skew then $\tilde{u}_1 < \mu_1$ ($\tilde{u}_1 > \mu_1$).

$\mu_1 = 1$: positive skew x (coeff. of skewness $= 1$)



Preliminary findings

Let $\tilde{u}_1$ be the value of u_1 that minimises $\text{var}(\widehat{\mu}_1)$.

- If covariate values $x_1, \ldots, x_n$ are symmetrically distributed then $\tilde{u}_1 = \mu_1$ (quantile regression).
- If $x_1, \ldots, x_n$ are positive (negative) skew then $\tilde{u}_1 < \mu_1$ ($\tilde{u}_1 > \mu_1$).

... but the loss in efficiency from using $\tilde{u}_1 = \mu_1$ is small.

Extensions:

- More general models.
- Effect of model mis-specification due to low threshold;

Adjustment for spatial dependence (for wave height data)

Independence log-likelihood:

$$l_{IND}(\theta) \;=\; \underbrace{\sum_{j=1}^{k}}_{\text{(storms)}} \; \underbrace{\sum_{i=1}^{72}}_{\text{(space)}} \log f_{ij}(y_{ij};\theta) = \sum_{j=1}^{k} l_j(\theta).$$

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

$$\widehat{\theta} \to N(\theta_0, I^{-1}\, V\, I^{-1}),$$

- I = Fisher expected information: $-\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\, l_{IND}(\theta_0)\right)$;
- $V = \mathrm{var}\left(\frac{\partial}{\partial\theta}\, l_{IND}(\theta)\right) = \sum_j \mathrm{var}\left(U_j(\theta_0)\right) = \sum_j \mathrm{E}\left(U_j^2(\theta_0)\right)$,

  where

$$U_j(\theta) = \frac{\partial l_j(\theta)}{\partial\theta}.$$

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

Estimate

- I by Fisher observed information, evaluated at $\widehat{\theta}$;
- V by $\displaystyle\sum_{j=1}^{k} U_j^2\left(\widehat{\theta}\right)$.

Let $H_A = \left(-\widehat{I}^{-1}\,\widehat{V}\,\widehat{I}^{-1}\right)^{-1}$ and $H_I = -\widehat{I}$.

Chandler and Bate (2007):

$$l_{ADJ}(\theta) = l_{IND}(\widehat{\theta}) + \frac{(\theta-\widehat{\theta})' H_A (\theta-\widehat{\theta})}{(\theta-\widehat{\theta})' H_I (\theta-\widehat{\theta})}\left(l_{IND}(\theta) - l_{IND}(\widehat{\theta})\right),$$

- Adjust $l_{IND}(\theta)$ so that its Hessian is H_A at $\widehat{\theta}$ rather than H_I.
- This adjustment preserves the usual asymptotic distribution of the likelihood ratio statistic.

Simulation study

30 years of daily data on a spatial grid.

- Spatial dependence : mimics that of wave height data.
- Temporal dependence : moving maxima : extremal index 1/2.
- Spatial variation: location μ linear in longitude and latitude.

- ξ: $-0.2, 0.1, 0.4, 0.7$.
- Thresholds: 90th, 95th, 99th percentiles.
- SE adjustment: data from distinct years are independent.
- Simulations with no covariate effects and/or no spatial dependence for comparison.

Findings of simulation study

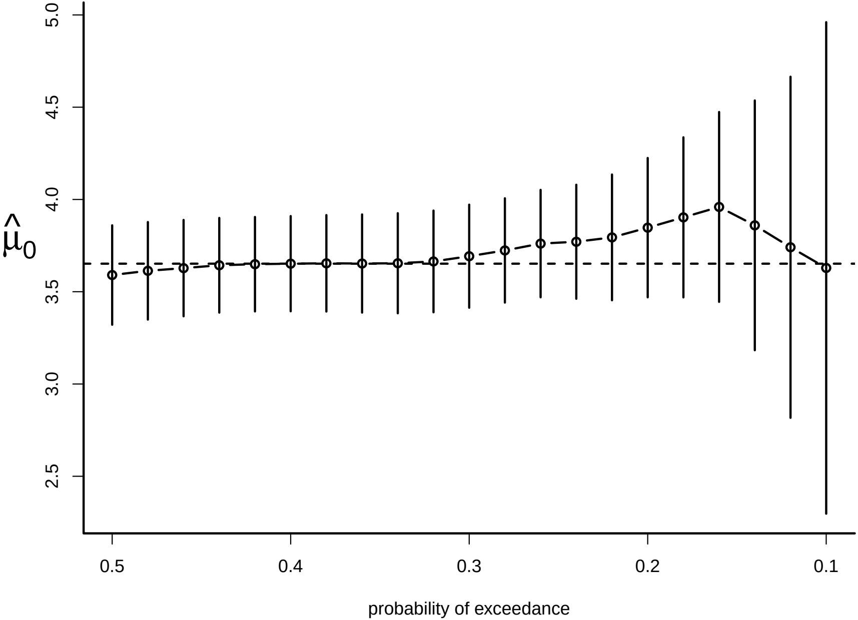
- Slight underestimation of standard errors : uncertainty in threshold ignored.
- Uncertainties in covariate effects of threshold are negligible compared to the uncertainty in the level of the threshold.
- Estimates of regression effects from QR and EV models are very close : both estimate extreme quantiles from the same data.
- To a large extent fitting the EV model accounts for uncertainty in the covariate effects at the level of the threshold.

Summary of modelling of wave height data

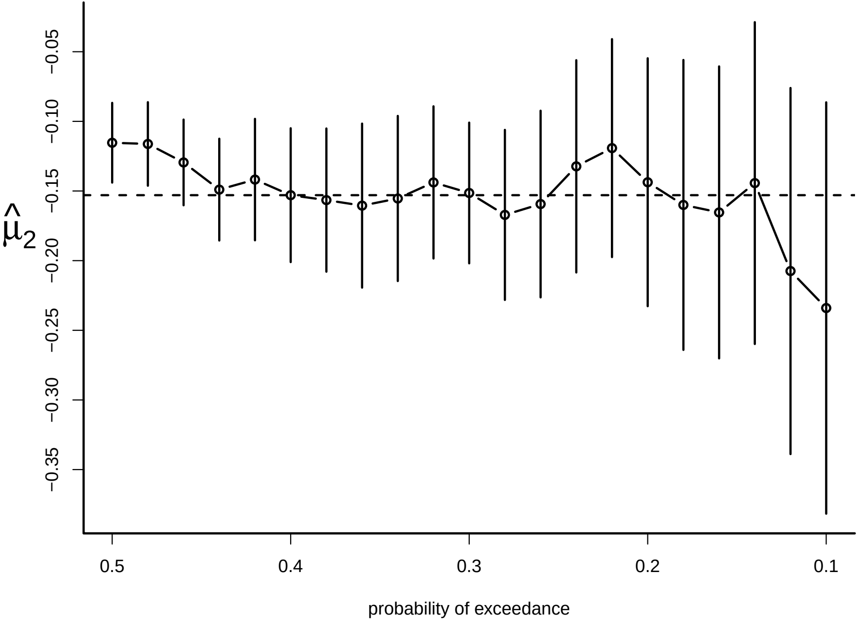
Threshold selection:

- Choice of p: look for stability in parameter estimates.
- Based on μ (and u) quadratic in longtiude and latitude, σ and ξ constant . . .

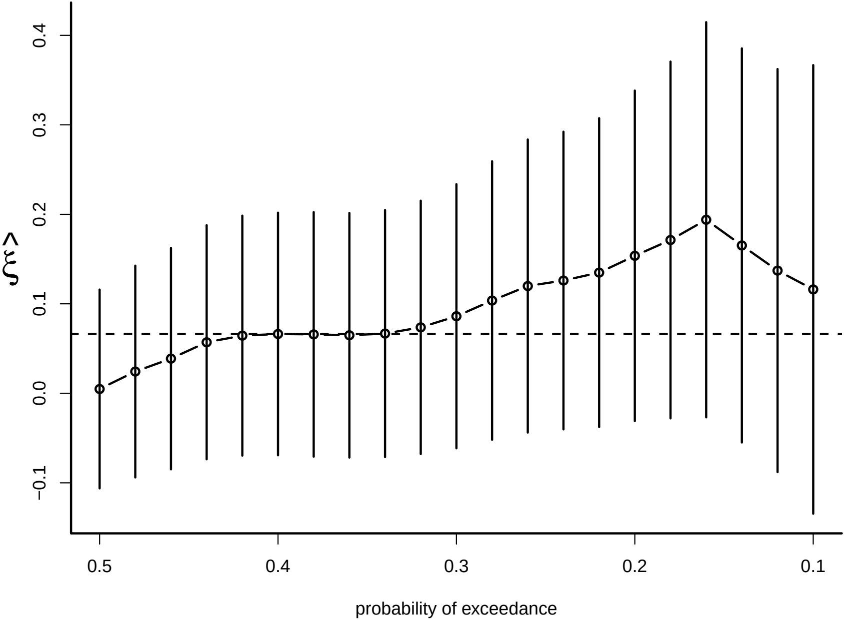
Threshold selection : μ intercept



Threshold selection : μ coefficient of latitude



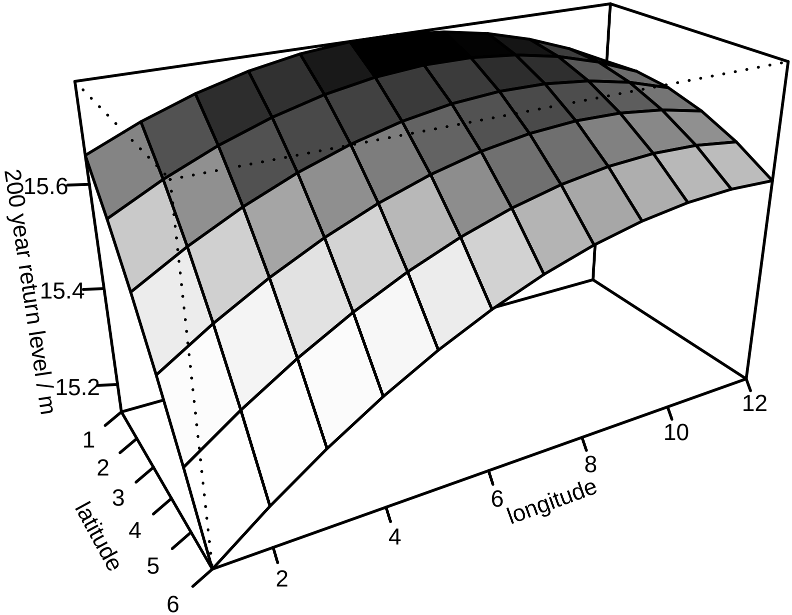
Threshold selection : ξ



Summary of modelling of wave height data

- Choice of p: look for stability in parameter estimates.
  Use $p = 0.4$.
- Model diagnostics : slight underestimation at very high levels, but consistent with estimated sampling variability.
- QR model and EV model agree closely.
- $\widehat{\xi} = 0.066$, with 95% confidence interval $(-0.052, 0.223)$.
- Estimated 200 year return level at (long=7, lat=1) is 15.78m with 95% confidence interval $(12.90, 22.28)$m.

Conditional 200 year return levels



Conclusions

Quantile regression:

- a simple and effective strategy to set thresholds for non-stationary EV models;
- supported by simulation study;
- theoretical work is on-going;

Kyselý, J., et al. (2010) use quantile regression to set a time-dependent threshold for peaks-over-threshold GP modelling of data simulated from a climate model.

Spatial dependence

- adjust inferences for dependence; or
- model dependence explicitly.

References

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

Kyselý, J., Picek, J. and Beranová, R. (2010) Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold *Global and Planetary Change*, **72**, 55-68.

Northop, P. J. and Jonathan, P. Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights. Revisions completed for *Environmetrics*.

Thank you for your attention.