Non-stationary extremes with splines

Philip Jonathan

philip.jonathan@shell.com
www.lancs.ac.uk/∼jonathan

Shell Projects & Technology
Lancaster University

Hejnice 2013

Thanks

Thanks for contributions by Shell colleagues:

- Kevin Ewans, Graham Feld, David Randell, Yanyun Wu

... and Lancaster students:

- Kaylea Haynes, Emma Ross, Elena Zanini

Outline

1. Background
   - Motivation
   - Australian North West Shelf
2. Extreme value analysis: challenges
   - Univariate challenges
   - Multivariate challenges
3. Non-stationary extremes
   - Penalised B-splines
   - Quantile regression model for extreme value threshold
   - Poisson model for rate of threshold exceedance
   - Generalised Pareto model for size of threshold exceedance
   - Return values
4. Current developments
   - Extremal dependence
   - Conditional extremes
   - Spatial extremes

Contents

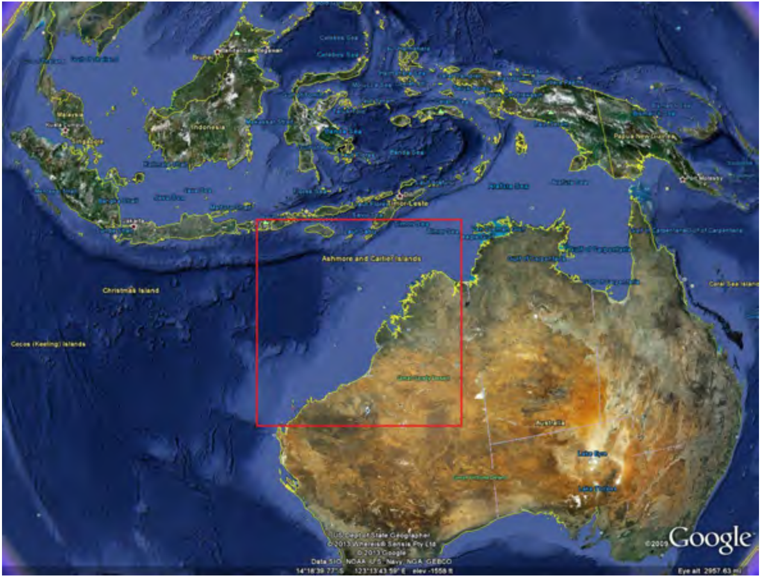
Motivation

- **Rational** design an assessment of marine structures:
  - Reducing **bias** and **uncertainty** in estimation of structural reliability
  - Improved understanding and communication of risk

  - For new (e.g. floating) and existing (e.g. steel and concrete) structures
  - Climate change

- Other applied fields for extremes in industry:
  - Corrosion and fouling
  - Economics and finance

Australian North West Shelf



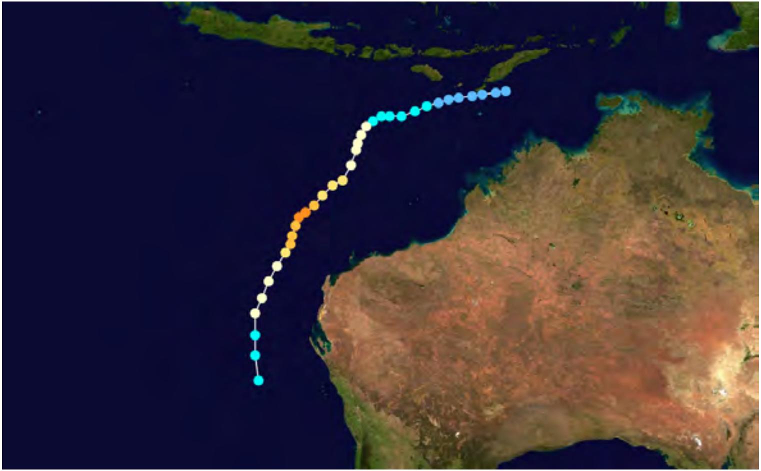
Australian North West Shelf

- Model **storm peak significant wave height** H_S
- Wave climate is dominated by westerly **monsoonal swell** and **tropical cyclones**
- Cyclones originate from Eastern Indian Ocean, Timor and Arafura Sea

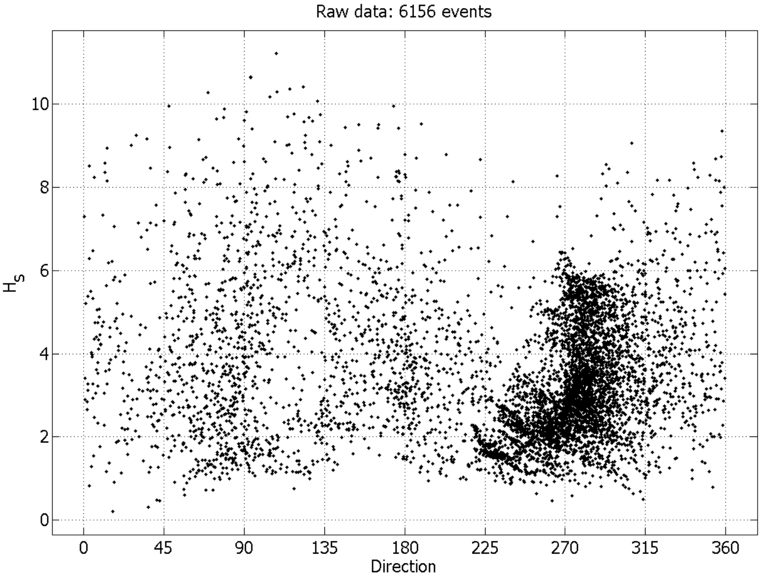
- Sample of **hindcast** storms for period 1970-2007
- 9×9 rectangular spatial grid over $5^o \times 5^o$ longitude-latitude domain
- **Spatial** and **directional** variability in extremes present
- **Marginal** spatio-directional model

Cyclone Narelle January 2013: spatio-directional

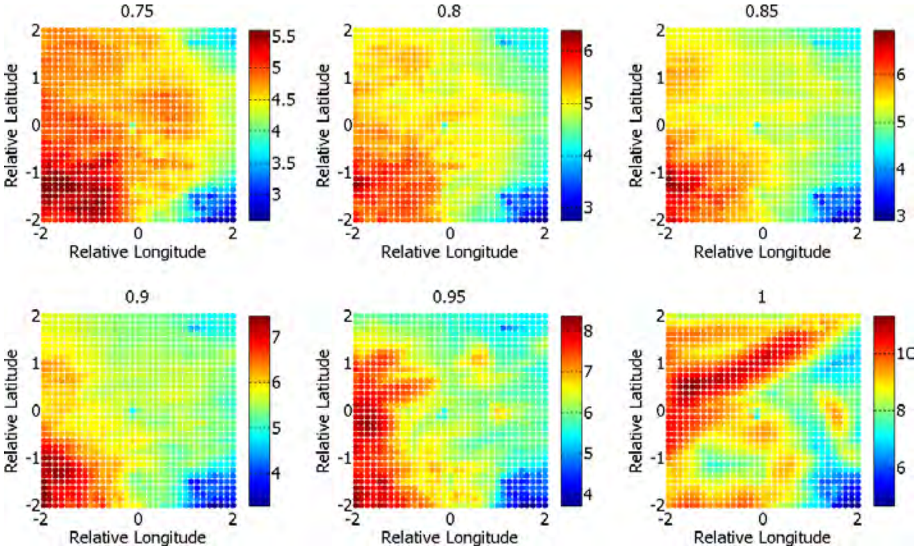
Cyclone Narelle January 2013: cyclone track



Storm peak H_S by direction



Quantiles of storm peak H_S spatially



Contents

Extreme value analysis: *univariate* challenges

- **Covariates** and **non-stationarity**:
  - Location, direction, season, time, water depth, ...
  - Multiple / multidimensional covariates in practice
- **Cluster** dependence:
  - Same events observed at many locations (pooling)
  - Dependence in time (Chavez-Demoulin and Davison 2012)
- **Scale** effects:
  - Modelling X or $f(X)$? (Reeve et al. 2012)
- **Threshold** estimation:
  - Scarrott and MacDonald [2012]
- **Parameter** estimation
  - Maximum likelihood, moments, Hill, ...
- **Measurement** issues:
  - Field measurement uncertainty greatest for extreme values
  - Hindcast data are simulations based on pragmatic physics, calibrated to historical observation

Extreme value analysis: *multivariate* challenges

- **Spatial extremes using componentwise maxima**:
  - $\Leftrightarrow$ max-stability $\Leftrightarrow$ multivariate regular variation
  - Assumes all components extreme
  - $\Rightarrow$ Perfect independence or asymptotic dependence **only**
  - Composite likelihood for spatial extremes (Davison et al. 2012)
- **Extremal dependence**: (Ledford and Tawn 1997)
  - Assumes regular variation of joint survivor function
  - Gives more general forms of extremal dependence
  - $\Rightarrow$ Asymptotic dependence, asymptotic independence (with +ve, -ve association)
  - Hybrid spatial dependence model (Wadsworth and Tawn 2012)
- **Conditional extremes**: (Heffernan and Tawn 2004)
  - Assumes, given one variable being extreme, convergence of distribution of remaining variables
  - Allows some variables not to be extreme
  - Not equivalent to extremal dependence
- Application:
  - ... *a huge gap in the theory and practice of multivariate extremes* ... (Beirlant et al. 2004)

Contents

Model components

- Sample $\{\dot{z}_i\}_{i=1}^{\dot{n}}$ of $\dot{n}$ storm peak significant wave heights observed at locations $\{\dot{x}_i, \dot{y}_i\}_{i=1}^{\dot{n}}$ with storm peak directions $\{\dot{\theta}_i\}_{i=1}^{\dot{n}}$
- Model components:
  1. **Threshold** function ϕ above which observations $\dot{z}$ are assumed to be extreme estimated using quantile regression
  2. **Rate of occurrence** of threshold exceedances modelled using Poisson model with rate $\rho(\triangleq \rho(\theta, x, y))$
  3. **Size of occurrence** of threshold exceedance using generalised Pareto (GP) model with shape and scale parameters ξ and σ

Model components

- Rate of occurrence and size of threshold exceedance functionally **independent** (Chavez-Demoulin and Davison 2005)
  - Equivalent to non-homogeneous Poisson point process model (Dixon et al. 1998)
- Smooth functions of covariates estimated using penalised B-splines (Eilers and Marx 2010)
  - Slick linear algebra (c.f. generalised linear array models, Currie et al. 2006)

Penalised B-splines

- Physical considerations suggest model parameters ϕ, ρ, ξ and σ vary smoothly with covariates θ, x, y
- Values of $(\eta =)\phi, \rho, \xi$ and σ all take the form:

$$\eta = B\beta_\eta$$

  for **B-spline** basis matrix B (defined on index set of covariate values) and some β_η to be estimated
- Multidimensional basis matrix B formulated using Kronecker products of marginal basis matrices:

$$B = B_\theta \otimes B_x \otimes B_y$$

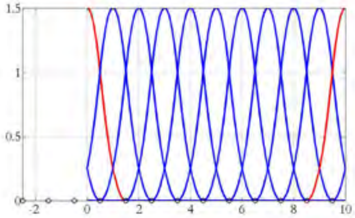
- Roughness R_η defined as:

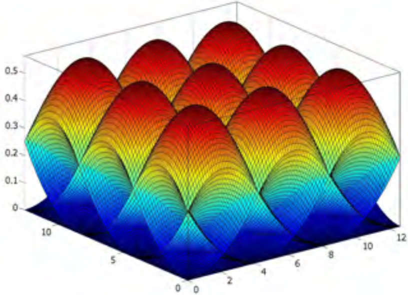
$$R_\eta = \beta_\eta' P \beta_\eta$$

  where effect of P is to difference neighbouring values of β_η

Penalised B-splines

- **Wrapped** bases for periodic covariates (seasonal, direction)
- **Multidimensional** bases easily constructed. **Problem size** sometimes prohibitive
- Parameter **smoothness** controlled by roughness coefficient λ: cross validation chooses λ optimally





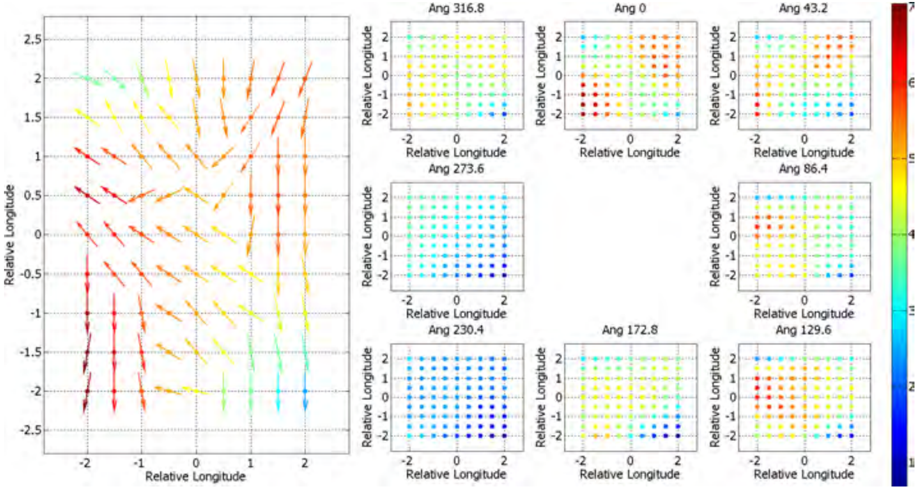
Quantile regression model for extreme value threshold

- Estimate smooth quantile $\phi(\theta, x, y; \tau)$ for non-exceedance probability τ of z (storm peak H_S) using quantile regression by minimising **penalised** criterion ℓ_ϕ^* with respect to basis parameters:

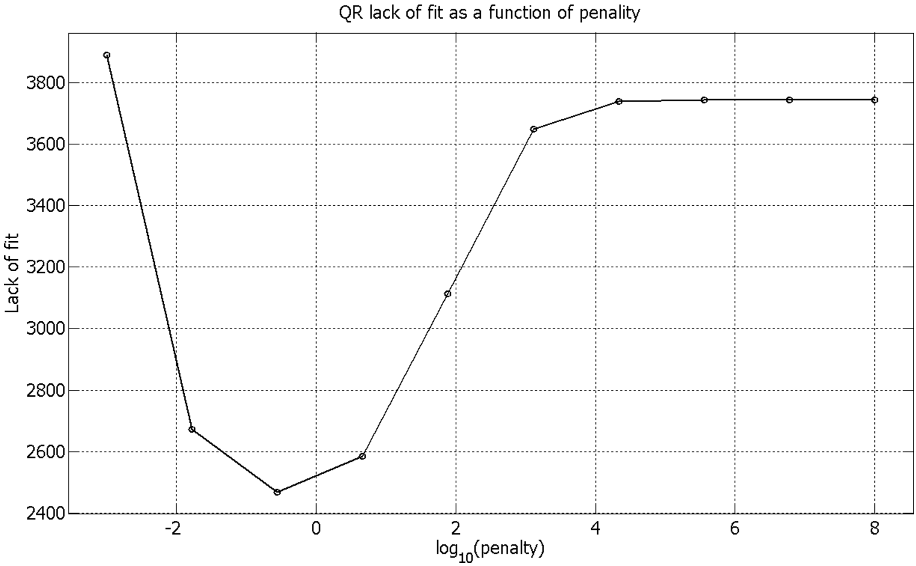
$$\begin{aligned} \ell_\phi^* &= \ell_\phi + \lambda_\phi R_\phi \\ \ell_\phi &= \{\tau \sum_{r_i \geq 0}^{n} |r_i| + (1-\tau) \sum_{r_i < 0}^{n} |r_i|\} \end{aligned}$$

  for $r_i = z_i - \phi(\theta_i, x_i, y_i; \tau)$ for $i = 1, 2, ..., n$, and **roughness** R_ϕ controlled by roughness coefficient λ_ϕ
- (Non-crossing) quantile regression formulated as linear programme (Bollaerts et al. 2006)

Spatio-directional 50% quantile threshold



Cross-validation for optimal roughness



Poisson model for rate of threshold exceedance

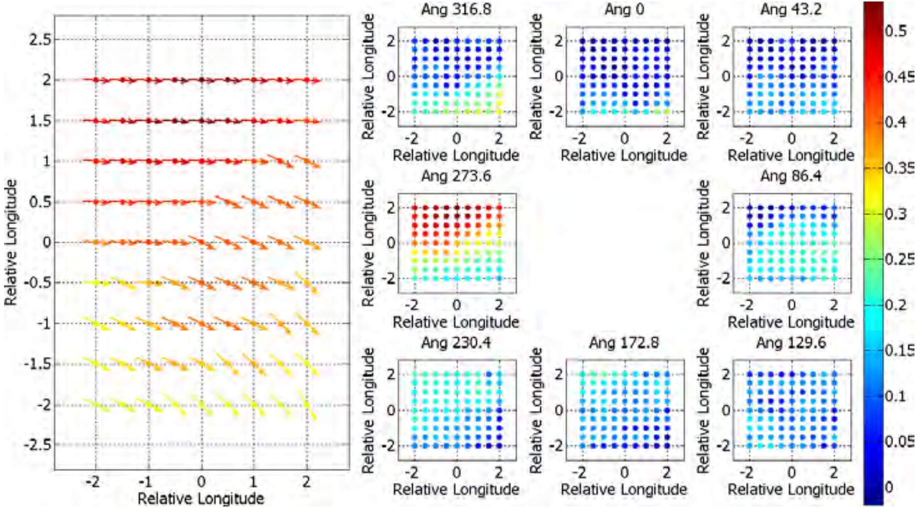
- Poisson model for rate of occurrence of threshold exceedance estimated by minimising roughness penalised log likelihood:

$$\ell^*_\rho = \ell_\rho + \lambda_\rho R_\rho$$

- (Negative) penalised Poisson log-likelihood (and approximation):

$$\begin{aligned} \ell_\rho &= -\sum_{i=1}^{n} \log \rho(\theta_i, x_i, y_i) + \int \rho(\theta, x, y) d\theta dx dy \\ \hat{\ell}_\rho &= -\sum_{j=1}^{m} c_j \log \rho(j\Delta) + \Delta \sum_{j=1}^{m} \rho(j\Delta) \end{aligned}$$

- $\{c_j\}_{j=1}^m$ counts of threshold exceedances on index set of m ($>> 1$) bins partitioning covariate domain into intervals of volume Δ
- λ_ρ estimated using cross validation


Relative Longitude
Relative Longitude
Ang 316.8
Ang 0
Ang 43.2
Ang 273.6
Ang 86.4
Ang 230.4
Ang 172.8
Ang 129.6
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0

Generalised Pareto model for size of threshold exceedance

- Generalise Pareto model for size of threshold exceedance estimated by minimising roughness penalised log-likelihood:

$$\ell^*_{\xi,\sigma} = \ell_{\xi,\sigma} + \lambda_\xi R_\xi + \lambda_\sigma R_\sigma$$

- (Negative) conditional generalised Pareto log-likelihood:

$$\ell_{\xi,\sigma} = \sum_{i=1}^{n} \log \sigma_i + \frac{1}{\xi_i} \log(1 + \frac{\xi_i}{\sigma_i}(z_i - \phi_i))$$

- Parameters: **shape** ξ, **scale** σ
- Threshold ϕ set prior to estimation
- λ_ξ and λ_σ estimated using cross validation. In practice set $\lambda_\xi = \kappa\lambda_\sigma$ for fixed κ

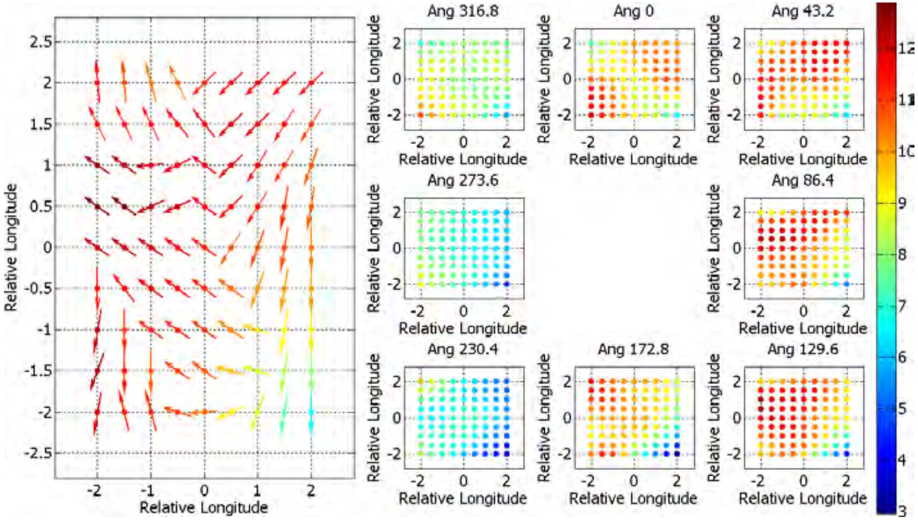
Return values

- Return value z_T of storm peak significant wave height corresponding to return period T (years) evaluated from estimates for ϕ, ρ, ξ and σ:

$$z_T = \phi - \frac{\sigma}{\xi}(1 + \frac{1}{\rho}(\log(1 - \frac{1}{T}))^{-\xi})$$

- z_{100} corresponds to 100–year return value, denoted H_{S100}
- Alternative: estimation of return values by simulation under model

Spatio-directional 100-year return value H_{S100}



Contents

Current developments

- Non-stationarity
  - Spatio-directional, seasonal-directional and spatio-seasonal-directional
- Computational efficiency
  - Sparse and **slick** matrix manipulations
- Quantifying uncertainty
  - Bootstrapping, Bayesian (Nasri et al. 2013, Oumow et al. 2012)
- Spatial dependence
  - Composite likelihood: model componentwise maxima
  - Censored likelihood: block maxima $\rightarrow$ threshold exceedances
  - Hybrid model: **full range** of extremal dependence
- Interpretation within **structural design framework**
- Non-stationary **conditional** extremes
  - Spline representations for parameters of marginal and conditional extremes models (Jonathan et al. 2013)

Types of extremal dependence

Extremal dependence

- Bivariate random variable (X, Y)
- $\chi = \lim_{x\to\infty} Pr(X > x|Y > x)$
- *asymptotically independent* if $\chi = 0$
- *asymptotically dependent* if $\chi > 0$

- Extremal dependence models:
    - Admit asymptotic independence.
- But have issues with:
    - Thresholds
    - Covariates
    - High dimensions

- Ideas from theory of **regular variation** (see Bingham et al. 1987)

Limit assumption 1 on joint tail

- (X_F, Y_F) with Frechet marginals ($Pr(X_F < f) = e^{-\frac{1}{f}}$).
- Assume $Pr(X_F > f, Y_F > f)$ is **regularly varying at infinity**:

$$lim_{f\to\infty}\frac{Pr(X_F > sf, Y_F > sf)}{Pr(X_F > f, Y_F > f)} = s^{-\frac{1}{\eta}} \text{ for some fixed } s > 0$$

- This suggests:

$$\begin{aligned}
Pr(X_F > sf, Y_F > sf) &\approx s^{-\frac{1}{\eta}} Pr(X_F > f, Y_F > f)\\
Pr(X_G > g+t, Y_G > g+t) &= Pr(X_F > e^{g+t}, Y_F > e^{g+t})\\
&\approx e^{-\frac{t}{\eta}} Pr(X_F > e^{g}, Y_F > e^{g})\\
&= e^{-\frac{t}{\eta}} Pr(X_G > g, Y_G > g)
\end{aligned}$$

  on Gumbel scale X_G: $Pr(X_G < g) = \exp(-e^{-g})$.
- η is known as the **coefficient of tail dependence**.
- η and χ **characterise** extremal dependence between two variables.

Limit assumption 2 on joint tail

- Ledford and Tawn [1997] motivated by Bingham et al. [1987]
- Assume model $Pr(X_F > f, Y_F > f) = \ell(f) f^{-\frac{1}{\eta}}$
  - $\ell(f)$ is a **slowly-varying** function, $lim_{f \to \infty} \frac{\ell(sf)}{\ell(f)} = 1$
- Then:

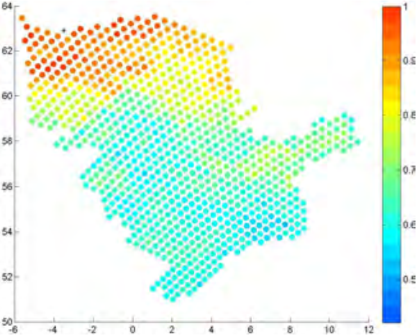
$$\begin{aligned} Pr(X_F > f | Y_F > f) &= \frac{Pr(X_F > f, Y_F > f)}{Pr(Y_F > f)} \\ &= \ell(f) f^{-\frac{1}{\eta}} (1 - e^{-\frac{1}{f}})^{-1} \\ &\sim \ell(f) f^{1-\frac{1}{\eta}} \\ &\sim \ell(f) Pr(Y_F > f)^{\frac{1}{\eta}-1} \end{aligned}$$

- At $\eta < 1$ (or $lim_{f \to \infty} \ell(f) = 0$), X_F and Y_F are **As.Ind.**!
- η **easily estimated from a sample** by noting that L_F, the **minimum** of X_F and Y_F is approximately GP-distributed:

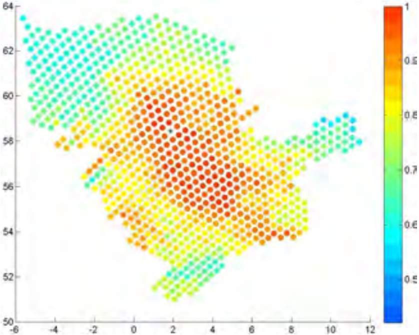
$$Pr(L_F > f + s | L_F > f) \sim (1 + \frac{s}{f})^{-\frac{1}{\eta}} \text{ for large } f$$

Characterising pairwise spatial dependence using η

Northern North Sea



Central North Sea



- Asymptotic independence if $\eta < 1$
- Asymptotic dependence $\eta = 1$ valid locally only
- Non-stationary region of asymptotic dependence

Conditional extremes

Limit assumption on conditional tail

Limit assumption on conditional tail

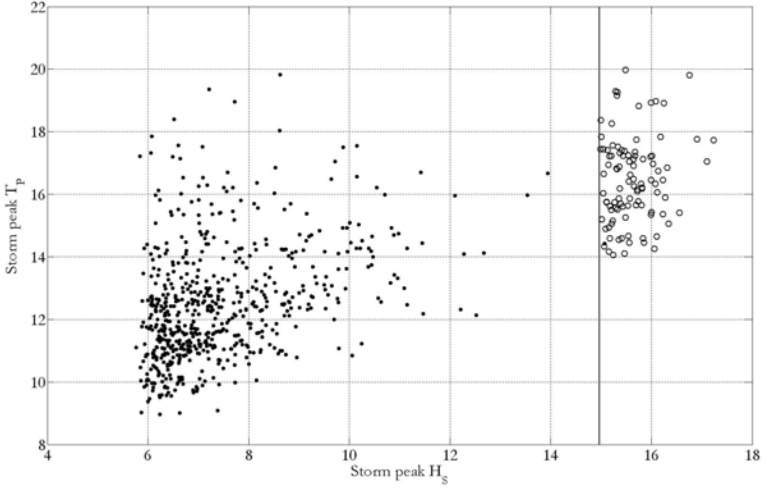
- Model conditional (and hence joint) extremes of two variables
- Heffernan and Tawn [2004]

- Sample $\{x_{i1}, x_{i2}\}_{i=1}^{n}$ of variate X_1 and X_2
- (X_1, X_2) transformed to (Y_1, Y_2) on **standard Gumbel** scale
- Model $(Y_2|Y_1 = y) = ay + y^b Z$ for **large** y and **positive** dependence
- Model $(Y_1|Y_2 = y)$ similarly

- Appropriate for most known distributional forms, but not all
- Simulation to sample joint distribution of (Y_1, Y_2) (and (X_1, X_2))

- Encompasses **both** asymptotic dependence and asymptotic independence
- Extends naturally (pairwise) to **high dimensions**
- But: consistency of $(Y_2|Y_2)$ and $(Y_1|Y_2)$ **not** ensured

Simple stationary conditional extremes



Non-stationary conditional extremes

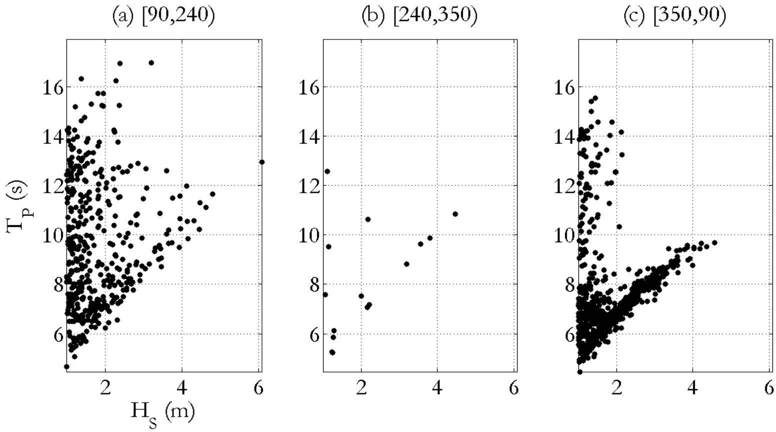
On **Gumbel** scale, extend with common covariate θ:

$$(Y_2|Y_1 = y, \theta) = \alpha_\theta y + y^{\beta_\theta}(\mu_\theta + \sigma_\theta Z) \text{ for } y > \phi_\theta(\tau)$$

where:

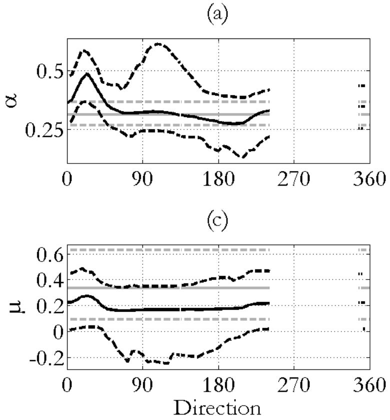
- $\phi_\theta(\tau)$ is a high non-stationary quantile of Y_1 on Gumbel scale, for non-exceedance probability τ, above which the model fits well
- $\alpha_\theta \in [0, 1]$, $\beta_\theta \in (-\infty, 1]$, $\sigma_\theta \in [0, \infty)$
- Z is a random variable with **unknown** distribution G, assumed Normal for estimation

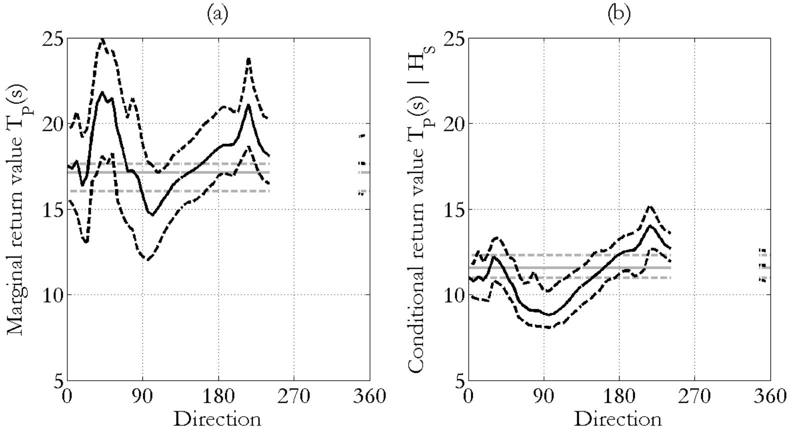
South Atlantic Ocean sample



Single directional covariate. Three directional sectors identified by consideration of fetch conditions, with differing sample characteristics

South Atlantic Ocean parameter estimates





More at *www.lancs.ac.uk/∼jonathan/NSCE13.pdf*

Spatial extremes

Modelling of component-wise maxima

Component-wise maxima

- Beirlant et al. [2004] is a nice introduction.
- No obvious way to order multivariate observations.
- Theory based on **component-wise maximum**, M.
  - For sample $\{x_{ij}\}_{i=1}^{n}$ in p dimensions:
  - $M_j = max_{i=1}^{n}\{x_{ij}\}$ for each j.
  - M probably not a sample point!
- $P(M \leq x) = \prod_{j=1}^{p} P(X_j \leq x_j) = F^n(x)$
  - Assume: $F^n(a_n x + b_n) \xrightarrow{D} G(x)$
  - Therefore also: $F_j^n(a_{n,j} x_j + b_{n,j}) \xrightarrow{D} G_j(x_j)$

Homogeneity

- Limiting distribution with Frechet marginals, G_F
  - $G_F(z) = G(G_1^{\leftarrow}(e^{-\frac{1}{z_1}}), G_2^{\leftarrow}(e^{-\frac{1}{z_2}}), ..., G_p^{\leftarrow}(e^{-\frac{1}{z_p}}))$

- $V_F(z) = -\log G_F(z)$ is the **exponent measure** function
- $V_F(sz) = s^{-1} V_F(z)$ **homogeneity order -1**
- $V_F(1)$ is known as the **extremal coefficient** (and $V(1) = 2 - \chi$)

Homogeneity order -1 is equivalent to asymptotic dependence (or **perfect** independence):

$$
\begin{aligned}
P(X > sf, Y > sf) &= 1 - (P(X > sf) + P(Y > sf) \\
&+ P(X \leq sf, Y \leq sf)) \\
&= (1 - P(X \leq sf, Y \leq sf)) - 2P(X > sf) \\
&= (1 - \exp(-V(sf, sf))) - 2(1 - \exp(-1/(sf))) \\
&\approx V(sf, sf) = s^{-1} V(f, f) \text{ for large } f \\
&= s^{-1} P(X > f, X > f) \text{ so that } \eta = 1
\end{aligned}
$$

Composite likelihood for spatial dependence

- Composite likelihood $l_C(\theta)$ assuming Frechet marginals:

$$l_C(\theta) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \log f(z_i, z_j; \theta)$$

$$f(z_i, z_j) = (\frac{\partial V(z_i, z_j)}{\partial z_i}\frac{\partial V(z_i, z_j)}{\partial z_j} - \frac{\partial^2 V(z_i, z_j)}{\partial z_i \partial z_j})e^{-V(z_i, z_j)}$$

- Lots of possible exponent measures with simple bivariate parametric forms with pre-specified functions (e.g. of distance) whose parameters must be estimated:
  - Smith (Spatial Gaussian process)
  - Schlather (Extremal Gaussian process)
  - Geometric Gaussian
  - Brown-Resnick model
  - Davison and Gholamrezaee
  - Wadsworth & Tawn (Hybrid Gaussian-Gaussian process)
- See Davison et al. [2012].

Smith process

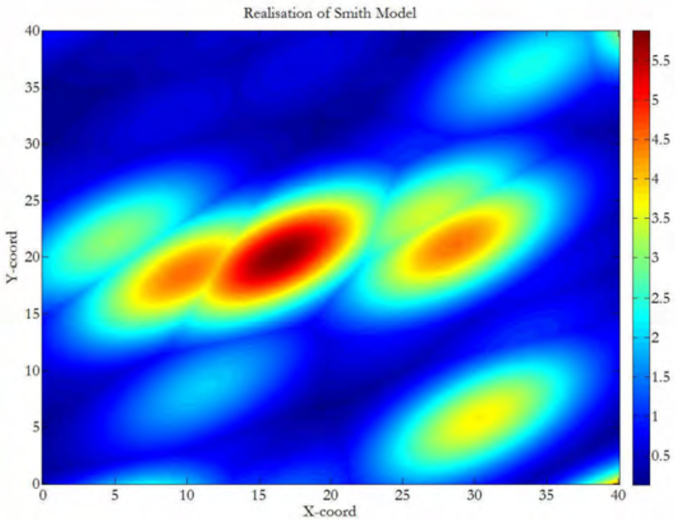
$$
\begin{aligned}
V(z_i, z_j) &= \frac{1}{z_i}\Phi(\frac{\alpha(h)}{2} + \frac{1}{\alpha(h)}\log(\frac{z_j}{z_i})) \\
&+ \frac{1}{z_j}\Phi(\frac{\alpha(h)}{2} + \frac{1}{\alpha(h)}\log(\frac{z_i}{z_j}))
\end{aligned}
$$

with pre-specified $\alpha(h) = (h'\Sigma^{-1}h)^{1/2}$ of distance h, where:

$$
\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}
$$

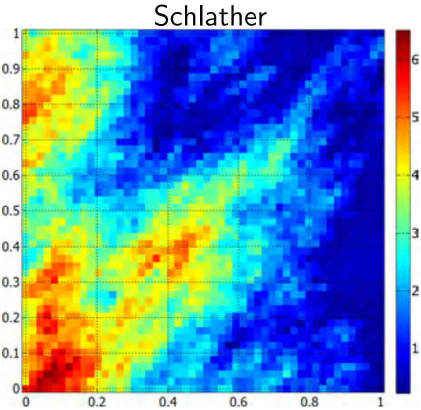
and σ_1^2, σ_{12} and σ_2^2 must be estimated.

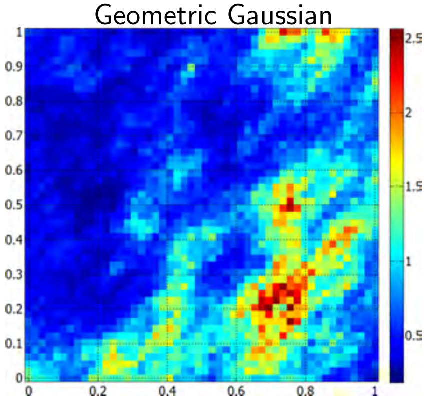
Realisation from Smith process



For case $\sigma_1^2 = 20$, $\sigma_{12} = 15$ and $\sigma_2^2 = 30$. Standard Frechet marginals.

Realisations: Schlather and geometric Gaussian processes





Useable spatial extremes

- Non-stationary spatial processes
  - parameterise in terms of covariates
- Modelling of threshold exceedances more efficient than block maxima
  - censored likelihood
- Cannot assume asymptotic dependence
  - hybrid model admits asymptotic dependence and asymptotic independence
- Computational efficiency

Summary

1. Background
   - Motivation
   - Australian North West Shelf
2. Extreme value analysis: challenges
   - Univariate challenges
   - Multivariate challenges
3. Non-stationary extremes
   - Penalised B-splines
   - Quantile regression model for extreme value threshold
   - Poisson model for rate of threshold exceedance
   - Generalised Pareto model for size of threshold exceedance
   - Return values
4. Current developments
   - Extremal dependence
   - Conditional extremes
   - Spatial extremes

References

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes: theory and applications*. Wiley, 2004.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*. Cambridge University Press, 1987.

K Bollaerts, P H C Eilers, and M Aerts. Quantile regression with monotonicity restrictions using P-splines and the L1 norm. *Statistical Modelling*, 6:189–207, 2006.

V. Chavez-Demoulin and A.C. Davison. Generalized additive modelling of sample extremes. *J. Roy. Statist. Soc. Series C: Applied Statistics*, 54:207, 2005.

V. Chavez-Demoulin and A.C. Davison. Modelling time series extremes. *REVSTAT - Statistical Journal*, 10:109–133, 2012.

I. D. Currie, M. Durban, and P. H. C. Eilers. Generalized linear array models with applications to multidimensional smoothing. *J. Roy. Statist. Soc. B*, 68:259–280, 2006.

A. C. Davison, S. A. Padoan, and M. Ribatet. Statistical modelling of spatial extremes. *Statistical Science*, 27:161–186, 2012.

J. M. Dixon, J. A. Tawn, and J. M. Vassie. Spatial modelling of extreme sea-levels. *Environmetrics*, 9:283–301, 1998.

P H C Eilers and B D Marx. Splines, knots and penalties. *Wiley Interscience Reviews: Computational Statistics*, 2:637–653, 2010.

J. E. Heffernan and J. A. Tawn. A conditional approach for multivariate extreme values. *J. R. Statist. Soc. B*, 66:497, 2004.

P. Jonathan, K. C. Ewans, and D. Randell. Joint modelling of environmental parameters for extreme sea states incorporating covariate effects. *Coastal Engineering*, 79:22–31, 2013.

A. W. Ledford and J. A. Tawn. Modelling dependence within joint tail regions. *J. R. Statist. Soc. B*, 59:475–499, 1997.

B. Nasri, S. El Adlouni, and T. B. M. J. Ouarda. Bayesian estimation for GEV B-Spline model. *Open Journal of Statistics*, 3: 118–128, 2013.

B. Oumow, M. de Carvalho, and A. C. Davison. Bayesian P-spline mixture modeling of extreme forest temperatures. *Available at www.mat.puc.cl/∼ mdecarvalho*, 2012.

D.T. Reeve, D. Randell, K.C.Ewans, and P. Jonathan. Accommodating measurement scale uncertainty in extreme value analysis of North Sea storm severity. *Ocean Eng.*, 53:164–176, 2012.

C. Scarrott and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT - Statistical Journal*, 10:33–60, 2012.

J.L. Wadsworth and J.A. Tawn. Dependence modelling for spatial extremes. *Biometrika*, 99:253–272, 2012.