Modelling extreme environments

Philip Jonathan
Shell Technology Centre Thornton, Chester, UK

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

SUTGEF meeting, Oxford
September 2011

Outline

- Motivation.
- Modelling challenges.
- Basics.
- Covariate effects in extremes.
- Multivariate extremes.
- Current developments.
- Conclusions.

Motivation

Katrina in the Gulf of Mexico.

Katrina damage.

Cormorant Alpha in a North Sea storm.

”L9” platform in the Southern North Sea.

A wave seen from a ship.

Black Sea coast.

Motivation

- **Rational** design an assessment of marine structures:
  - Reducing **bias** and **uncertainty** in estimation of structural reliability.
  - Improved understanding and communication of risk.
  - Climate change.
- Other applied fields for extremes in industry:
  - Corrosion and fouling.
  - Finance.
  - Networks.

Sanity check

- **All models are wrong, some models are useful.**
- George Box, *http://en.wikipedia.org/wiki/George_E._P._Box*
- How can we make models as useful as possible?
- Consistency between physical, engineering and statistical insights.

Modelling challenges

- **Covariate** effects:
  - Location, direction, season, ...
  - Multiple covariates in practice.
- **Cluster** dependence:
  - e.g. storms independent, observed (many times) at many locations.
  - e.g. dependent occurrences in time.
- **Scale** effects:
  - Modelling x^2 gives different estimates c.f. modelling x.
- **Threshold** estimation.
- **Parameter** estimation.
- **Measurement** issues:
  - Field measurement uncertainty greatest for extreme values.
  - Hindcast data are simulations based on pragmatic physics, calibrated to historical observation.

- **Multivariate** extremes:
  - Waves, winds, currents, forces, moments, displacements, ...
  - Componentwise maxima $\Leftrightarrow$ max-stability $\Leftrightarrow$ regular variation:
    - Assumes **all** components extreme.
    - $\Rightarrow$ Perfect independence or asymptotic dependence **only**.
  - Extremal dependence:
    - Assumes regular variation of joint survivor function.
    - Gives rise to more general forms of extremal dependence.
    - $\Rightarrow$ Asymptotic dependence, asymptotic independence.
  - Conditional extremes:
    - Assumes, given one variable being extreme, convergence of distribution of remaining variables.
    - Not equivalent to extremal dependence.
    - Allows some variables not to be extreme.
  - Inference:
    - ... *a huge gap in the theory and practice of multivariate extremes* ... (Beirlant et al. 2004)

Basics

Degenerate cdf of block maximum

- $F(X) = Pr(X \leq x)$, cumulative distribution function (cdf)
- $M_n = \max_i\{X_i\}$, **block** maximum
- $Pr(M_n \leq x) = [Pr(X \leq x)]^n$, cdf of maximum
- As $n \uparrow \infty$, $Pr(M_n \leq x)$ becomes **degenerate** ($= 0$ everywhere except at the maximum value of X, x^F)
- What do we do to make $Pr(M_n \leq x)$ useful?

Generalised extreme value distribution

- Try shifting and scaling the random variable to make its tail more stable (this is like the **central limit theorem**)

- $Y_n = a_n^{-1}(max_i\{X_i\} - b_n)$
- $Pr(Y_n \leqslant y) = [Pr(X \leq b_n + a_n y)]^n$
- As $n \uparrow \infty$, $Pr(Y_n \leqslant y)$ is **almost always** well behaved (we have **max-stable** distribution)

$$Pr(Y_n \leqslant y) \quad \rightarrow \quad \exp\{(1 + \xi \frac{y-\mu}{\sigma})_+^{-\frac{1}{\xi}}\} \text{ as } n \rightarrow \infty \text{ for } \xi \neq 0$$

$$(\quad \rightarrow \quad \exp\{\exp(-\frac{y-\mu}{\sigma})\} \text{ when } \xi = 0)$$

- **Generalised extreme value** distribution (GEV)

Domain of attraction

- **All** max-stable distributions converge to the GEV for some value of shape parameter, ξ
  - **Any** max-stable distribution is within the **domain of attraction** (DOA) of GEV for some ξ

- The Weibull distribution converges to GEV with:
  - $\xi = 0$
  - $\bar{F} = kx^{\alpha} exp{-}cx^{\tau}$
  - $a_n = \frac{1}{c\tau}(c^{-1}\log n)^{(1/\tau)-1}$
  - $b_n = (c^{-1}\log n)^{1/\tau}$ to leading order

- Note: this theory is analogous to central limit theorem. There is nothing mysterious here.
  - If you are happy that the mean of random variables with arbitrary distributions converges to a Gaussian $\Rightarrow$ you should be equally happy with GEV for block maxima!

GEV $\Rightarrow$ GP

- Y_n is max-stable, the maximum of n events (i.e. a **block** maximum), each with distribution function F

- So, if n is large enough, $F^n(y) \approx \exp(-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-\frac{1}{\xi}})$
- $n \log F(y) \approx -\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-\frac{1}{\xi}}$ (log both sides)
- $Pr(Y_n > y) = 1 - F(y) \approx \frac{1}{n}\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-\frac{1}{\xi}}$ (Taylor expansion, $\log x = -(1-x)$)
- $Pr(Y_n > y | Y_n > u) = \frac{1-F(y)}{1-F(u)} \approx \left(1+\xi\frac{y-u}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}$ (simple re-arrangement, where $\tilde{\sigma} = \sigma + \xi(u-\mu)$)
- This is the **generalised Pareto** (GP) distribution.

- **Threshold exceedences** from max-stable distributions are GP distributed.
- **Block maxima** from max-stable distributions are GEV distributed

Poisson + GP $\Rightarrow$ GEV

- If occurrence rate of exceedences are Poisson, we can write:

$$
\begin{aligned}
Pr(\text{max in period } \leqslant z) &= \sum_{k=0}^{\infty}(k \text{ storms in period}) F^k(z) \\
&= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \exp(-\lambda) F^k(z) \\
&= \exp(-\lambda(1 - F(z)))
\end{aligned}
$$

- But threshold exceedences are GP-distributed, so:

$$
Pr(\text{max in period } \leqslant z) = \exp(-\lambda \left(1 + \xi \frac{z-u}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}})
$$

- λ is expected number of exceedences, $\tilde{\sigma} = \sigma + \xi(u - \mu)$.
- Set λ to be $\left(1 + \xi \frac{u-\mu}{\sigma}\right)^{-\frac{1}{\xi}}$ (w.l.o.g) $\Rightarrow$ GEV

Take-aways

- We should model tails of distributions with GEV and GP distributions
  - **Threshold exceedences** from max-stable distributions are GP distributed.
  - **Block maxima** from max-stable distributions are GEV distributed
  - Motivation for GEV and GP is **asymptotic theory**
  - We can only justify fitting GEV and GP when we are **clearly in the tail**

- Weibull is a **restricted** choice of distribution for modelling corresponding to $\xi = 0$.
  - Physics (e.g. Miche) tells us that Weibull cannot be correct
  - Weibull might be easier to fit to data (since it is more restricted), but this doesn't necessarily make it better

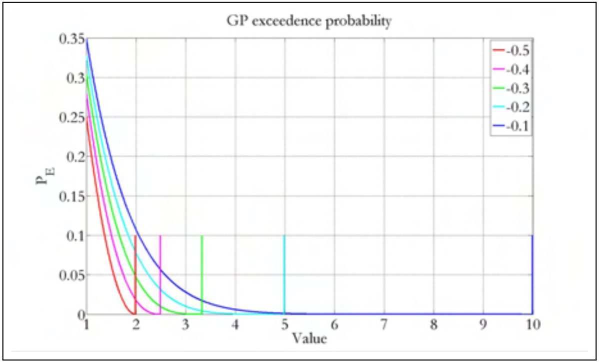
Effect of ξ

- $Pr(X > x|X > u) = (1 + \xi\frac{x-u}{\sigma})^{-\frac{1}{\xi}}$
- If $\xi < 0$, there is a **finite** upper end-point x^F which cannot be exceeded
- If $\xi \geq 0$, the upper end-point $x^F = \infty$

- For ocean waves, observation and physics suggests that x^F is **finite.** e.g Miche:
  - $\frac{1}{2}k_L H_{MAX} = 0.14\pi \tanh(k_L d)$
  - In deep water, Taylor expansion yields $k_L H_{MAX} = 0.8$ limit
  - In shallow water, Taylor expansion yields $\frac{H_{MAX}}{d} = 0.8$ limit

- Weibull distribution has the upper end-point $x^F = \infty$, inconsistent with physics

Effect of $\xi < 0$



- As $\xi \uparrow 0$, then $x_F \uparrow \infty$

Changing threshold

- Consider changing threshold from u to v, $v > u$
- Then $Pr(X > x|X > u) = (1 + \xi\frac{x-u}{\sigma})^{-\frac{1}{\xi}}$

$$
\begin{aligned}
Pr(X > x|X > v) &= \frac{Pr(X > x)}{Pr(X > v)} \\
&= \frac{Pr(X > x|X > u)}{Pr(X > v|X > u)} \\
&= \frac{(1 + \xi\frac{x-u}{\sigma})^{-\frac{1}{\xi}}}{(1 + \xi\frac{v-u}{\sigma})^{-\frac{1}{\xi}}} \\
&= (1 + \xi\frac{x-v}{\sigma + \xi(v-u)})^{-\frac{1}{\xi}}
\end{aligned}
$$

- ξ is **unchanged**, σ **varies linearly** with gradient ξ a.a.f.o. threshold

Return values

- GP or GEV model with parameters ξ, σ, u
- p-year return value x_p is defined by:

$$1 - \frac{1}{p} = \exp\{-\lambda \left(1 + \xi \frac{x_p - u}{\sigma}\right)^{-\frac{1}{\xi}}\}$$

- λ is the expected number of exceedences **per annum.**

- Quantile q of the p-year maximum $x_p(q)$ is defined by:

$$q = \exp(-p\lambda \left(1 + \xi \frac{x_p(q) - u}{\sigma}\right)^{-\frac{1}{\xi}})$$

- $p\lambda$ is the expected number of exceedences **in p years.**

Covariates: outline

- Sample $\{x_i, \theta_i\}_{i=1}^{n}$ of variate x and covariate θ.
- Non-homogeneous Poisson process model for **threshold exceedences**
- Davison and Smith [1990], Davison [2003], Chavez-Demoulin and Davison [2005]
- Rate of occurrence of threshold exceedence and size of threshold exceedence are functionally **independent**.
- Other equivalent interpretations.
- Time, season, space, direction, GCM parameters ...

- Generalised Pareto density (and negative conditional log-likelihood) for **sizes** of threshold excesses:

$$
\begin{aligned}
f(x_i;\xi_i,\sigma_i,u) &= \frac{1}{\sigma_i}(1+\frac{\xi_i}{\sigma_i}(x-u_i))^{-\frac{1}{\xi}-1} \text{ for each } i \\
l_E(\xi,\sigma) &= -\sum_{i=1}^{n} log(f(x_i;\xi_i,\sigma_i,u_i))
\end{aligned}
$$

- Parameters: **shape** ξ, **scale** σ are functions of covariate θ.
- Threshold u set prior to estimation.

- (Negative) Poisson process log-likelihood (and approximation) for **rate of occurrence** of threshold excesses:

$$
\begin{aligned}
l_N(\mu) &= \int_{i=1}^{n} \mu dt - \sum_{i=1}^{n} \log \mu_i \\
\widehat{l}_N(\mu) &= \delta \sum_{j=1}^{m} \mu(j\delta) - \sum_{j=1}^{m} c_j \log \mu(j\delta)
\end{aligned}
$$

- $\{c_j\}_{j=1}^{m}$ counts the number of threshold exceedences in each of m bins partitioning the covariate domain into intervals of length δ
- Parameter: **rate** μ, a function of covariate θ.

- Overall:

$$I(\xi, \sigma, \mu) \quad = \quad I_E(\xi, \sigma) + I_N(\mu)$$

  with all of ξ, σ and μ smooth with respect to t.

- We can estimate μ independently of ξ and σ.

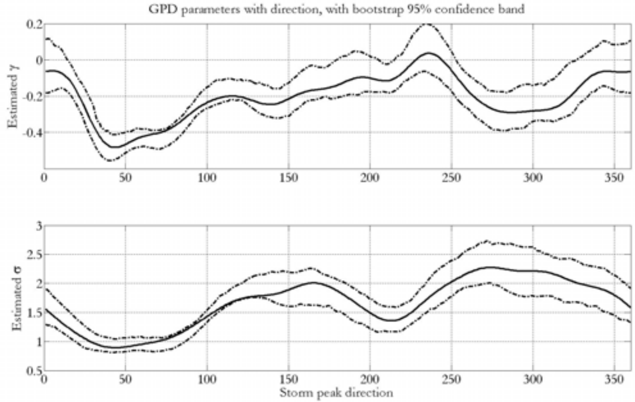
- We can impose smoothness on parameters in various ways.
- In a frequentist setting, we can use **penalised likelihood**:

$$\ell(\theta) \quad = \quad l(\theta) + \lambda R(\theta)$$

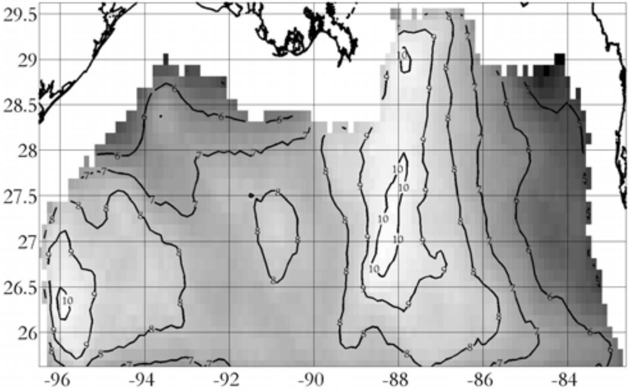
  - $R(\theta)$ is parameter roughness (usually **quadratic form** in parameter vector) w.r.t. covariate θ.
  - λ is roughness tuning parameter
- In a Bayesian setting, we can impose a **random field prior** structure (and corresponding posterior) on parameters:

$$\begin{aligned} f(\theta|\alpha) &= \exp\{-\alpha \sum_{i=1}^{n} \sum_{\theta_j \text{ near } \theta_i} (\theta_i - \theta_j)^2\} \\ \log f(\xi, \sigma|x, \alpha) &= l(\xi, \sigma, \mu|x) \\ &- \sum_{i=1}^{n} \sum_{\theta_j \text{ near } \theta_i} \{\alpha_\xi(\xi_i - \xi_j)^2 + \alpha_\sigma(\sigma_i - \sigma_j)^2\} \end{aligned}$$

Covariates: applications



Fourier directional model for GP shape and scale at Northern North Sea location, with 95% bootstrap confidence band.



Spatial model for 100-year storm peak significant wave height in the Gulf of Mexico (not to scale), estimated using a **thin-plate spline** with directional pre-whitening.

Multivariate: outline

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 will probably not be a sample point!

- $P(M \leqslant x) = \prod_{j=1}^{p} P(X_j \leqslant x_j) = F^n(x)$
  - We 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 of exponent measure implies asymptotic dependence (or perfect independence)!

Composite likelihood for spatial dependence

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

$$
\begin{aligned}
l_C(\theta) &= -\sum_{i=1}^{n}\sum_{j=1}^{n}\log f(z_i, z_j; \theta) \\
f(z_i, z_j) &= \left(\frac{\partial V(z_i, z_j)}{\partial z_i}\frac{\partial V(z_i, z_j)}{\partial z_j} - \frac{\partial^2 V(z_i, z_j)}{\partial z_i \partial z_j}\right)e^{-V(z_i, z_j)}
\end{aligned}
$$

- Exponent measure has simple bivariate parametric form, e.g. :

$$
V(z_i, z_j) = \left(\frac{1}{z_i} + \frac{1}{z_j}\right)\left(1 - \frac{\alpha(h)}{2}\left(1 - \left(1 - 2\frac{(\rho(h)+1)z_i z_j}{z_i^2 + z_j^2}\right)^2\right)\right)
$$

with two pre-specified functions α and ρ of distance h whose parameters must be estimated.

- Component-wise maxima has some pros:
  - Most widely-studied branch of multivariate extremes.
  - Composite likelihood offers some promise, but is itself an approximation.
- And many cons:
  - Hotch-potch of methods.
  - Does not accommodate asymptotic independence.
  - Threshold selection!
  - Covariates!
- Parametric forms.

Extremal dependence

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

- Extremal dependence models:
    - Admit asymptotic independence.
- But have issues with:
    - Threshold selection.
    - Covariates!

- Bingham et al. [1987]

- (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: $Pr(X_G < g) = \exp(-e^{-g})$.

- Ledford and Tawn [1996] 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:

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

- 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:

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

Conditional extremes

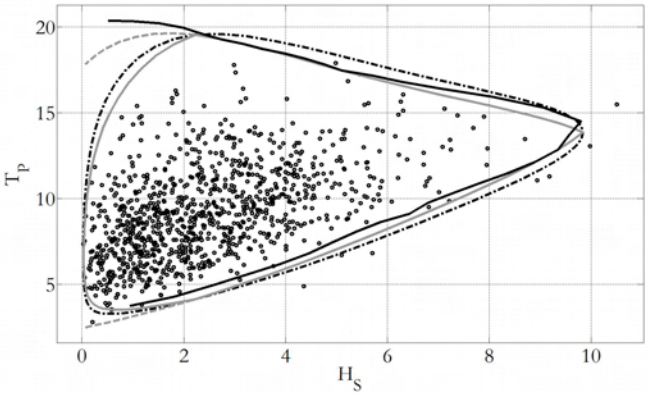
- Heffernan and Tawn [2004]

- Sample $\{x_{i1}, x_{i2}\}_{i=1}^{n}$ of variate X_1 and X_2.
- (X_1, X_2) need to be transformed to (Y_1, Y_2) on the same **standard Gumbel** scale.
- Model the **conditional** distribution of Y_2 given a large value of Y_1.
- **Asymptotic** argument relies on X_1 (and Y_1) being **large**.

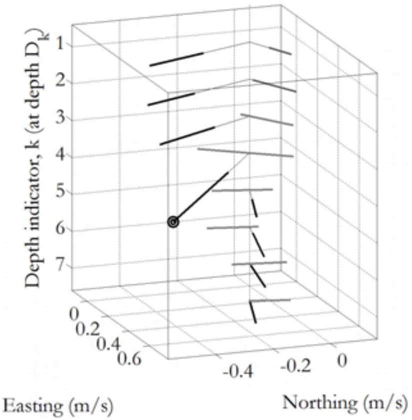
- Applies to almost all known forms of multivariate extreme value distribution, but not all.

- $(X_1, X_2) \overset{PIT}{\Rightarrow} (Y_1, Y_2)$.
- $(Y_2 | Y_1 = y_1) = a y_1 + y_1^b Z$ for large values y_1 and +ve dependence.
- Estimate a, b and Normal approximation to Z using regression.
- $(Y_1, Y_2) \overset{PIT}{\Rightarrow} (X_1, X_2)$.
- Simulation to sample joint distribution of (Y_1, Y_2) (and (X_1, X_2)).
- Pros:
    - Extends naturally to high dimensions
    - c.f. copulas
- Cons:
    - Threshold selection for (large number of) models.
    - Covariates!
    - Consistency of $Y_2 | Y_1$ and $Y_1 | Y_2$ not guaranteed.

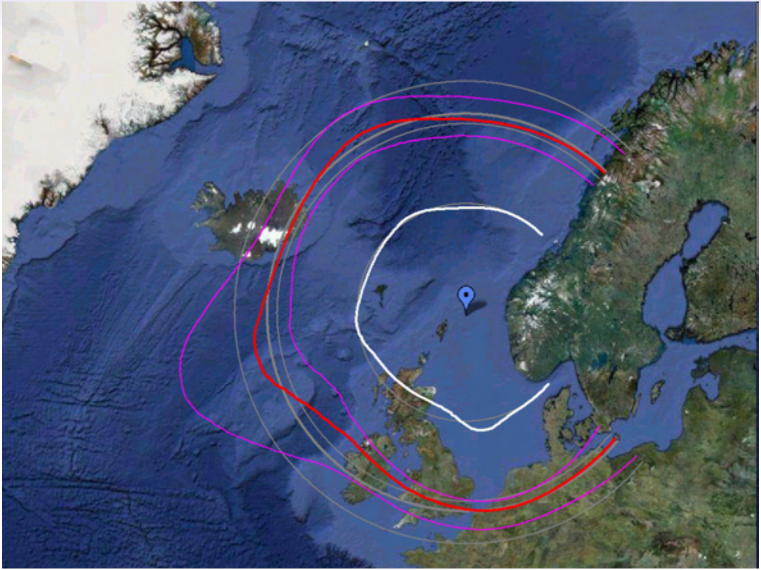
Multivariate: applications



Environmental **design contours** derived from a conditional extremes model for storm peak significant wave height, H_S, and corresponding peak spectral period, T_P.

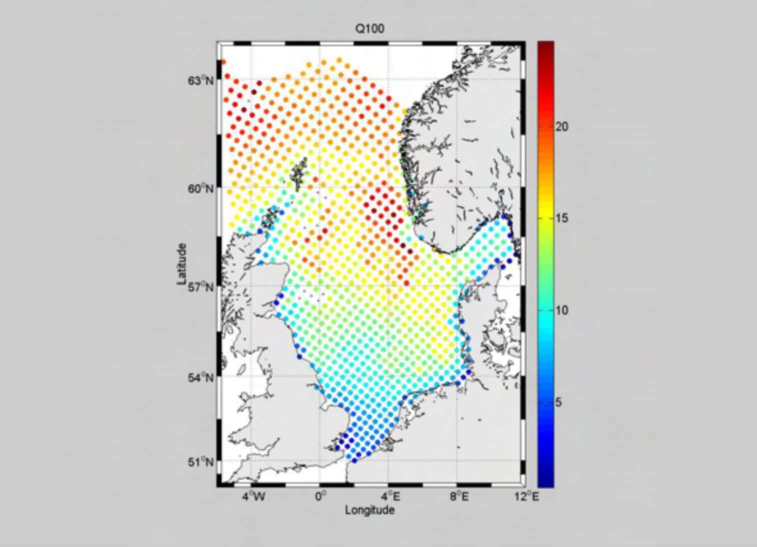


Current profiles with depth (a 32-variate conditional extremes analysis) for a North-western Australia location.

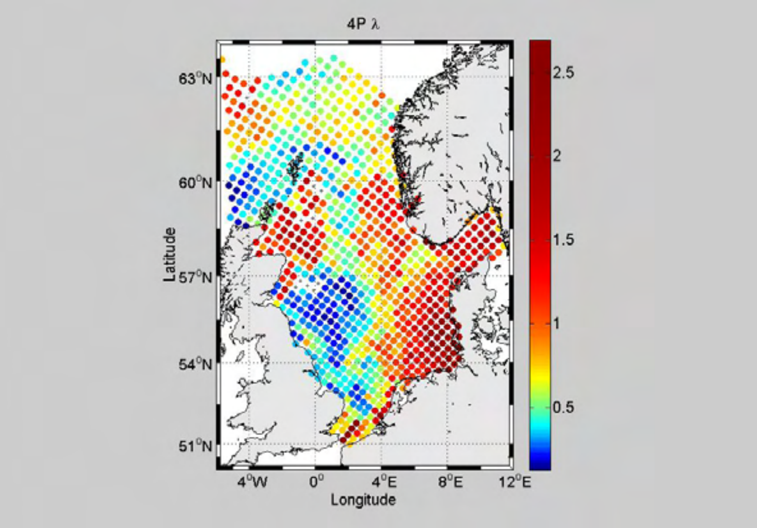


Fourier **directional** model for conditional extremes at a Northern North Sea location.

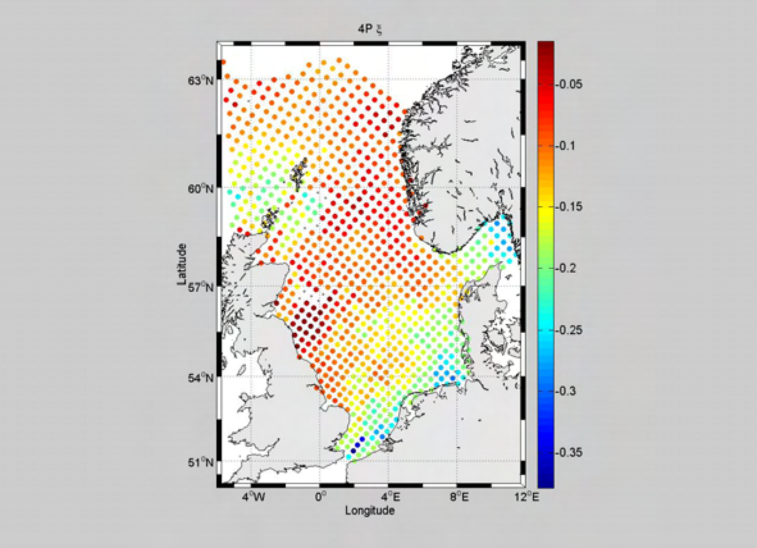
Current developments



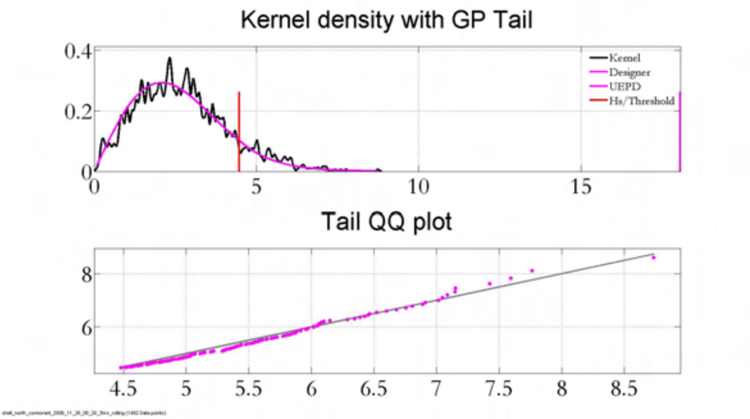
Extreme quantiles from **Bayesian** model incorporating scale uncertainty via a **Box-Cox** transformation, point-wise for North Sea.



Box-Cox scale λ, point-wise for North Sea.



Generalised Pareto shape, point-wise for North Sea.



A **Weibull-GP** model for the distribution of waves in **shallow water.**

- **p-spline** approaches to spatio-temporal and spatio-directional extreme value models.
  - Easy specification of multi-covariate roughness.
- **Composite likelihood** approaches to (asymptotically dependent) joint extremes.
- **Laplace approximation** as alternative to MCMC.
- **Statistical down-scaling** to estimate climate change effects on structural safety.
- **Mixture modelling** for elimination of threshold selection

Thanks

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

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.

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

A. C. Davison. *Statistical models*. Cambridge University Press, 2003.

A.C. Davison and R. L. Smith. Models for exceedances over high thresholds. *J. R. Statist. Soc. B*, 52:393, 1990.

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

A. W. Ledford and J. A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83:169–187, 1996.