Modelling extreme environments

Philip Jonathan
Shell Technology Centre Thornton, Chester, UK
Lancaster University, UK

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

Robust12
Němčičky, Září 2012

Acknowledgements

- Kevin Ewans
- Kaylea Haynes
- David Randell
- Yanyun Wu

Outline

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

Review (Jonathan and Ewans) at *www.lancs.ac.uk/∼jonathan*.

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.

Praha 1872.

Praha 2002.

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.
  - Network traffic.

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.
  - estimated using e.g. extremal index (Ledford and Tawn 2003)
- **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 $\Leftrightarrow$ multivariate 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 (with +ve, -ve association).
  - 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)

Covariates: outline

- Sample $\{x_i, t_i\}_{i=1}^n$ of variate x and covariate t.
- 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 ...

Quantile regression models threshold

- Data $\{\theta_i, x_i\}_{i=1}^n$, τ^{th} conditional quantile $\psi(\tau, \theta)$.

  Fourier basis:

  $$\psi(\tau, \theta) = \sum_{k=0}^{p} \alpha_{c\tau k} \cos(k\theta) + \alpha_{s\tau k} \sin(k\theta) \text{ and } \alpha_{s\tau 0} \triangleq 0$$

  Spline basis:

  $$\psi(\tau, \theta) = \sum_{k=0}^{p} \Phi_{\theta k} \beta_{\tau k}$$

- Estimated by minimising **penalised** criterion Q_τ^* with respect to basis parameters (α or β):

  $$Q_\tau^* = \{\tau \sum_{r_i \geq 0}^{n} |r_i| + (1 - \tau) \sum_{r_i < 0}^{n} |r_i|\} + \lambda R_{\psi\tau}$$

  for $r_i = x_i - \psi(\tau, \theta_i)$ for $i = 1, 2, ..., n$, and **roughness** $R_{\psi\tau}$.

GP models size of threshold exceedances

- 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.
- Threshold u set prior to estimation.

Poisson models rate of threshold exceedances

- (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

- Overall:

$$l(\xi, \sigma, \mu) \quad = \quad l_E(\xi, \sigma) + l_N(\mu)$$

  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**:

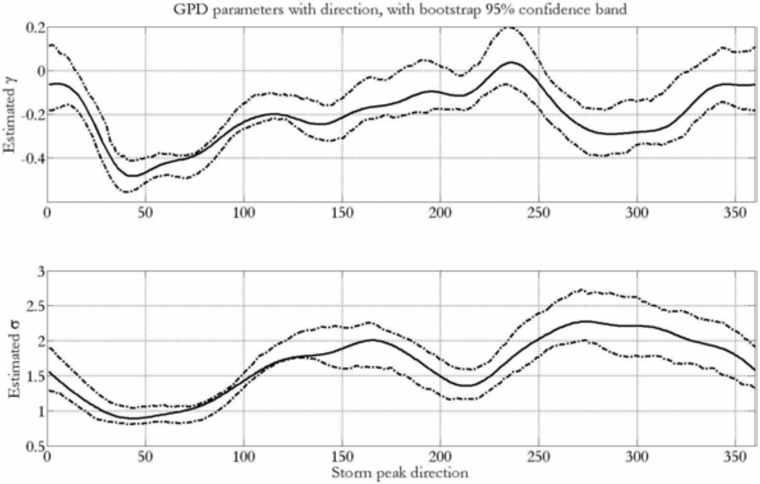
$$\ell(\theta) \quad = \quad l(\theta) + \lambda R(\theta)$$

  - $R(\theta)$ is parameter roughness (usually quadratic form in parameter vector)
  - λ is roughness tuning parameter

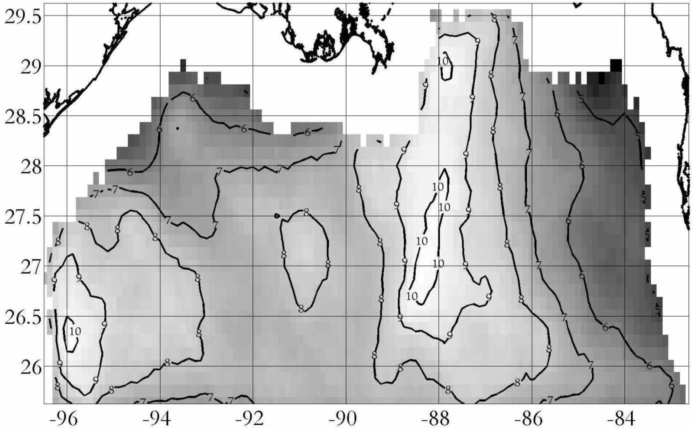
- In a Bayesian setting, we can impose a **random field prior** structure (and corresponding posterior) on parameters:

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

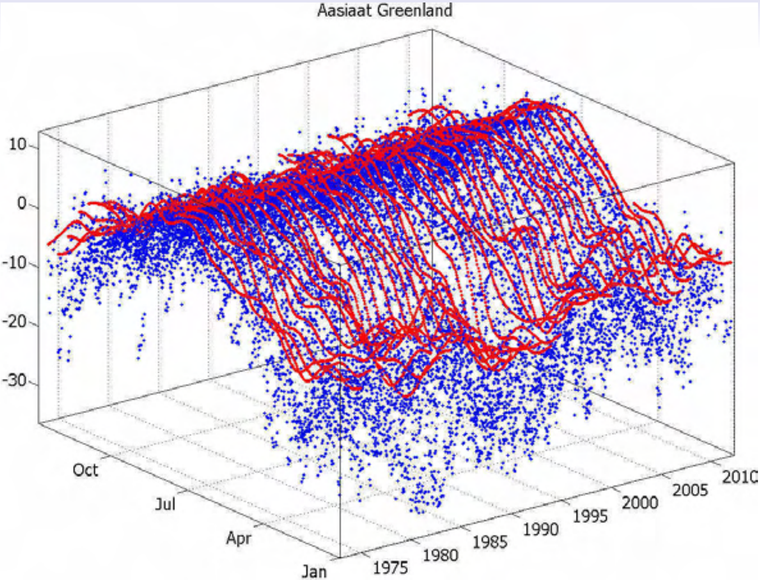
Covariates: applications



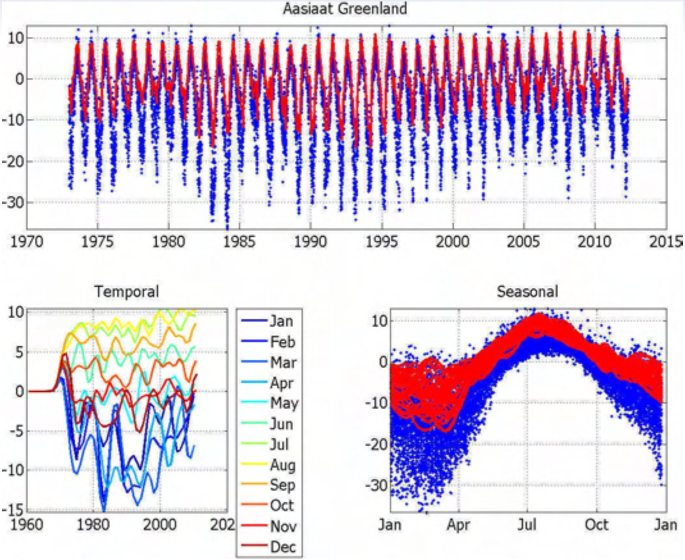
Fourier directional model for GP shape and scale at Northern North Sea location, with 95% bootstrap confidence band.



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.



Seasonal-temporal model of 90%ile of air temperature at Greenland location using spline quantile regression.



Seasonal-temporal model of 90%ile of air temperature at Greenland location using spline quantile regression.

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:

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

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

Smith model

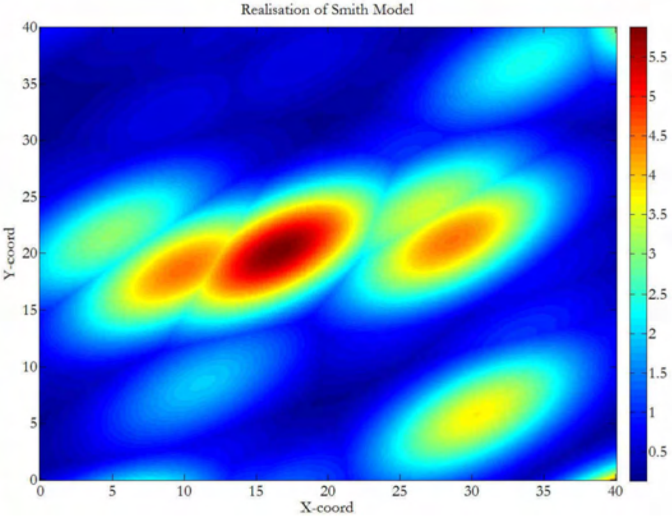
$$V(z_i, z_j) = \frac{1}{z_i}\Phi\left(\frac{\alpha(h)}{2} + \frac{1}{\alpha(h)}\log\left(\frac{z_j}{z_i}\right)\right) + \frac{1}{z_j}\Phi\left(\frac{\alpha(h)}{2} + \frac{1}{\alpha(h)}\log\left(\frac{z_i}{z_j}\right)\right)$$

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.

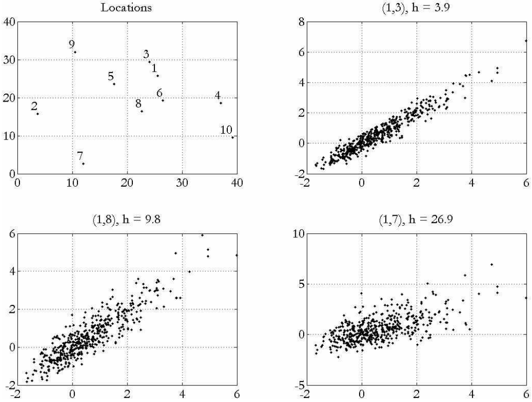
Realisation from Smith model



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

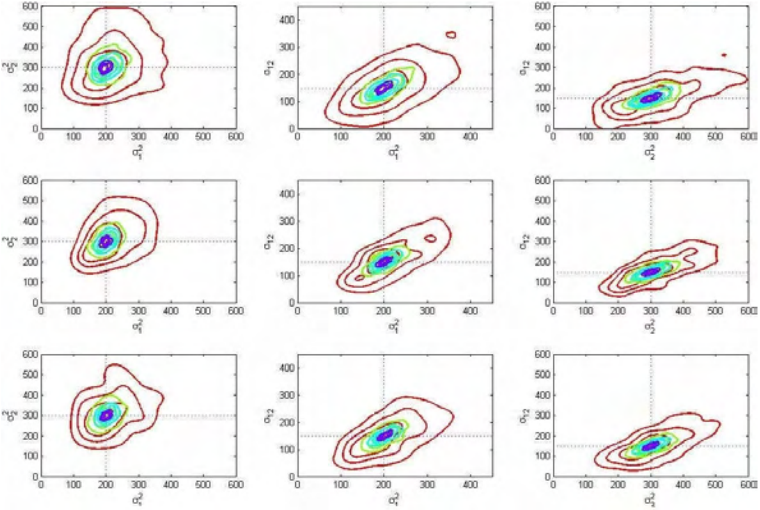
Simulation from Smith model

Simulated samples of size $N = 10$, 50, 100 and 500 corresponding to $K = 10$, 50 and 100 spatial locations, for $\sigma_1^2 = 200$, $\sigma_{12} = 150$ and $\sigma_2^2 = 300$ with standard Frechet marginals. Locations at random on 40×40 grid.



Sample size $N = 500$, $K = 10$ locations.

Maximum composite likelihood estimates



25%, 50% and 75% percentiles of MCLE estimates for $N = 10$ (Red), 50 (Green), 100 (Turquoise) and 500 (Purple) observations over $K = 10$ (Top), 50 (Centre), and 100 (Bottom) sites.

- Component-wise maxima has some pros:
  - Most widely-studied branch of multivariate extremes.
  - Composite likelihood offers some promise; Bayesian inference feasible.
- 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!

- Ideas from theory of **regular variation** (see 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:

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

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

η is known as the **coefficient of tail dependence**.

- 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:

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

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.

- $(X_1, X_2) \overset{PIT}{\Rightarrow} (Y_1, Y_2)$.
- $(Y_2 | Y_1 = y_1) = ay_1 + y_1^b Z$ 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
- Cons:
  - Threshold selection for (large number of) models.
  - Covariates!
  - Consistency of $Y_2 | Y_1$ and $Y_1 | Y_2$ not guaranteed.

Conditional extremes with covariates

On Gumbel scale, by analogy with Heffernan & Tawn (2004) we propose the following conditional extremes model:

$$(Y_k | Y_j = y_j, \Phi = \phi) = \alpha_\phi y_j + y_j^{\beta_\phi}(\mu_\phi + \sigma_\phi Z) \text{ for } y_j > \psi_j^G(\theta_j, \tau_{j*}^G)$$

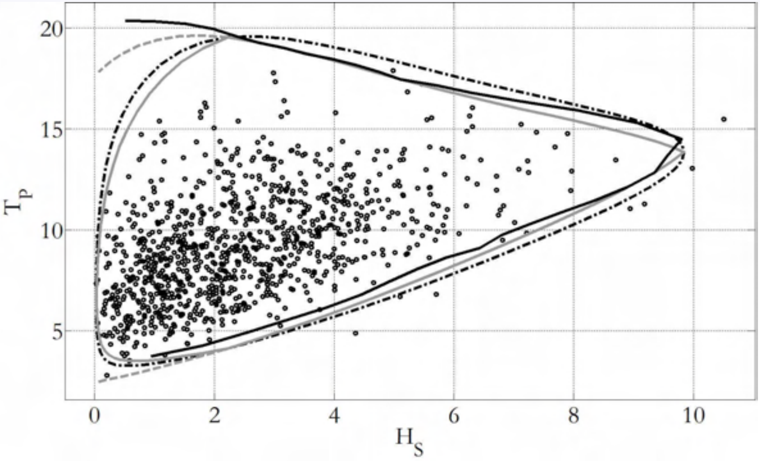
where:

- $\psi_j^G(\theta_j, \tau_{j*}^G)$ is a high directional quantile of Y_j on Gumbel scale, above which the model fits well
- $\alpha_\phi \in [0,1]$, $\beta_\phi \in (-\infty, 1]$, $\sigma_\phi \in [0, \infty)$
- Z is a random variable with **unknown** distribution G
- Z will be assumed to be approximately Normally distributed for the purposes of parameter estimation

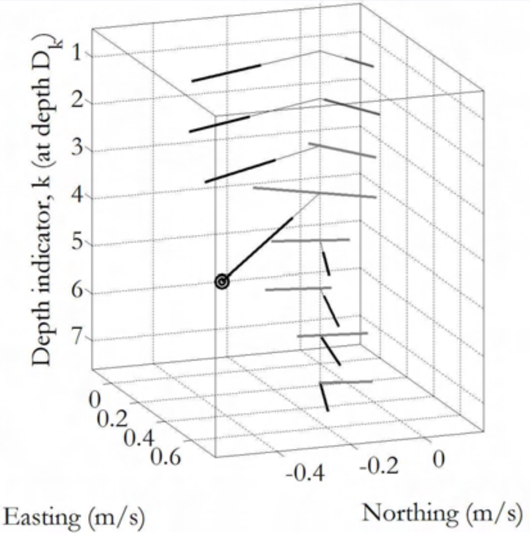
Settings:

- In a (H_S, T_P) case, $\phi \triangleq \theta_j \triangleq \theta_k$, and dependence is assumed a function of absolute covariate
- In a $(H_S, \mathit{WindSpeed})$ case, $\phi = \theta_k - \theta_j$, and dependence is assumed a function of relative covariate

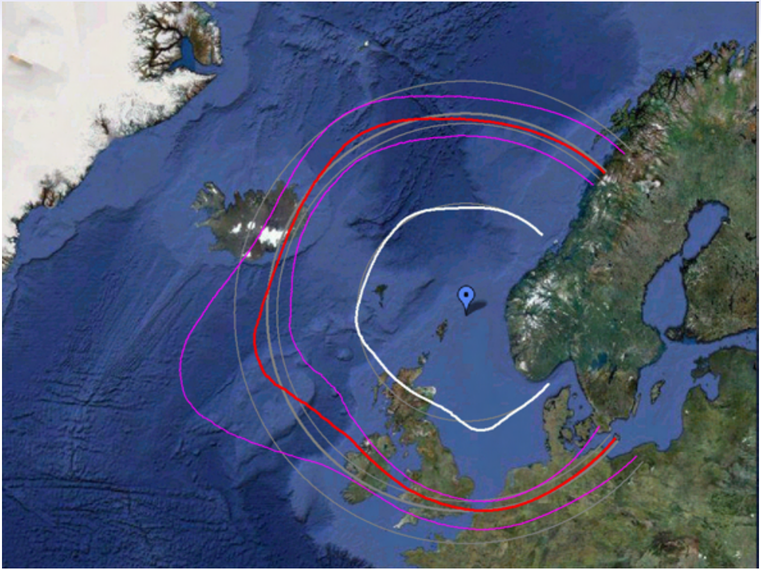
Multivariate: applications



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

- **p-spline** and **random field** approaches to spatio-temporal and spatio–directional extreme value models.
- **Composite likelihood**: model (asymptotically dependent) componentwise–maxima.
- **Censored likelihood**: allows extension from block-maxima to threshold exceedances.
- **Hybrid spatial dependence model**: incorporation of asymptotic independence using inverted multivariate extreme value distribution.

Děkuji za pozornost!

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

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

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.

P. Jonathan and K. C. Ewans. Statistical modelling of extreme ocean environments for marine design. *Ocean Eng. ∼2013, under review, draft at www.lancs.ac.uk/∼jonathan*.

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

A. W. Ledford and J. A. Tawn. Diagnostics for dependence within time series extremes. *J. Roy. Statist. Soc. B*, 65:521–543, 2003.