COVARIATE EFFECTS IN MARGINAL AND CONDITIONAL EXTREMES

Philip Jonathan, **David Randell**, Kevin Ewans, Graham Feld
Shell Statistics and Chemometrics

Motivation: extremes in met-ocean

- **Rational** and **consistent** design an assessment of **marine structures**:
    - Reduce bias and uncertainty in estimation of **return values**
- Non-stationary **marginal** and **conditional** extremes:
    - Multiple locations, multiple variables, time-series
    - **Multidimensional** covariates
- Improved **understanding** and **communication** of risk:
    - Incorporation within **well-established** engineering design practices
    - **"Knock-on" effects** of "improved" inference
    - New and existing structures

- Other current applications in Shell:
    - Geophysics: seismic hazard assessment
    - Asset integrity: corrosion & fouling

Extremes in met-ocean: **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 $f(X)$? (Harris 2004)
- **Threshold** estimation:
  - Scarrott and MacDonald [2012]
- **Parameter** estimation
- **Measurement** issues:
  - Field measurement uncertainty greatest for extreme values
  - Hindcast data are simulations based on pragmatic physics, calibrated to historical observation

Extremes in met-ocean: **multivariate** challenges

- **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)
  - Point process / multivariate GP process
- **Extremal dependence**: (Ledford and Tawn 1997)
  - Assumes regular variation of joint survivor function
  - Yields more general forms of extremal dependence
  - $\Rightarrow$ Asymptotic dependence, asymptotic independence (with +ve, -ve association), "hidden regular variation"
  - "Ray" extensions
  - 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
  - Extensions

Extremes in met-ocean: illustrations

- This talk is about statistical modelling

- Illustration 1: **Marginal** directional-seasonal (with a twist)
- Illustration 2: Marginal spatio-directional
- Illustration 3: Directional **conditional**

Illustration 1: **directional-seasonal**

- Marginal model for **single** North Sea location
- Response is **storm peak significant wave height,** H_S^{sp}
- Wave climate is dominated by **extra-tropical storms**
- Directional and seasonal variability in extremes present:
  - **Fetch** variability (Atlantic, Norwegian Sea, North Sea)
  - Land shadows (Norway, UK)
  - **Winter** storms more energetic

- **Within-storm** evolution of significant wave height, H_S in time given H_S^{sp}
- Distributions for extreme **wave height**, **crest elevation** and **surge** given H_S

- Sample of **hindcast** storms for period of ≈50 years
- Animation: Link

Directional variability

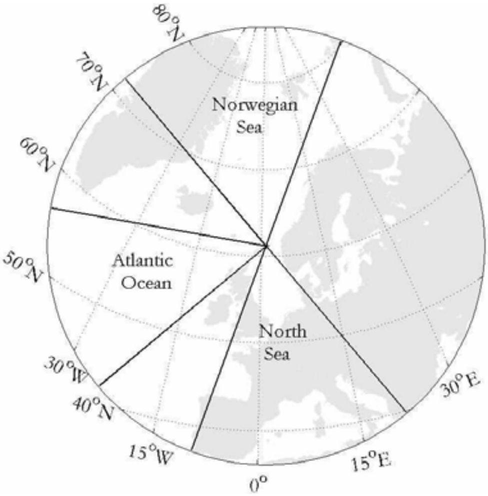


Figure: Fetch and land shadows

Storm peak significant wave height, H_S^{sp}

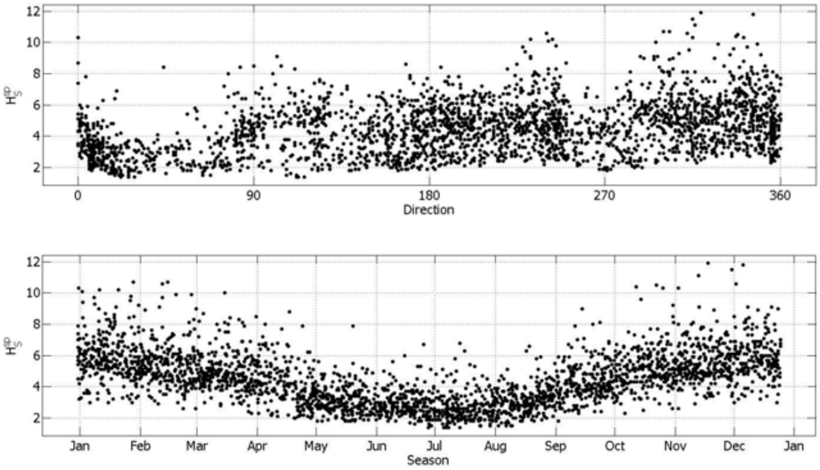


Figure: Storm peak significant wave height H_S^{sp} on storm direction θ^{sp} (upper panel) and storm season ϕ^{sp} (lower panel)

Quantiles of H_S^{sp}

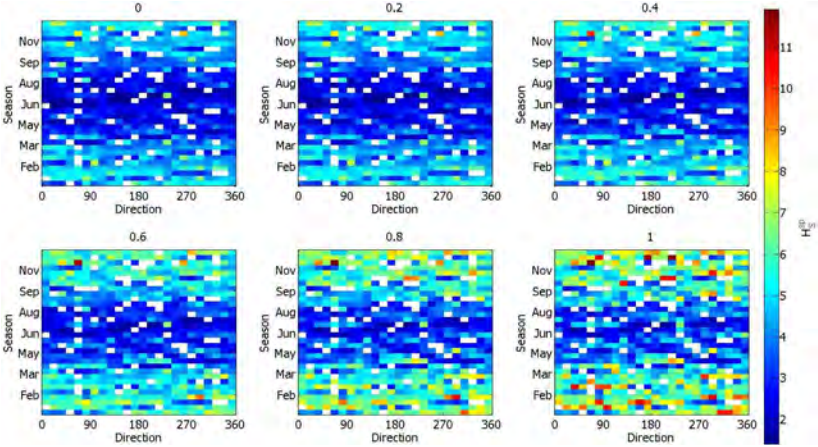


Figure: Empirical quantiles of storm peak significant wave height, H_S^{sp} by storm direction, θ^{sp}, and storm season, θ^{sp}. Empty bins are coloured white

Storm model **zoals Johan**

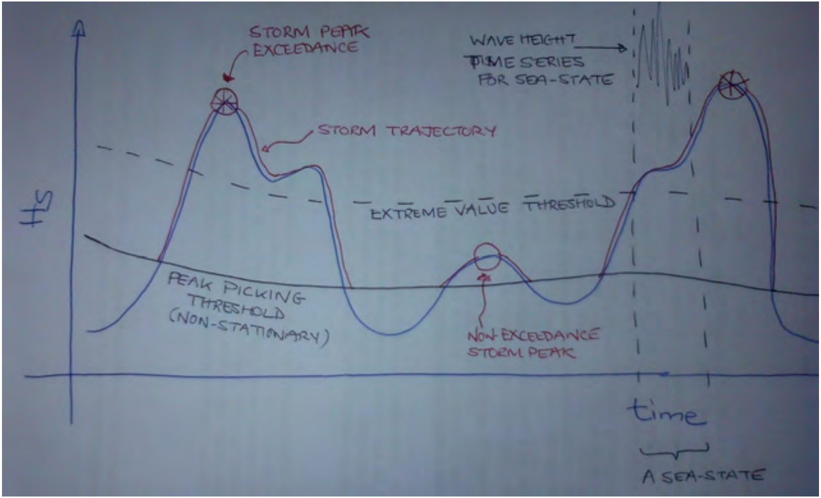


Figure: $H_S \approx 4\times$ standard deviation of ocean surface profile at a location corresponding to a specified period (typically three hours)

Storm trajectories of significant wave height, H_S

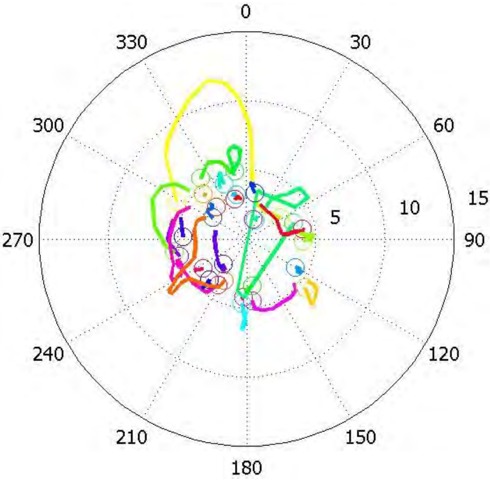


Figure: Storm trajectories of significant wave height, H_S on wave direction θ for 30 randomly-chosen storm events (in different colours). A circle marks the start of each intra-storm trajectory

Model components

- Sample $\{\dot{z}_i\}_{i=1}^{\dot{n}}$ of $\dot{n}$ **storm peak** significant wave heights observed with storm peak directions $\{\dot{\theta}_i\}_{i=1}^{\dot{n}}$ and storm peak seasons $\{\dot{\phi}_i\}_{i=1}^{\dot{n}}$

- Model components (all non-stationary w.r.t θ, ϕ):
  1. **Threshold** function ψ 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 ρ
  3. **Size of occurrence** of threshold exceedance using generalised Pareto (GP) model with shape and scale parameters ξ and σ

(Drop *sp* superscripts where convenient)

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
- Smooth functions of covariates estimated using penalised B-splines (Eilers and Marx 2010)
- Large number of parameters to estimate:
  - Slick linear algebra (c.f. generalised linear array models, Currie et al. 2006)
  - Efficient optimisation

Penalised B-splines

- Physical considerations suggest model parameters ψ, ρ, ξ and σ vary smoothly with covariates θ, ϕ
- Values of $(\eta =)\psi, \rho, \xi$ 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:

$$B = B_\theta \otimes B_\phi$$

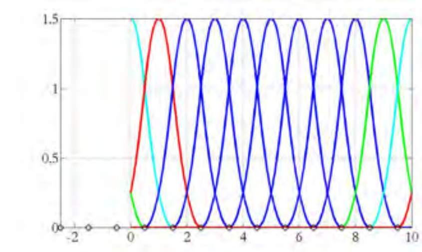
  (**exact** operations calculated without explicit evaluation)
- Roughness R_η defined as:

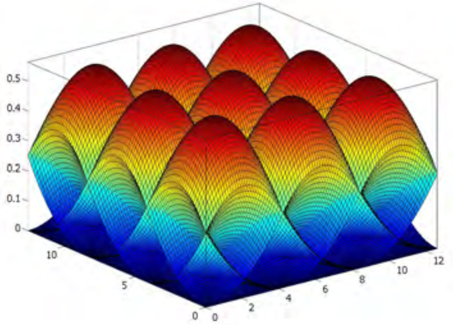
$$R_\eta = \beta'_\eta P \beta_\eta$$

  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** or similar chooses λ optimally
- Alternatives: random fields, Gaussian processes, ...





Quantile regression model for extreme value threshold

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

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

  for $r_i = z_i - \psi(\theta_i, \phi_i; \tau)$ for $i = 1, 2, ..., n$, and **roughness** R_ψ controlled by roughness coefficient λ_ψ
- (Non-crossing) quantile regression formulated as linear programme (Bollaerts et al. 2006)
- λ_ψ estimated using cross validation or similar

Directional-seasonal threshold, ψ

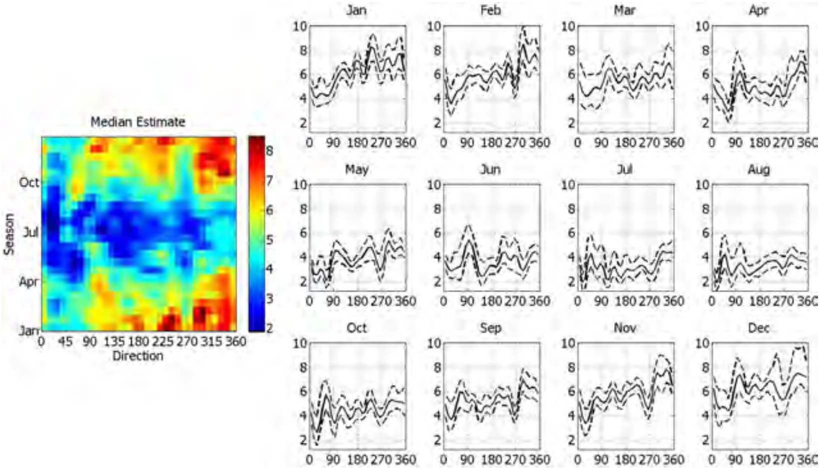


Figure: LHS: bootstrap median. RHS: 12 monthly directional

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):

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

- $\{c_j\}_{j=1}^m$ 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 or similar

Directional-seasonal exceedance rate, ρ

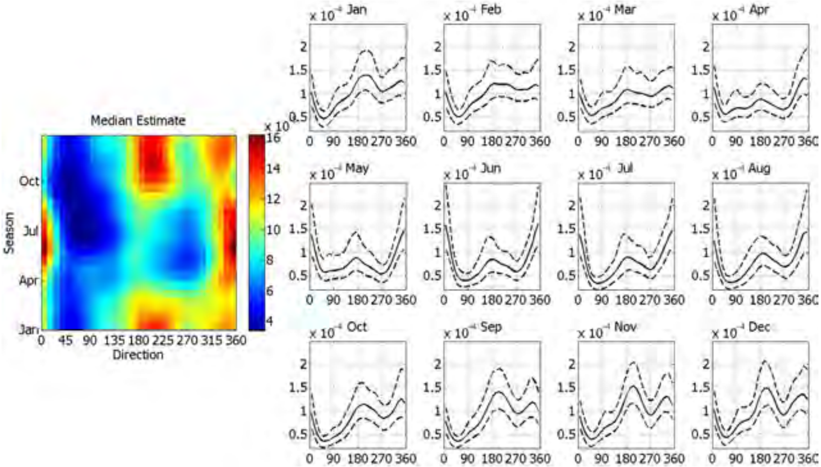


Figure: LHS: bootstrap median. RHS: 12 monthly directional

GP 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:

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

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

Directional-seasonal parameter plot for GP shape, ξ

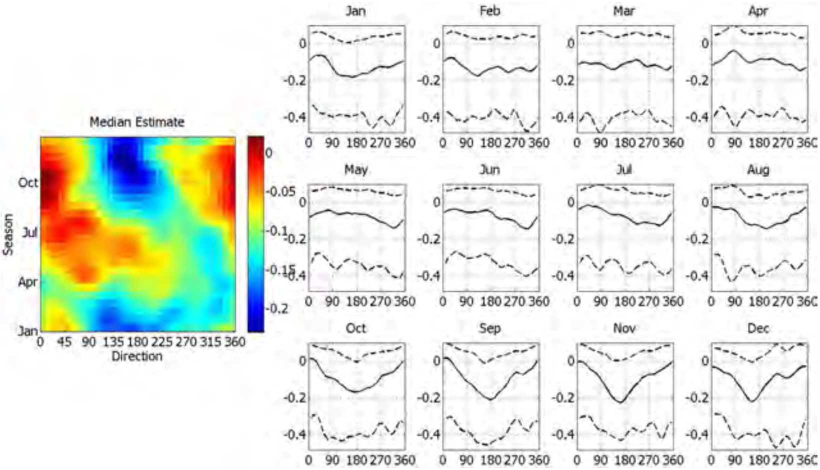


Figure: LHS: bootstrap median. RHS: 12 monthly directional

Directional-seasonal parameter plot for GP scale, σ

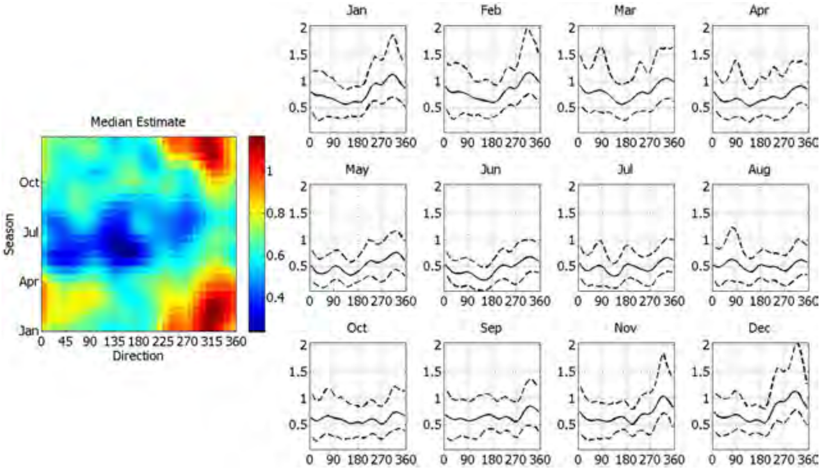


Figure: LHS: bootstrap median. RHS: 12 monthly directional

Return values

- Estimation of return values by simulation under the model
  - Number of events in period
  - Directions and seasons of each event
  - Size (or magnitude) of each event
  - H_{S100} is the maximum value of H_S^{sp} in a simulation period of 100–years

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

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

  - Implementation and interpretation **problematic**

CDFs for H_{S100}

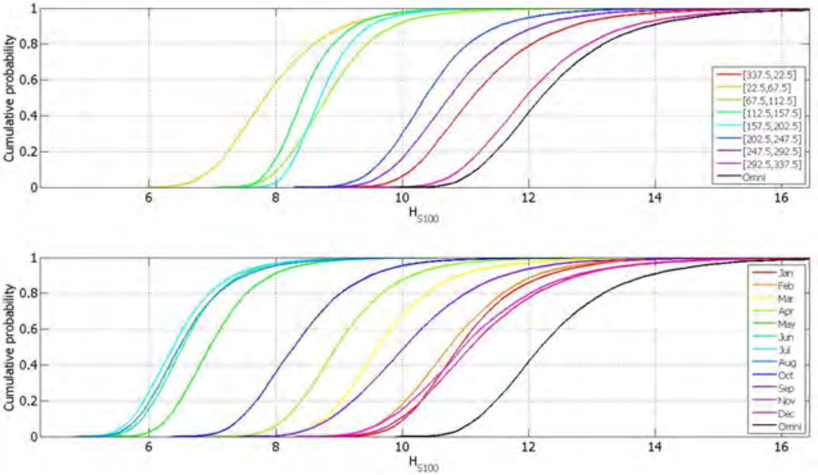


Figure: CDFs incorporating bootstrap uncertainty

Directional-seasonal return value plot for H_{S100}

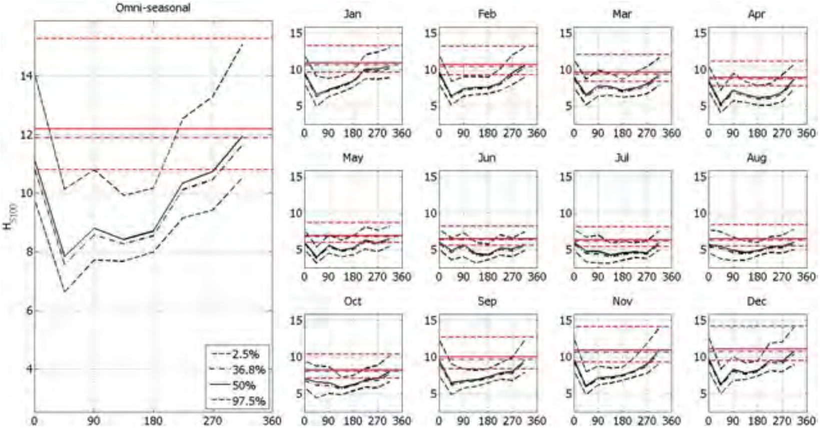


Figure: LHS: directional omni-seasonal return values. RHS: directional return values for calendar months

Directional-seasonal return value plot for H_{S100}

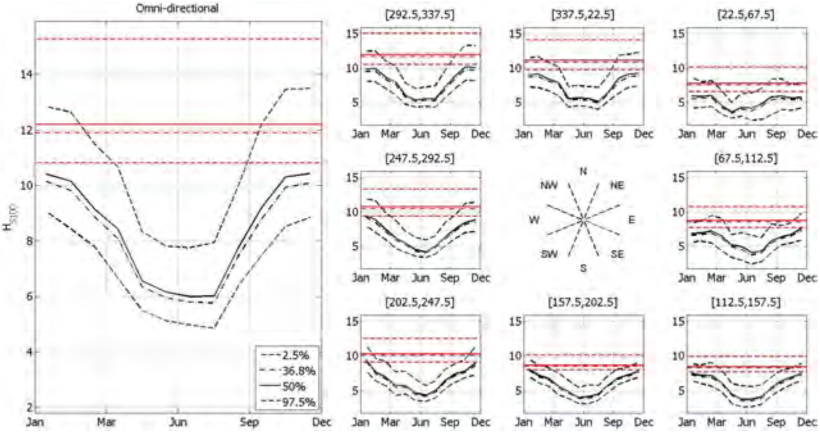


Figure: LHS: seasonal omni-directional return values. RHS: seasonal return values for directional octants

Within-storm variability

Figure: Cormorant Alpha platform in North Sea

Within-storm variability

Critical environmental variables:

- Storm peak significant wave height:
  - (Sea state) significant wave height
  - Maximum wave height
  - Maximum crest elevation
  - Peak total water level ($\approx$ wave + **surge** + tide)
- "Associated" values of wind speed and direction corresponding to peak significant wave height:
  - Maximum conditional structural loads and responses
  - Conditional extremes

Estimating within-storm variability

- Extreme value model allows simulation of H_S^{sp}, θ^{sp} and ϕ^{sp}
- Matching procedure used to estimate storm evolution $(H_S(t), \theta(t), \phi(t)) | (H_S^{sp}, \theta^{sp}, \phi^{sp})$ for sea state t
  - Essential in estimating return values for covariate bins other than that containing the storm peak
  - Opportunity for empirical modelling
- Empirical (physics-motivated) literature models for $H(t)|H_S(t)$ and $H_{max}(t)|H_S(t)$

The cumulative distribution function for the maximum wave height H_{max} in a sea-state of n_s waves with significant wave height $H_S = h_s$ is taken (see, for example, Forristall 2000, Prevosto et al. 2000) to be given by:

$$P(H_{max} \leq h_{max} | H_S = h_s, M = n_s) = (1 - \exp(-\frac{1}{\beta}(\frac{h_{max}}{h_s/4})^{\alpha}))^{n_s}$$

with $\alpha = 2.13$ and $\beta = 8.42$. The number of waves n_s in a particular sea state is estimated by dividing the length of the sea-state (in seconds) by its zero-crossing period, T_Z

Directional-seasonal return value plot for H_{max100}

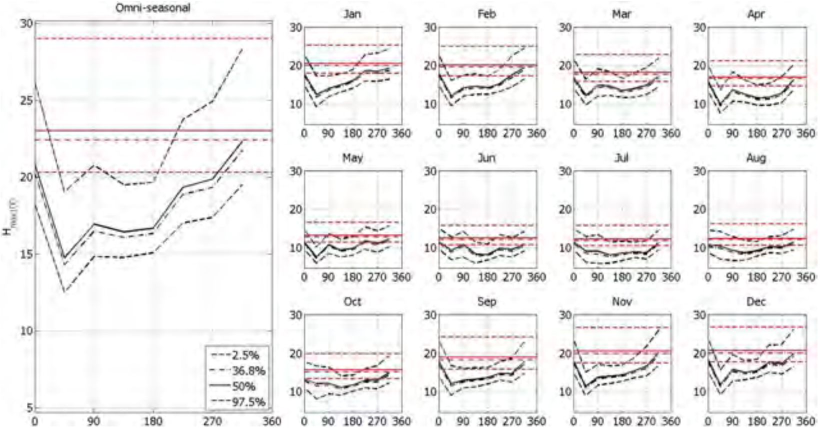


Figure: LHS: directional omni-seasonal return values. RHS: directional return values for calendar months

Validation of model for (within-storm) H_S

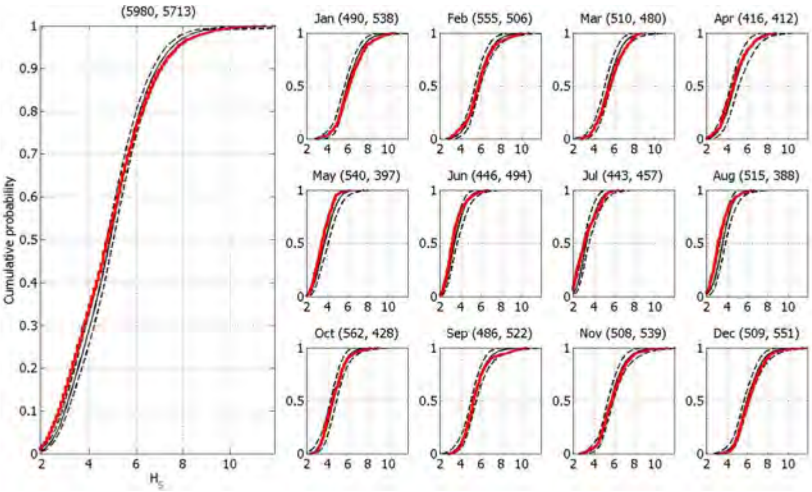


Figure: CDFs for H_S for original sample and for 1000 sample realisations under the model corresponding to the same time period as the original sample

Illustration 2: **spatio-directional** (briefly)

Figure: Katrina

Illustration 2: **spatio-directional**

- Longitude, latitude and direction as covariates
  - Physics: direction and season correlated
  - Gulf of Mexico (GoM), North West Shelf of Australia (NWS) applications here
- Marginal per location
- Estimation of spatial smoothness
  - Sample is spatially dependent
  - Vertical adjustment / sandwich estimator
  - Bootstrap

GoM spatio-directional H_S^{sp}

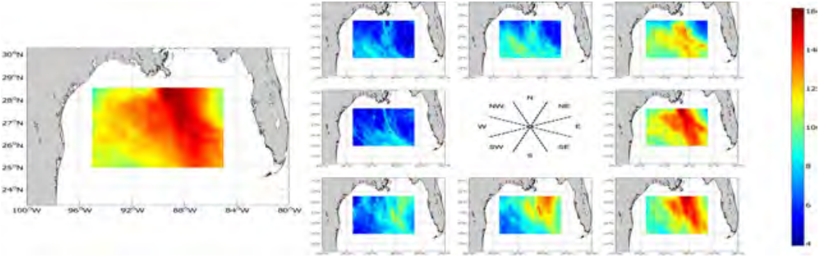


Figure: $\approx$ 17000 locations $\times$ 32 directional bins for Gulf of Mexico. Plot for quantile (withheld) of 100-year maximum storm peak significant wave height, H_S^{sp}

NWS spatio-directional H_S^{sp}

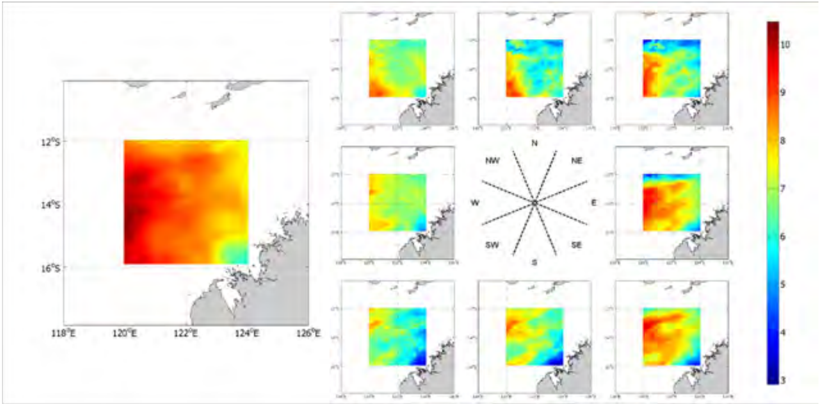


Figure: North West Shelf of Australia. See Jonathan et al. [2014]

Illustration 3: directional **conditional**

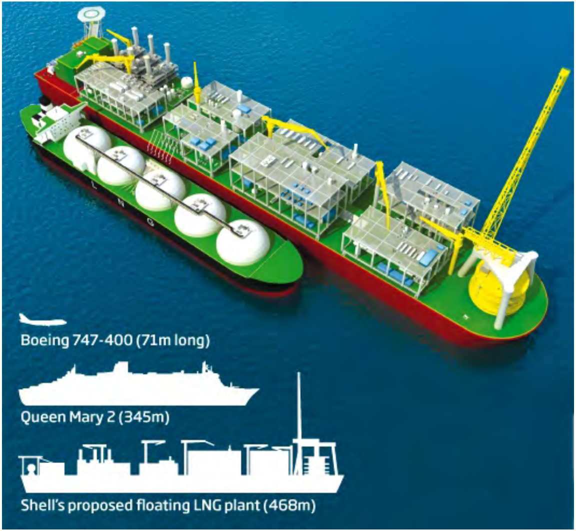


Figure: Floating LNG tanker

Floating LNG tanker

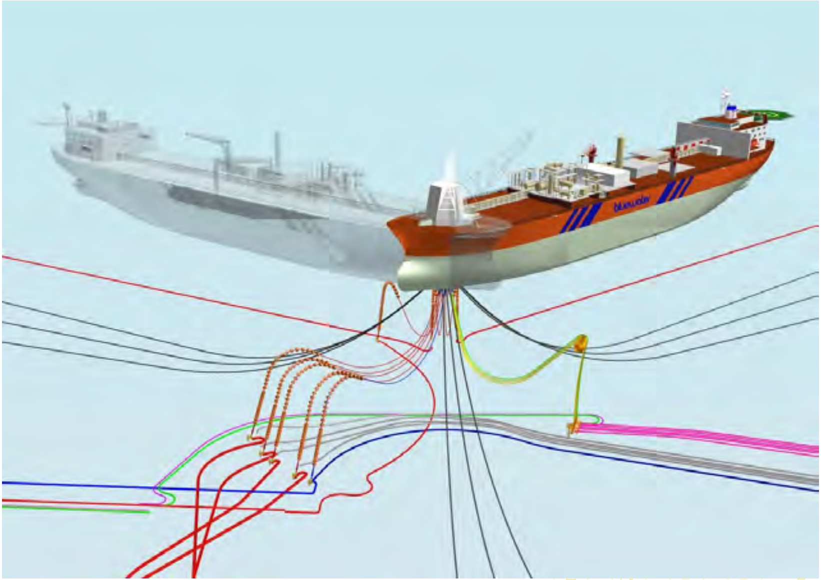


Illustration 3: directional **conditional**

Problem structure:

- Bivariate sample $\{\dot{x}_{ij}\}_{i=1,j=1}^{n,2}$ of random variables $\dot{X}_1, \dot{X}_2$
- Covariate values $\{\theta_{ij}\}_{i=1,j=1}^{n,2}$ associated with each individual

- For some choices of variables $\dot{X}$, e.g. $\dot{X}_1 = H_S$, $\dot{X}_2 = T_P$, $\theta_{i1} \triangleq \theta_{i2}$
- For other choices, e.g. $\dot{X}_1 = H_S$, $\dot{X}_2 =$WindSpeed, $\theta_{i1} \neq \theta_{i2}$ in general
- We will assume $\theta_{i1} = \theta_{i2} = \theta_i$

Objective:

- Objective: model the joint distribution of extremes of $\dot{X}_1$ and $\dot{X}_2$ as a function of θ

(Drop subscripts wherever possible for convenience)

Inference: outline

- Follows conditional extremes (Heffernan and Tawn 2004)
- Model $\dot{X}_1$ and $\dot{X}_2$ marginally as a function of θ:
  - Quantile regression (QR) below threshold
  - Generalised Pareto (GP) above threshold
- Transform to standard Gumbel variates X_1 and X_2
- Model X_2 given large values of X_1 using non-stationary extension of conditional extremes model (incorporating θ)
- Simulate for long return periods:
  - Generate samples of joint extremes on Gumbel scale
  - Transform to original scale
- Simulate structure variables $f(\dot{X}_1, \dot{X}_2)$ as needed

Non-stationary conditional extremes

On Gumbel scale, by analogy with Heffernan and Tawn [2004] we propose the following conditional extremes model:

$$(X_k|X_j = x_j, \theta) = \alpha_\theta x_j + x_j^{\beta_\theta}(\mu_\theta + \sigma_\theta Z) \text{ for } x_j > \phi^G_{j\tau'}(\theta)$$

where:

- $\phi^G_{j\tau'}(\theta)$ is a high directional quantile of X_j on Gumbel scale, 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
- Z will be assumed to be approximately Normally distributed for the purposes of parameter estimation

α_θ, β_θ, μ_θ and σ_θ are functions of direction with B-spline parameterisations

Jon's example

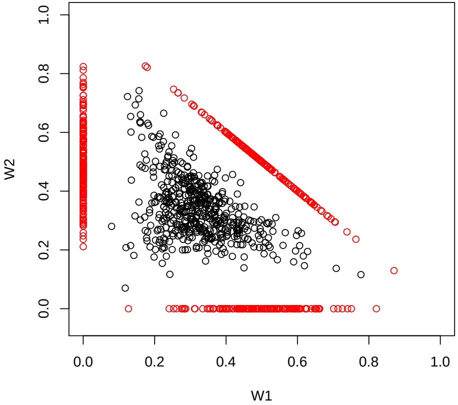


Figure: Physics / covariates to identify dependence structure?

Met-ocean analogy

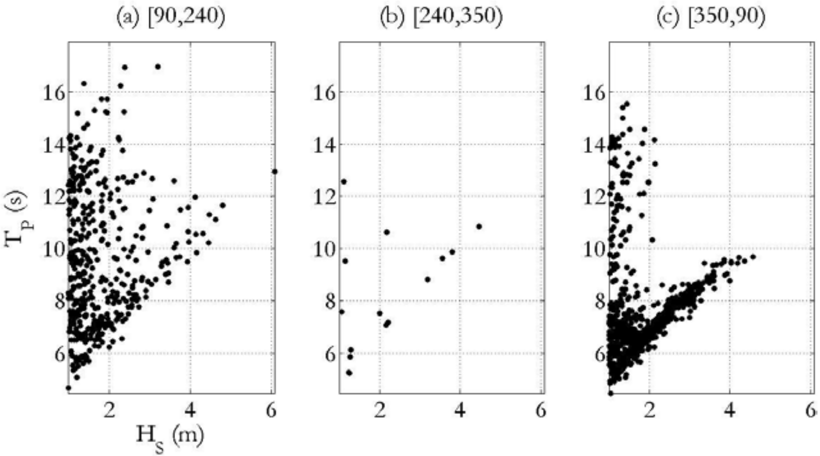


Figure: **Wind-sea** and (multiple) **swell** phenomena for offshore Brazil. T_P is spectral peak period, $1/T_P$ is that frequency at which most energy in propagated by the ocean surface

Simulation study

- Bivariate distribution with Normal dependence transformed marginally to standard Gumbel:

$$(X_1(\theta), X_2(\theta)) \;=\; -\log(-\log(\Phi_{\Sigma(\theta)}(X_{1N}, X_{2N})))$$

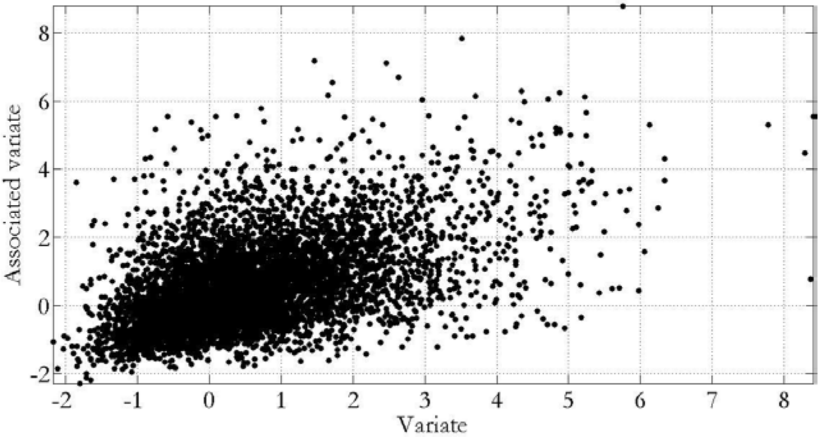
- For large x:

$$\begin{aligned}(X_2(\theta)|X_1(\theta) = x) &\approx \rho^2(\theta)x + x^{1/2}W(\theta)\\ W(\theta) &\sim \text{Normal}\end{aligned}$$

- 6 directional intervals: $\rho^2 = 0.6, 0.9, 0.5, 0.1, 0.7, 0.3$
- Sample size 1000×6

- Marginal forms known, estimate conditional model only
- Parameter estimates: $\alpha = \rho^2$ and $\beta = 1/2$.

Study 1: sample



Study 1: partitioned

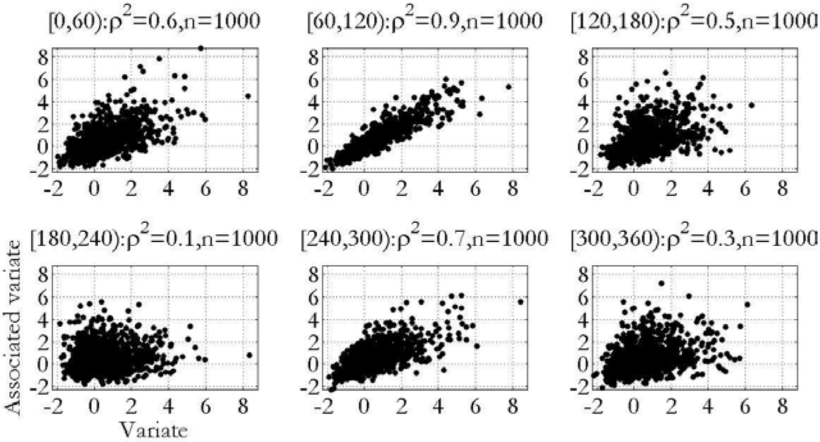


Figure: Simulation Case 1. Scatter plot per covariate interval. Values for intervals for covariate θ, parameter ρ^2 and sample size n are shown in each pane

Study 1: parameter estimates

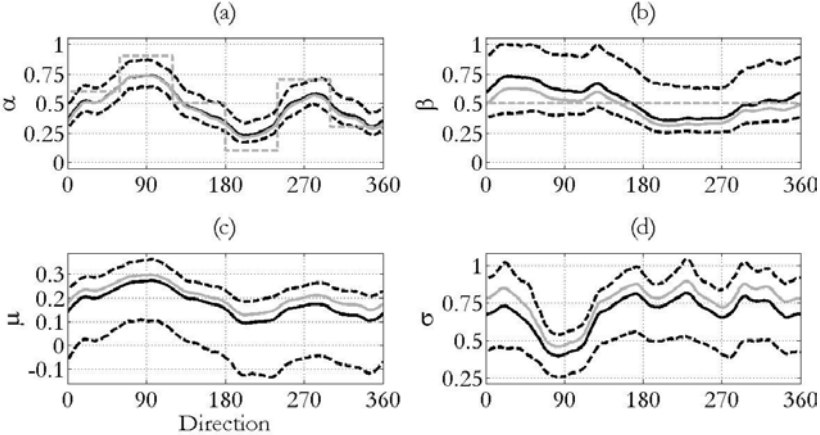


Figure: Simulation Case 1. Sample, bootstrap and true conditional extremes parameters with covariate. Sample estimate are given in solid grey. Median bootstrap estimates are given in solid black, with 95% bootstrap uncertainty bands in dashed black. True values of α and β in dashed grey

Study 1: return values

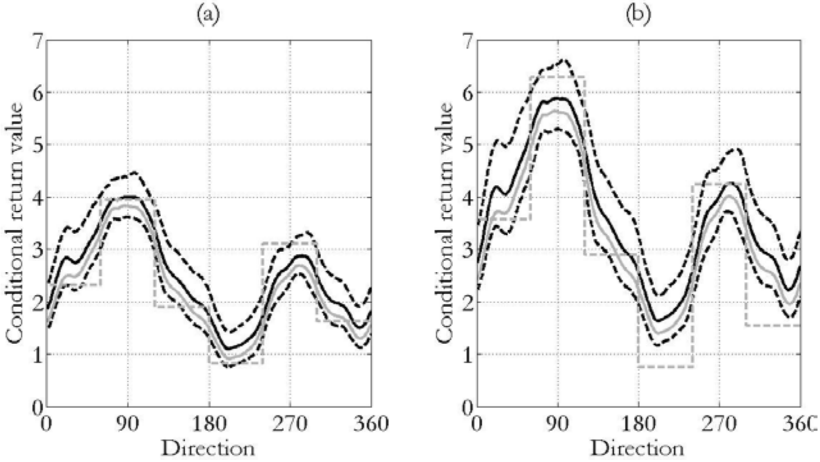


Figure: Simulation Case 1. Conditional return values of the associated variate $\dot{X}_2$, corresponding to values of the conditioning variate $\dot{X}_1$ with non-exceedance probabilities (for period of sample) of (a) 0.99 and (b) 0.999. Bootstrap median (solid) and 95% uncertainty band (dashed) in black. Estimate using actual sample in solid grey. True values in dashed grey

North Sea **directional** parameter estimates

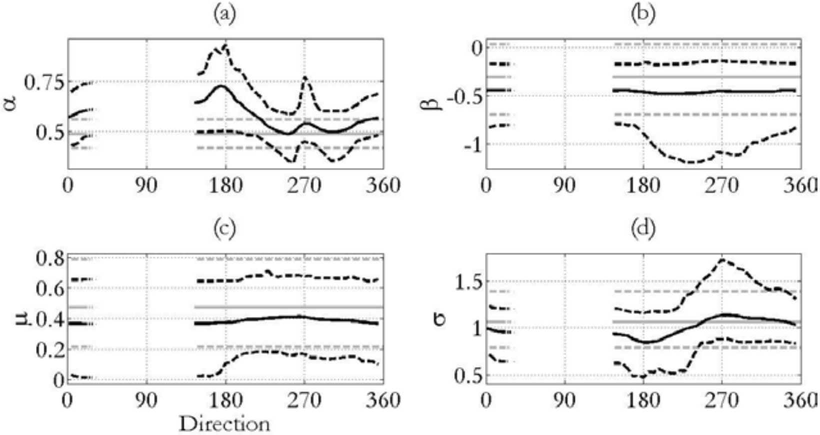


Figure: Northern North Sea. Non-stationary estimates for parameters α, β, μ and σ and their uncertainties (in black) as functions of covariate θ in terms of bootstrap median (solid) and 95% bootstrap uncertainty bands (dashed). Corresponding stationary estimates in grey

North Sea **directional** return values (closed-form)

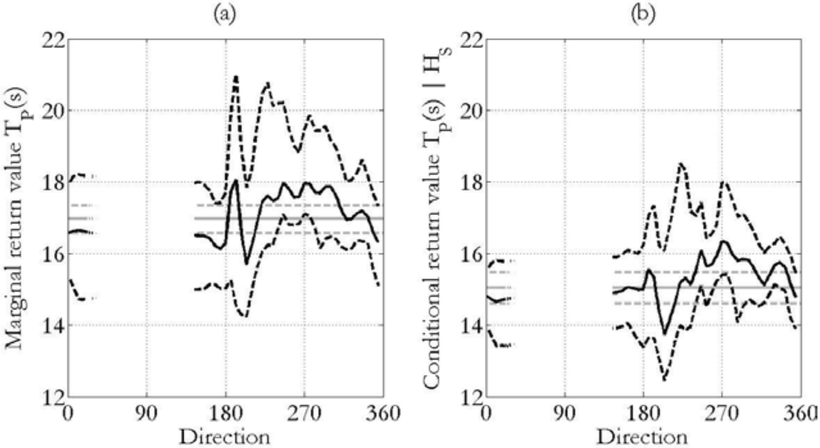


Figure: Northern North Sea. Estimates for (a) marginal return values of associated storm peak period with non-exceedance probabilities (for period of sample) of 0.999, and (b) conditional return value of associated storm peak period given a value of storm peak significant wave height with non-exceedance probabilities (for period of sample) of 0.999. Estimates as functions of covariate θ (black) in terms of median bootstrap value (solid) and 95% bootstrap uncertainty band (dashed). Corresponding estimates assuming no directional dependence in grey

North Sea marginal return values for T_P (simulation)

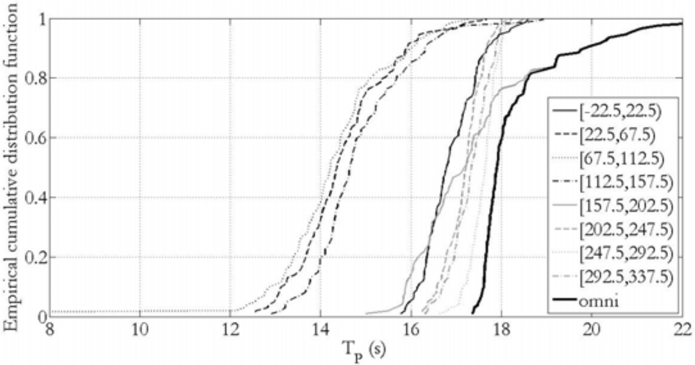


Figure: Omni-directional and sector **marginal** distributions of 100-year T_P^{sp}

North Sea **conditional** return values (simulation)

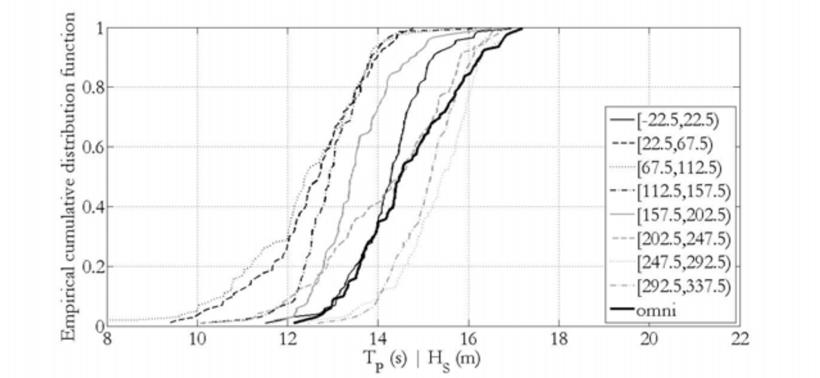


Figure: Omni-directional and sector **conditional** distributions of storm peak period, T_P^{sp} **given** 100-year H_S^{sp} using extension of model of Heffernan & Tawn incorporating non-stationarity

Non-stationary extremes: current developments

- Marginal models:
  - Other covariate representations
  - Extension to higher-dimensional covariates
- Computational efficiency:
  - More **sparse** and **slick** matrix manipulations, optimisation
  - **Parallel** implementation
- **Bayesian** formulation
- **Spatial** model:
  - Composite likelihood: model componentwise maxima
  - Non-stationary dependence
  - Censored likelihood: block maxima $\rightarrow$ threshold exceedances
  - Hybrid model: mix AD and **AI**?
- Non-stationary **conditional** extremes:
  - Multidimensional covariates
  - Multivariate response
- Incorporation within **structural design framework**

References

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.

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

G. Z. Forristall. Wave crest distributions: Observations and second-order theory. *Journal of Physical Oceanography*, 30:1931–1943, 2000.

R. I. Harris. Extreme value analysis of epoch maxima-convergence, and choice of asymptote. *Journal of Wind Engineering and Industrial Aerodynamics*, 92:897–918, 2004.

J. E. Heffernan and J. A. Tawn. A conditional approach for multivariate extreme values. *J. R. Statist. Soc. B*, 66: 497–546, 2004.

P. Jonathan, D. Randell, Y. Wu, and K. Ewans. Return level estimation from non-stationary spatial data exhibiting multidimensional covariate effects. *(Accepted by Ocean Engineering July 2014, draft at www.lancs.ac.uk/∼jonathan)*, 2014.

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

M. Prevosto, H. E. Krogstad, and A. Robin. Probability distributions for maximum wave and crest heights. *Coastal Engineering*, 40:329–360, 2000.

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.

Thank-you / Diolch yn fawr!