Distributions of return values for ocean wave characteristics using directional-seasonal extreme value analysis

Philip Jonathan, David Randell, Kevin Ewans, Graham Feld
Shell Stats & Chemometrics

Acknowledgement

- Shell colleagues (statistics, metocean)
- Colleagues in academia, especially at Lancaster University
- Shell interns and summer students

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

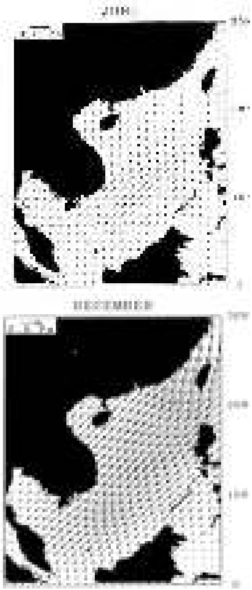
- **Componentwise maxima**:
  - $\Leftrightarrow$ max-stability $\Leftrightarrow$ multivariate regular variation
  - Assumes all components extreme
  - $\Rightarrow$ Perfect independence or asymptotic dependence **only**
  - Composite likelihood for spatial extremes (Davison et al. 2012)
  - Point process / 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

Marginal directional-seasonal extremes

googlemaps

Marginal directional-seasonal extremes

- Marginal model: **single** location
- Response: **storm peak significant wave height,** H_S^{sp}
- Wave climate: **monsoonal**
- Southwest monsoon ($\sim$ August, to northwest for us)
- Northeast monsoon ($\sim$ January, to east for us)
- Long **fetches** to Makassar Strait, Java Sea
- Land shadows of Borneo (northwest), Sulawesi (northeast), Java (south)



www.westernpacificweather.com

Marginal directional-seasonal extremes

- **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 1956 – 2012
- Variables: H_S, direction (from, clockwise from north), season and wave period information

- South China Sea platform (Storm Jangmi, $H_S = 3.6m$, $H \approx 6m$): Link

Directional and seasonal variability

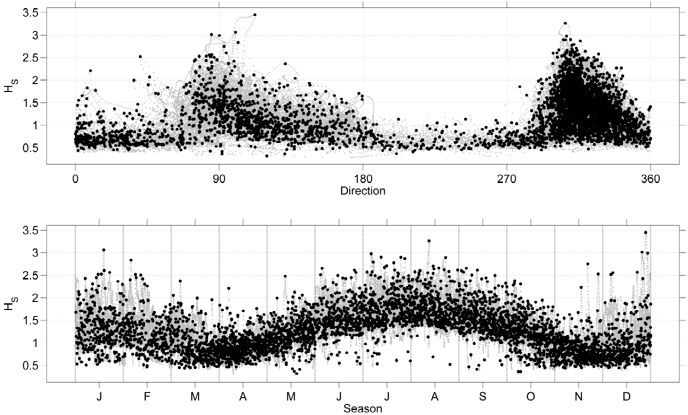


Figure: Storm peak significant wave height H_S^{sp} (black) on direction θ (upper panel) and season ϕ (lower panel). Also shown is sea-state significant wave height H_S (grey) on direction θ (upper panel) and season ϕ (lower panel). Southwest monsoon: August from northwest (315). Northeast monsoon: January from east (90).

Storm model

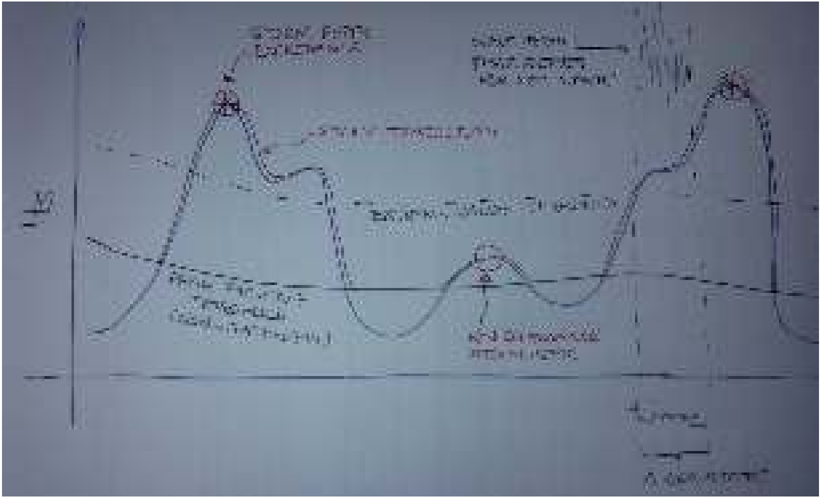


Figure: $H_S \approx 4\times$ standard deviation of ocean surface profile at a location corresponding to a specified period (typically three hours)

Quantiles of H_S^{sp}

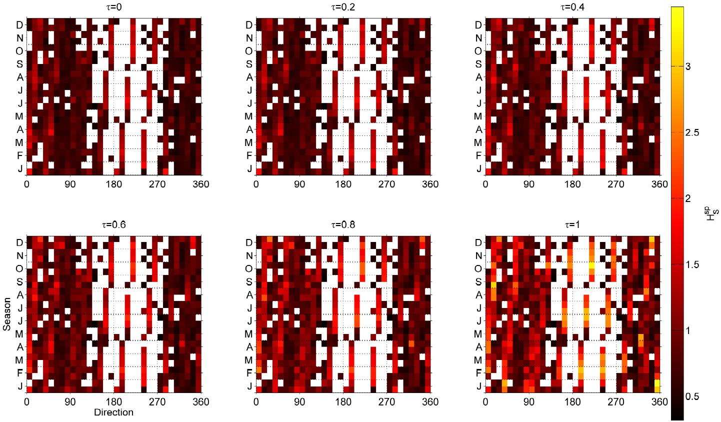


Figure: Empirical quantiles of storm peak significant wave height H_S^{sp} by direction θ and season ϕ, for threshold non-exceedance probabilities τ as listed. Empty bins are coloured white.

Storm **trajectories** of significant wave height, H_S

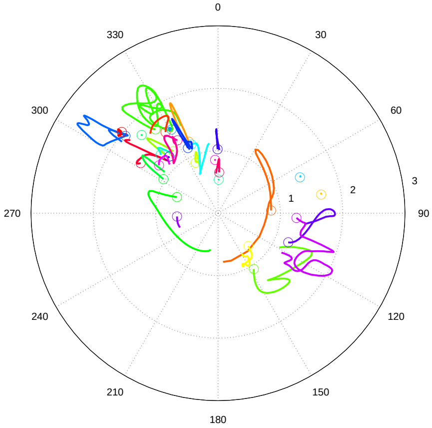


Figure: Storm trajectories of significant wave height H_S on wave direction θ for 30 randomly-chosen storm events (in different colours). A circle marks the start of each intra-storm trajectory.

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_u 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_u
  3. **Size of occurrence** of threshold exceedance using generalised Pareto (GP) model with shape and scale parameters ξ_u and σ_u

- Model estimated for **multiple** thresholds with non-exceedance probabilities τ_u, $u = 1, 2, 3, ...$

(Drop *sp* superscripts and *u* subscripts 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:

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

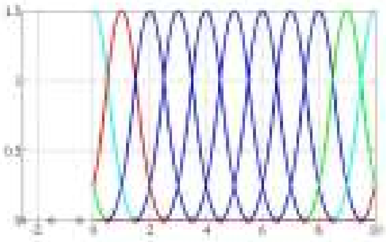
  (**exact** operations calculated without explicit evaluation)
- Roughness R_η defined as:

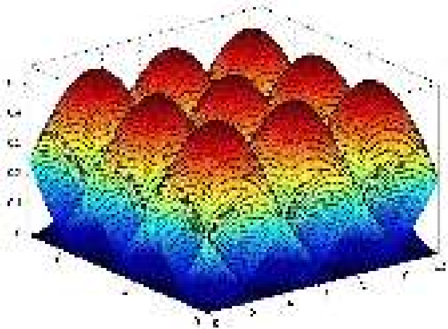
  $$R_\eta = \beta_\eta' P \beta_\eta$$

  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 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 extreme value **threshold**, ψ

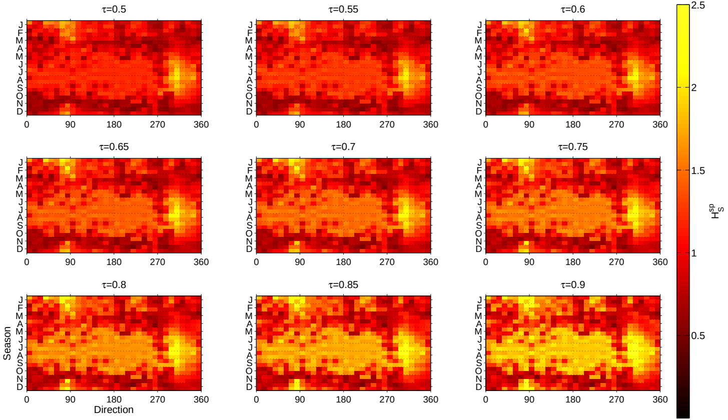


Figure: Directional-seasonal plots for extreme value thresholds, ψ, corresponding to equally spaced non-exceedance probabilities of H_S^{sp} on [0.5, 0.9] (left to right, then top to bottom).

Accommodating **multiple** thresholds

- Median **threshold ensemble** estimates

$$
\begin{aligned}
\tilde{\rho} &= \operatorname*{med}_{u} \left\{ \rho_u \frac{\tau_{\tilde{u}}}{1-\tau_u} \right\} \\
\tilde{\sigma} &= \operatorname*{med}_{u} \{ \sigma_u + \xi_u(\psi_{\tilde{u}} - \psi_u) \} \\
\tilde{\xi} &= \operatorname*{med}_{u} \{ \xi_u \}
\end{aligned}
$$

- Parameter estimates can be fairly compared
- $\tau_{\tilde{u}}$ set to 0.5

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):

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

- $\{c_j\}_{j=1}^m$ 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**, $\tilde{\rho}$

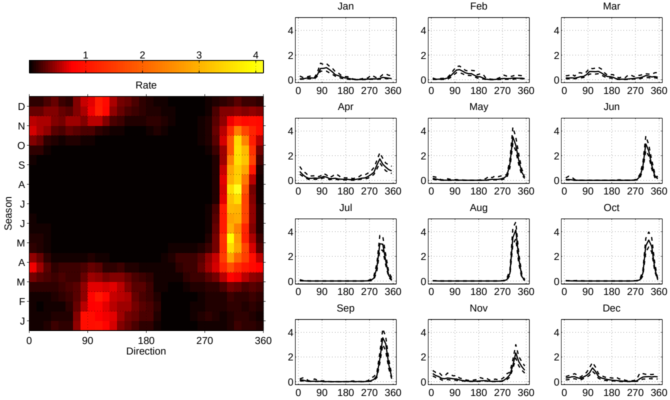


Figure: Directional-seasonal plot for median threshold ensemble rate $\tilde{\rho}(\times 1000)$ of threshold exceedance of H_S^{sp}. The left-hand panel shows $\tilde{\rho}(\times 1000)$ on θ^{sp} and ϕ^{sp}. The right hand panel shows 12 monthly directional estimates with 95% BCa bootstrap confidence intervals (dashed).

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 **scale**, $\tilde{\sigma}$

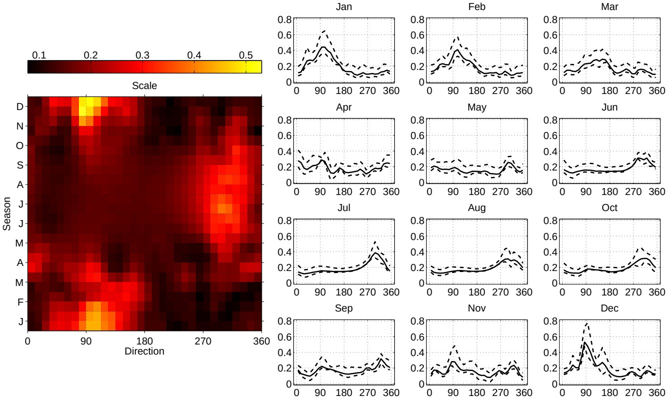


Figure: Directional-seasonal plot for median threshold ensemble generalised Pareto scale, $\tilde{\sigma}$. The left-hand panel shows $\tilde{\sigma}$ on θ and ϕ. The right hand panel shows 12 monthly directional estimates with 95% BCa bootstrap confidence intervals (dashed).

Directional-seasonal parameter plot for GP **shape**, $\tilde{\xi}$

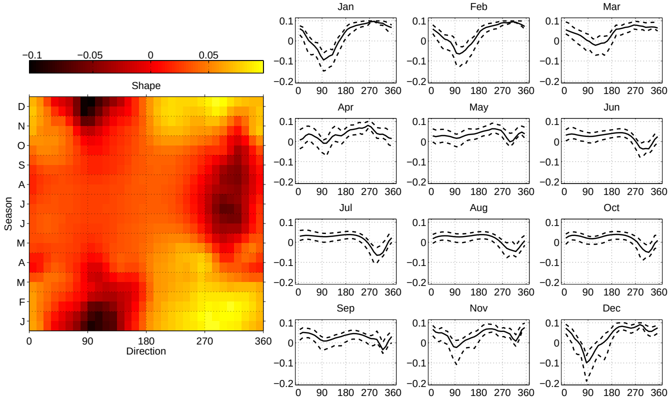


Figure: Directional-seasonal plot for median over threshold generalised Pareto shape, $\tilde{\xi}$. The left-hand panel shows $\tilde{\xi}$ on θ and ϕ. The right hand panel shows 12 monthly directional estimates with 95% BCa bootstrap confidence intervals (dashed).

Return values

- Estimation of return values by simulation under the model
  - Threshold level selected at random
  - 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**

Accommodating **multiple** thresholds

- **Threshold ensemble** estimates of return value distributions

$$\begin{aligned}\Pr(\tilde{Q} \leq x) &= \int_{\tau \in J_\tau} \Pr(Q \leq x|\tau) dF(\tau) \\ &\approx \frac{1}{n_\tau} \sum_{u=1}^{n_\tau} \Pr(Q \leq x|\tau_u)\end{aligned}$$

- The quantiles $\tilde{q}(p)$ are solutions to $\Pr(\tilde{Q} \leq x) = p$.
- Incorporates threshold variability in return value estimate

Return value plot for H_{S100}, $\tilde{q}(0.5)$

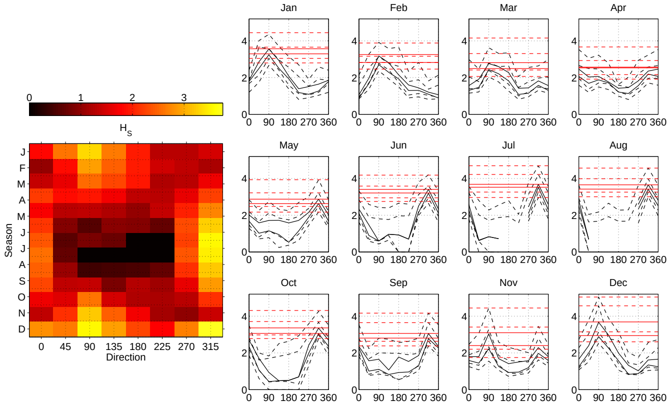


Figure: Directional-seasonal return value plot for 100-year significant wave height (in metres). The left-hand panel shows directional and seasonal variability of the median threshold ensemble estimate $\tilde{q}(0.5)$ for H_S. The right hand panel shows 12 monthly directional octant return values (in black) in terms of BCa 95% confidence limits for $\tilde{q}(0.5)$ (solid), $\tilde{q}(0.025)$ (dashed) and $\tilde{q}(0.975)$ (dashed). Also shown are the corresponding omni-directional estimates (in red).

Within-storm variability

Figure: Wave has removed boat landing gear

Within-storm variability

Critical environmental variables:

- Storm peak significant wave height:
  - (Sea state) significant wave height
  - Maximum wave height
  - **Maximum crest elevation**, C
  - Peak total water level ($\approx$ crest + 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 $C|H_S(t)$

The cumulative distribution function for the maximum crest elevation C in a sea-state parameterised by $\mathcal{S}$ of $n_\mathcal{S}$ waves with significant wave height $H_S = h_S$ is taken (see, e.g. Forristall 1978, 2000) to be given by:

$$\Pr(C \leq \eta|\mathcal{S}) = (1 - \exp(-\frac{\eta}{\alpha_\mathcal{S} h_S})^{\beta_\mathcal{S}})^{n_\mathcal{S}}$$

where all of $\alpha_\mathcal{S}$, $\beta_\mathcal{S}$ and $n_\mathcal{S}$ are functions of the sea-state parameters $\mathcal{S}$ estimated from observation.

Directional-seasonal return value plot for C_{100}

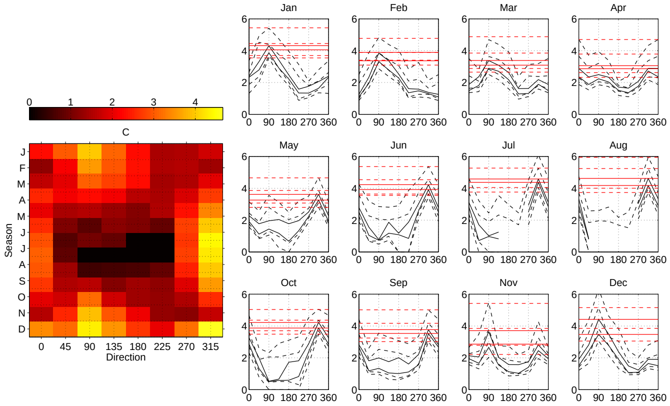


Figure: Directional-seasonal return value plot for 100-year crest elevation (in metres). The left-hand panel shows directional and seasonal variability of the median quantile over threshold $\tilde{q}(0.5)$ for C. The right hand panel shows 12 monthly directional octant return values (in black) in terms of BCa 95% confidence limits for $\tilde{q}(0.5)$ (solid), $\tilde{q}(0.025)$ (dashed) and $\tilde{q}(0.975)$ (dashed). Also shown are the corresponding omni-directional estimates (in red).

Validation of model for sea-state H_S

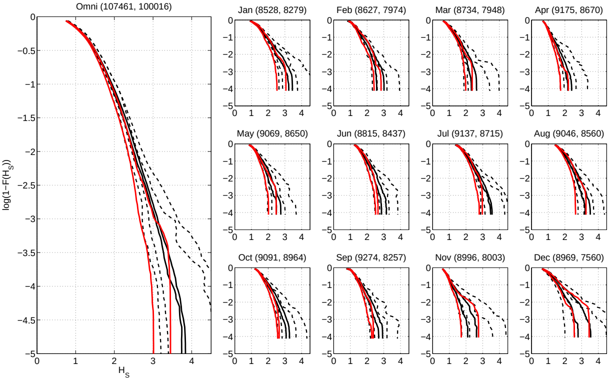


Figure: Illustration of validation of return value estimation for significant wave height by comparison of cumulative distribution functions (cdfs) for 1000 bootstrap resamples of the original sample with those from 1000 sample realisations under the model (incorporating intra-storm evolution of H_S) corresponding to the same time period as the original sample. The 12 right hand panels show empirical 95% bootstrap uncertainty bands for monthly omni-directional cdfs for the original sample (red), and BCa 95% confidence intervals for the 2.5%ile and 97.5%ile median over threshold estimates $\tilde{q}(0.025)$ and $\tilde{q}(0.975)$ (both dashed). Titles for plots, in brackets following the month name, are the numbers of actual and simulated events in each month. The left hand panel makes the equivalent omni-directional, omni-seasonal comparison.

Interval of threshold non-exceedance probability

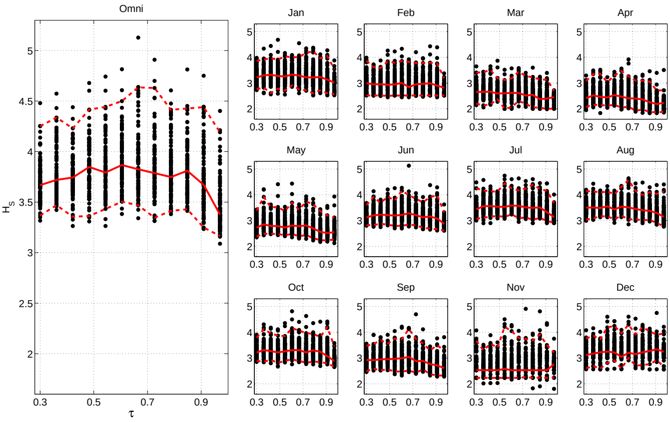


Figure: Estimates for 100-year maximum for H_S^{sp} from simulation under models corresponding to 100 bootstrap resamples for each of 15 choices of threshold non-exceedance probability, τ. Median estimates are connected by a solid red line. 2.5% and 97.5% iles are connected by dashed red lines. The left hand panel shows the omni-directional, omni-seasonal estimate. The right hand panels show 12 monthly omni-directional estimates.

Bootstrap threshold ensemble, $\breve{Q}$

- *Bootstrap threshold ensemble* return value $\breve{Q}$
- Estimate return value distribution $Q|B,\tau$ by simulation for threshold non-exceedance probability τ and bootstrap resample $B \in \mathcal{B}$

$$\begin{aligned} \Pr(\breve{Q} \le x) &= \int_{\tau \in J_\tau} \int_{B \in \mathcal{B}} \Pr(Q \le x|B,\tau) dF(B) dF(\tau) \\ &\approx \frac{1}{n_\tau}\frac{1}{n_B}\sum_{u=1}^{n_\tau}\sum_{b=1}^{n_B} \Pr(Q \le x|B_b,\tau_u) \end{aligned}$$

Bootstrap threshold ensemble $\breve{Q}$ and median threshold ensemble $\tilde{Q}$

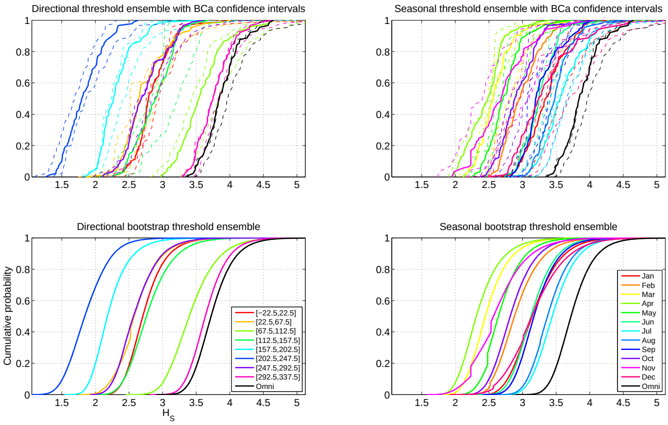


Figure: Empirical cumulative distribution functions (cdfs) for 100-year significant wave height from simulation under the directional-seasonal model. Left hand and right hand panels show directional and seasonal cdfs respectively. Upper panels shows median threshold ensemble estimates $\tilde{Q}$ with 95% BCa confidence intervals, and lower panels bootstrap threshold ensemble estimates $\breve{Q}$.

Non-stationary extremes: **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*, 54:207–222, 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. On the statistical distribution of wave heights in a storm. *J. Geophysical Research*, 83:2353–2358, 1978.

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. *Ocean Eng.*, 88:520–532, 2014.

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

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.

Marginal **spatio-directional**

Figure: Hurricane Katrina

Marginal 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
  - (Spatial) block bootstrap

GoM spatio-directional H_S^{sp}

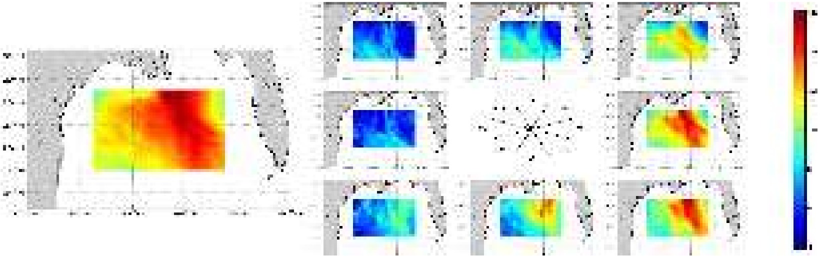


Figure: $\approx$ 17000 locations $\times$ 32 directional bins for Gulf of Mexico. Plot for quantile (withheld) of 100-year maximum storm peak significant wave height, H_S^{sp}

NWS spatio-directional H_S^{sp}

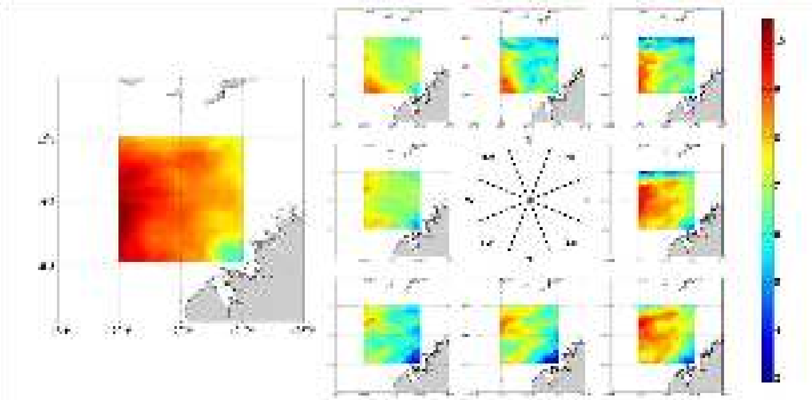


Figure: North West Shelf of Australia. See Jonathan et al. [2014]

Non-stationary **conditional** extremes

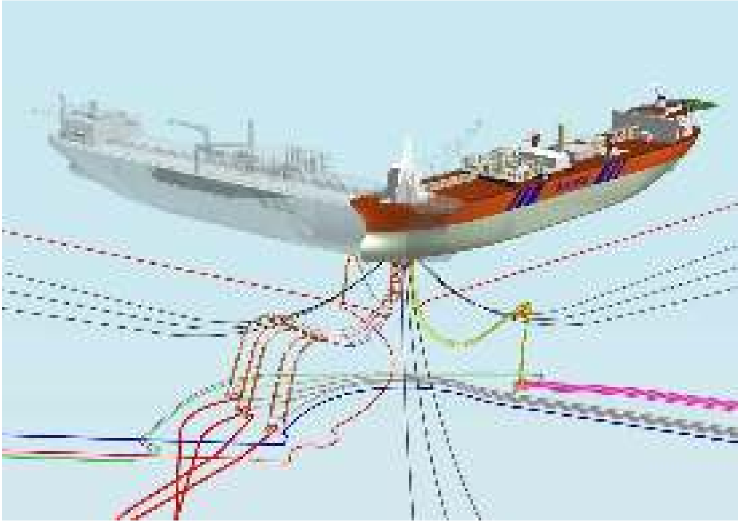


Figure: Floating LNG tanker (500m long!)

Non-stationary conditional extremes

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)

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

North Sea **marginal** return values for T_P (simulation)

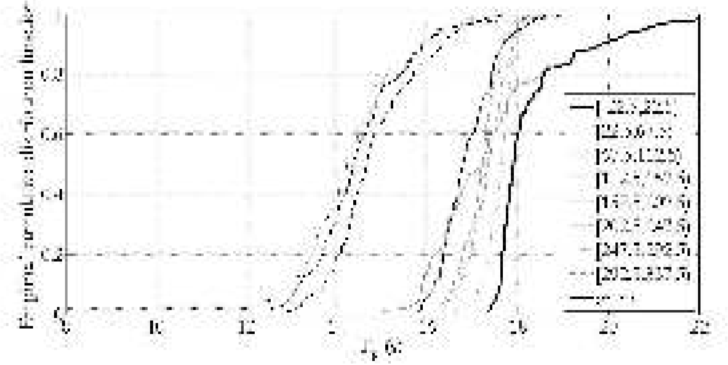


Figure: Omni-directional and sector **marginal** distributions of 100-year T_P^{sp}

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

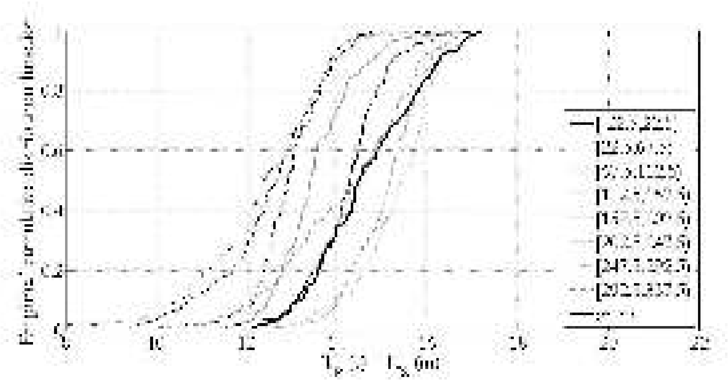


Figure: Omni-directional and sector **conditional** distributions of storm peak period, T_P^{sp} **given** 100-year H_S^{sp} using extension of model of Heffernan & Tawn incorporating non-stationarity