Extremes for metocean: Fundamentals

Philip Jonathan

Lancaster University

IOGP, London, September 2026
(slides at *www.lancs.ac.uk/∼jonathan)*

Motivation for masterclass

Problem set-up

- We think of the ocean environment as a "data generating process"
- Sometimes that process generates extreme values of particular concern
- The only way we can understand the extremes is by direct observation
- We fit an empirical tail model to data we collect
- We use the model to predict how often and how big the extremes will be

What are the issues?

- Which tail model should we fit? And why?
- How do we fit the tail model?
- How do we know we've achieved a good fit? How useful is the model?
- What if the process is non-stationary (ie there are covariates at play)?
- What if we want to fit a multivariate tail model?
- How do we tie this all together for engineering design?

Overview of masterclass

- A: Fundamentals
- B: Covariates
- C: Multivariate
- D: covXtreme
- E: Recent developments & take-aways

Overview of A: Fundamentals

- Motivating example
- Tail models commonplace in engineering
- Statistically-motivated tail models
- Fitting a tail model
- Key issues with GPD fitting of peaks-over-threshold
- Uncertainty quantification
- Basic statistical model fitting tools

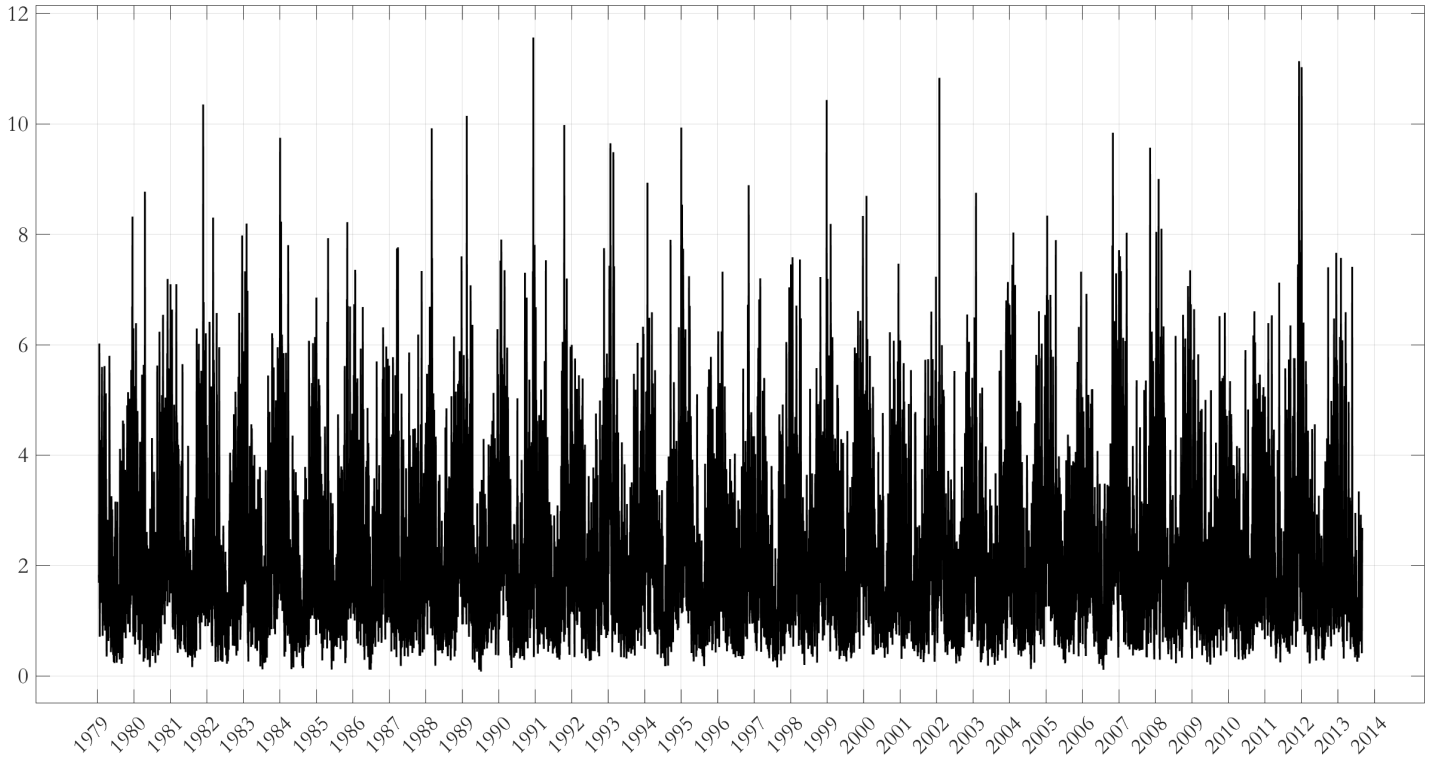
Sanity check

- All models are wrong, some models are useful
- George Box, `http://en.wikipedia.org/wiki/GeorgeE.P.Box`

- Consistent balance of engineering and statistical insights

Motivating application

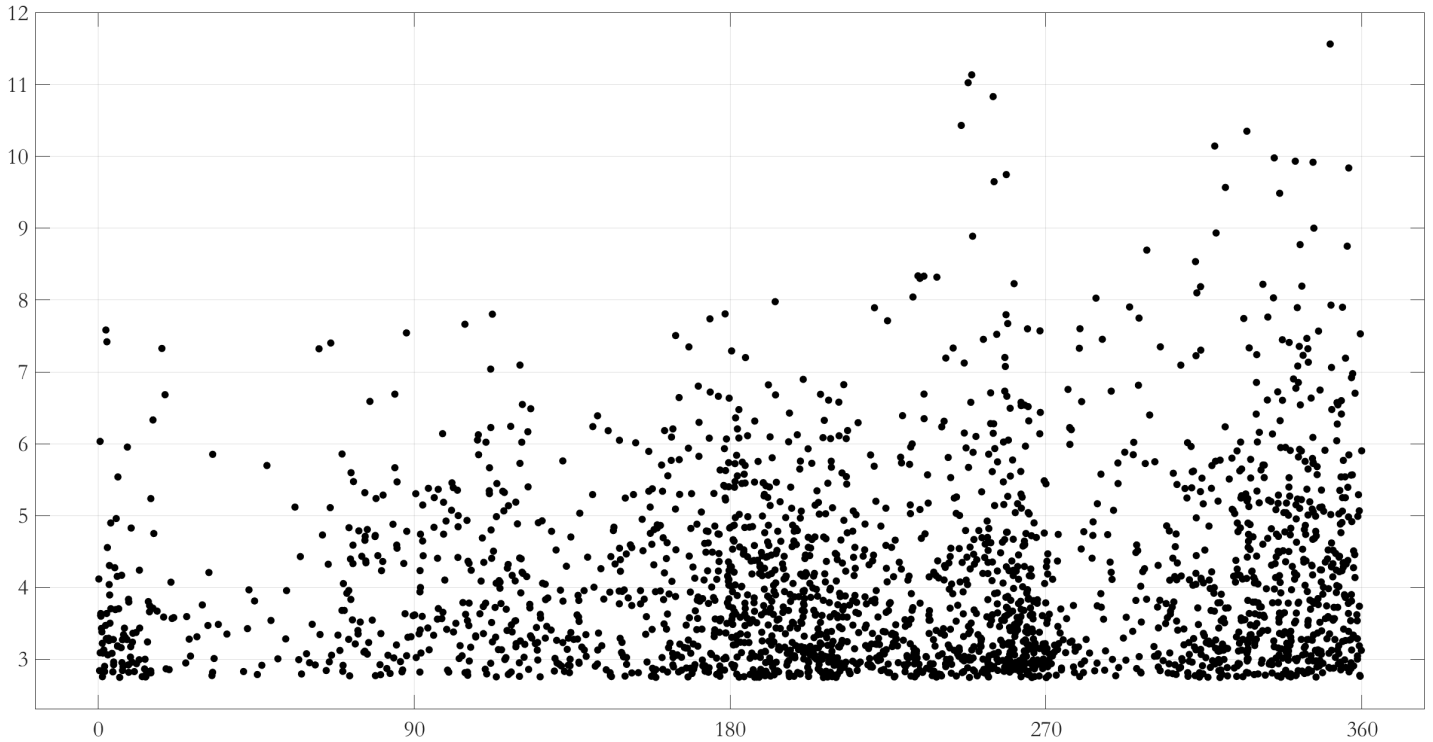
- 3-hour sea state H_S from a central North Sea location
- Data available from covXtreme repository



Time-series of H_S

Motivating application

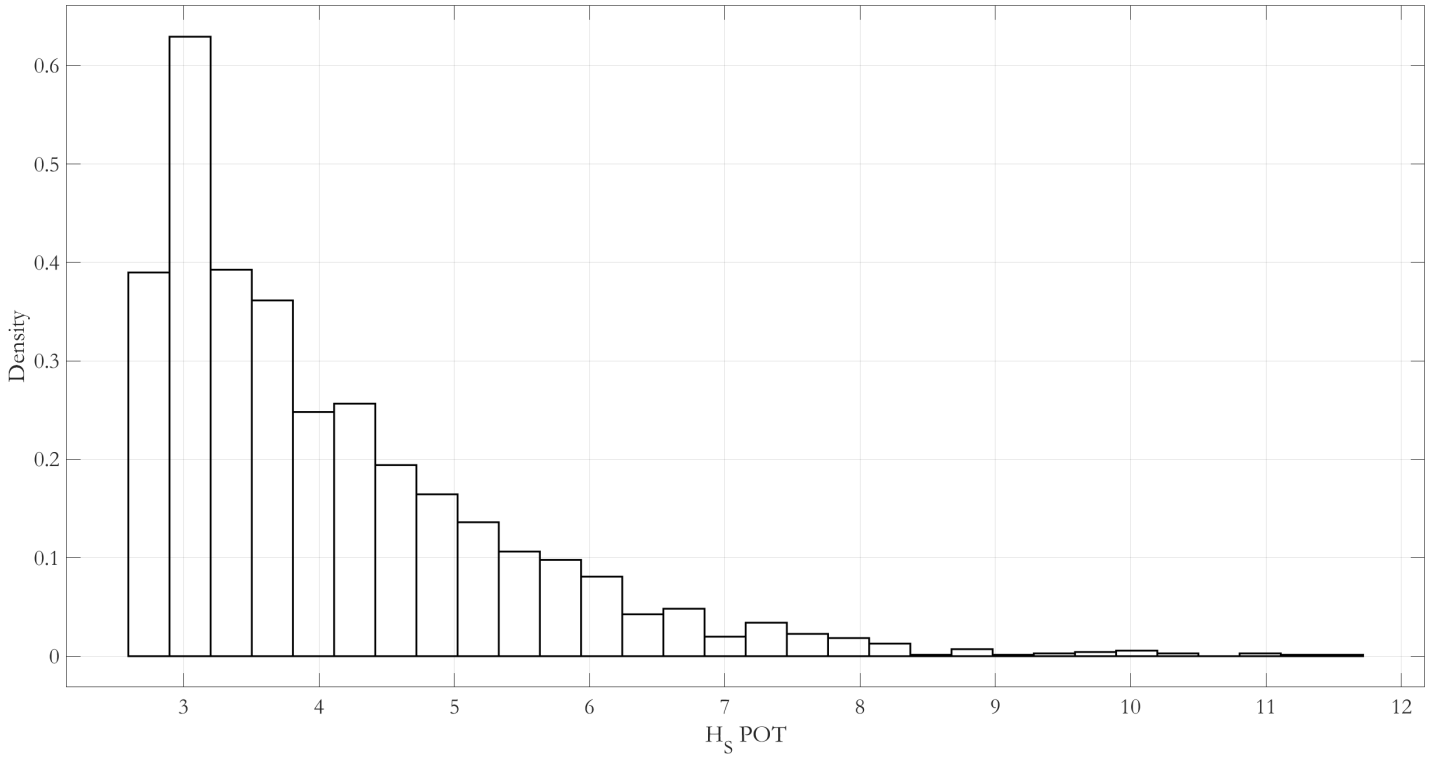
- Peak-picking algorithm used to isolate storm peaks H_S^{sp} over threshold



H_S POT on direction

Motivating application

- Density of H_S^{sp}



Density of H_S POT

Typical engineering approach

- Some variant of "Weibull" distribution
- For random variable X, e.g. H_S^{sp}

$$\mathbb{P}(X \leq x; \lambda, k) = F(x; \lambda, k) = \begin{cases} 1 - e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Shape $k > 0$, scale $\lambda > 0$
- There is no limit to how big X can get

- Widely used across engineering disciplines
  - Changing k (and λ) gives flexibility
  - Tail of density decreases for large x
- But there is no "fundamental" reason for using this distribution

Understanding the statistical approach

Degenerate distribution of block maxima

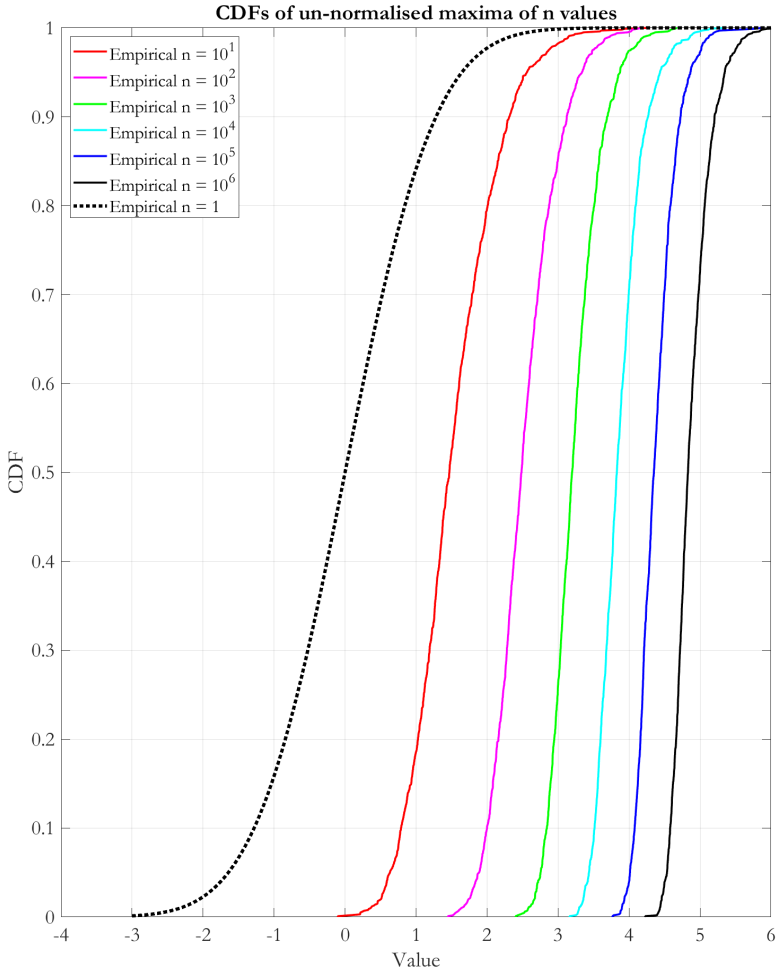
- $F(X) = \mathbb{P}(X \leq x)$, cumulative distribution function (cdf)
- $M_n = \max_i\{X_i\}$, block maximum
- $\mathbb{P}(M_n \leq x) = [\mathbb{P}(X \leq x)]^n$, cdf of maximum
- As $n \uparrow \infty$, $\mathbb{P}(M_n \leq x)$ becomes degenerate ($= 0$ everywhere except at the maximum value of X,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

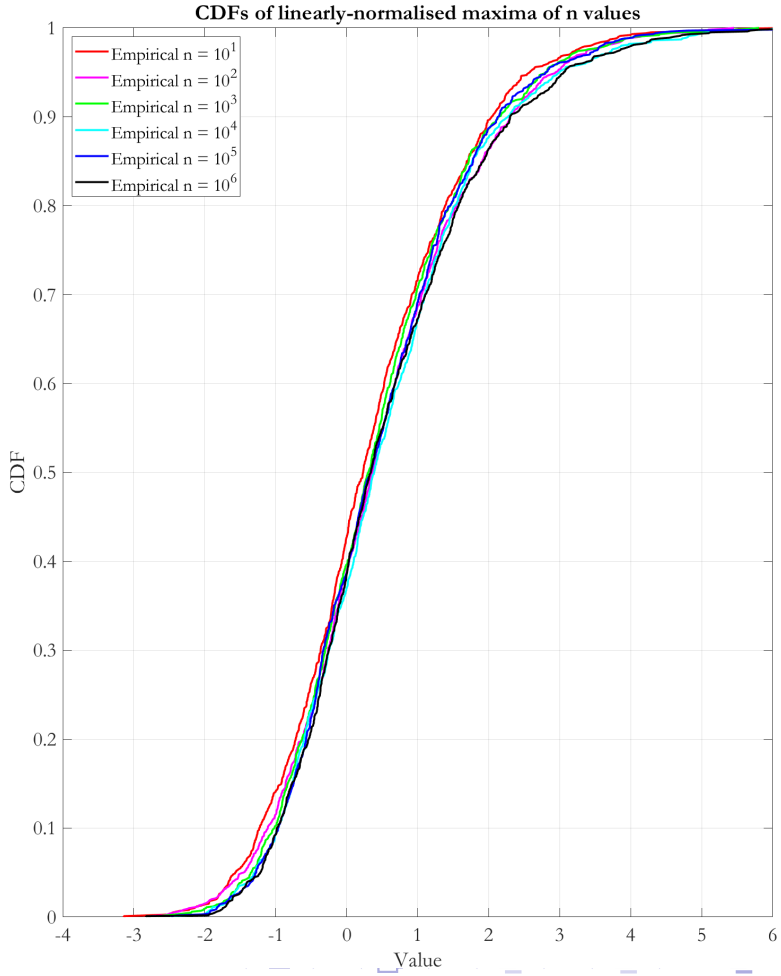
Non-degenerate distribution of shifted and scaled block maxima

- Try shifting and scaling the random variable to make its tail more stable (this is like the central limit theorem)
- $Y_n = a_n^{-1}(max_i\{X_i\} - b_n)$
- $\mathbb{P}(Y_n \leq y) = [\mathbb{P}(X \leq b_n + a_n y)]^n$
- As $n \uparrow \infty$, $\mathbb{P}(Y_n \leq y)$ is almost always well behaved once we find good a_n and b_n
- The tail of the distribution of Y_n is max-stable!

Understanding the statistical approach

- Left: distribution $[\mathbb{P}(X \leq x)]^n$ of un-normalised Gaussian maxima becomes degenerate
- Right: distribution $[\mathbb{P}(X \leq a_n y + b_n)]^n$ of linearly-normalised maxima converge to standard Gumbel





Understanding the statistical approach

The non-degenerate limiting distribution is unique

- $Y_n = a_n^{-1}(max_i\{X_i\} - b_n)$
- $\mathbb{P}(Y_n \leq y) = [\mathbb{P}(X \leq b_n + a_n y)]^n$

Generalised extreme value distribution (GEV)

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

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

- Location $\mu \in \mathbb{R}$, scale $\sigma > 0$, shape $\xi \in \mathbb{R}$

Maximum domain of attraction (MDA)

- A distribution whose (linearly normalised) maxima converges, belongs to the MDA of a max-stable distribution (for some ξ)
- We call these distributions "nice"; most continuous distributions are nice
- Once we find ξ for one choice of a_n and b_n, it must be the same for all choices of a_n and b_n

- The Weibull distribution converges to GEV with:
  - $\xi = 0$
  - $a_n = \frac{\lambda}{k}(\log n)^{1/k} - 1$
  - $b_n = \lambda(\log n)^{1/k}$ ($= F^{-1}(1 - 1/n)$ in general)

- Note: this theory is analogous to central limit theorem. There is nothing mysterious here!

Unified von Mises condition

- For any nice distribution, the ultimate value of GEV ξ is given by

$$\lim_{x \to x^*} \frac{d}{dx}\left(\frac{1}{g(x)}\right) = \xi, \quad g(x) = \frac{f(x)}{1 - F(x)} = \text{hazard}$$

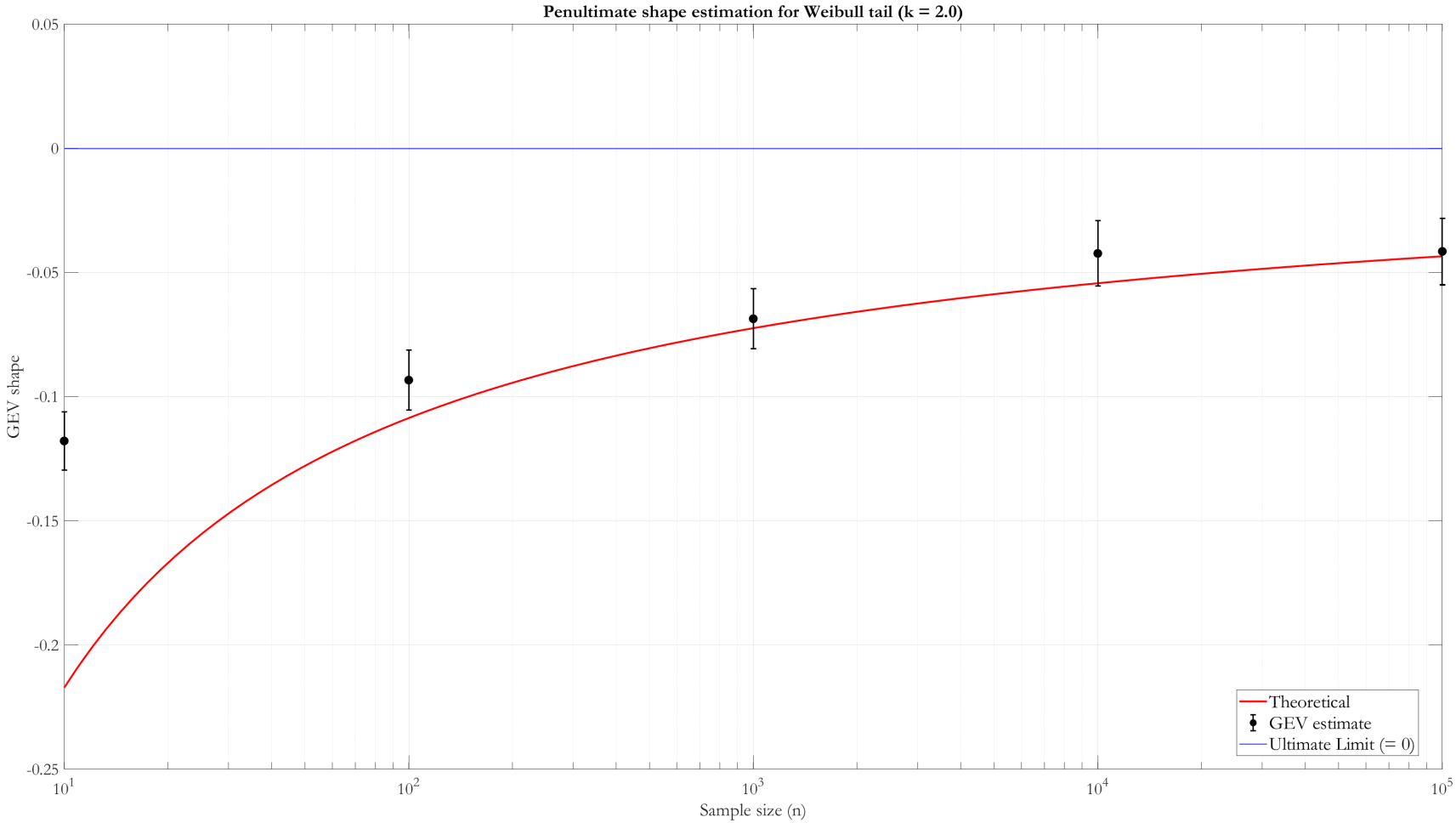
- The "penultimate approximation" ξ_n of ξ, for sample size n can be estimated using

$$\xi_n = \frac{d}{dx}\left(\frac{1}{g(x)}\right)_{x=b_n} \approx \frac{1-k}{k \ln n} \text{ for Weibull}$$

- Penultimate approximate gives better fit for samples from distributions which converge slowly to ultimate tail
- For finite Weibull sample ($n < \infty$) with $k > 1$ (e.g. $k \approx 2$, Rayleigh), we have $\xi_n < 0$, and we expect to fit a shorter tail with $x^* < \infty$
  - We should fit a GEV, but expecting $\hat{\xi} < 0$
  - We need to adjust our estimate of ξ_n for larger return value calculations!

Penultimate approximation for Weibull tail

- Weibull sample size n for $k = 2$ (Rayleigh)
  - Penultimate $\xi_n < 0$: we expect to fit a shorter tail expecting $\hat{\xi} < 0$
  - Hence expect finite upper end point $x^* < \infty$
  - Adjust estimate $\xi_{n'}$ to use in return value calculations when $n' > n$



What about POT?

- X has distribution function F in MDA of GEV(ξ)
- If n is large enough, $F^n(a_n x + b_n) \approx \exp(-\left(1 + \xi \frac{x-\mu'}{\sigma'}\right)^{-\frac{1}{\xi}})$ (GEV tail)
- So $F^n(x) \approx \exp(-\left(1 + \xi \frac{x-(a_n\mu'+b_n)}{\sigma' a_n}\right)^{-\frac{1}{\xi}}) = \exp(-\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}})$ say
- Then $n \log F(x) \approx -\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}$ (log both sides)
- $\mathbb{P}(X > x) = 1 - F(x) \approx \frac{1}{n}\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}$ ($\log y \approx -(1-y)$) for $y \approx 0$)
- $\mathbb{P}(X > x | X > u) = \frac{1-F(x)}{1-F(u)} \approx \left(1 + \xi \frac{x-u}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}$ (re-arrangement)
- where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ is the adjusted scale; common ξ for GEV and GP!

Generalised Pareto distribution (GP or GPD)

$$\mathbb{P}(X > x | X > u) = \left(1 + \xi \frac{x-u}{\sigma}\right)^{-\frac{1}{\xi}} \qquad \xi \in \mathbb{R}, \sigma > 0, x > u, x \leq x^*$$

- Block maxima from nice distributions $\to$ GEV as $n \to \infty$
- Threshold exceedences from nice distributions $\to$ GP as $u \to x^*$

Recovering GEV from GP

- X has a nice distribution function F, conditional tail $F_u(x) = \frac{F(x)}{1-F(u)}$
- Assume occurrence rate of exceedences over threshold u is Poisson-distributed

$$\begin{aligned}
\mathbb{P}(\text{maximum in period } \leq x | x > u) &= \sum_{k=0}^{\infty} (k \text{ storms in period}) F_u^k(x) \\
&= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \exp(-\lambda) F_u^k(x) \\
&= \exp\{-\lambda(1 - F_u(x))\}
\end{aligned}$$

- But threshold exceedences are GP, so $1 - F_u(x) = \left(1 + \xi \frac{x-u}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}$, so

$$\begin{aligned}
\mathbb{P}(\text{maximum in period } \leq x | x > u) &= \exp\{-\lambda \left(1 + \xi \frac{x-u}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}\} \\
&= \exp\{-(1 + \xi \frac{x-\mu}{\sigma})_+^{-1/\xi}\} \quad \text{(GEV)}
\end{aligned}$$

- when λ set to $\left(1 + \xi \frac{u-\mu}{\sigma}\right)^{-\frac{1}{\xi}}$ for $\tilde{\sigma} = \sigma + \xi(u - \mu)$ wlog; ξ unchanged

Take-aways

- Statistics suggests we should fit tails of distributions with GEV and GP
  - Motivation for GEV and GP is asymptotic theory
  - Can only justify fitting GEV and GP when we are in the tail
  - Threshold exceedences are GP if u large enough
  - Block maxima are GEV if n large enough

- "Weibull" is a restricted choice of distribution for modelling corresponding to $\xi = 0$ (and slow convergence)
  - Physics (e.g. constraints like Miche for $H_{\max}$ and water depth for H_S) tells us that Weibull cannot always be correct
  - Weibull might be easier to fit to data (since it is more restricted), but this doesn't necessarily make it better

- So how do we proceed?
  - Simulation studies for known data generating models
  - Understand characteristics of models under misspecification
  - Consistent balance of engineering and statistical insights

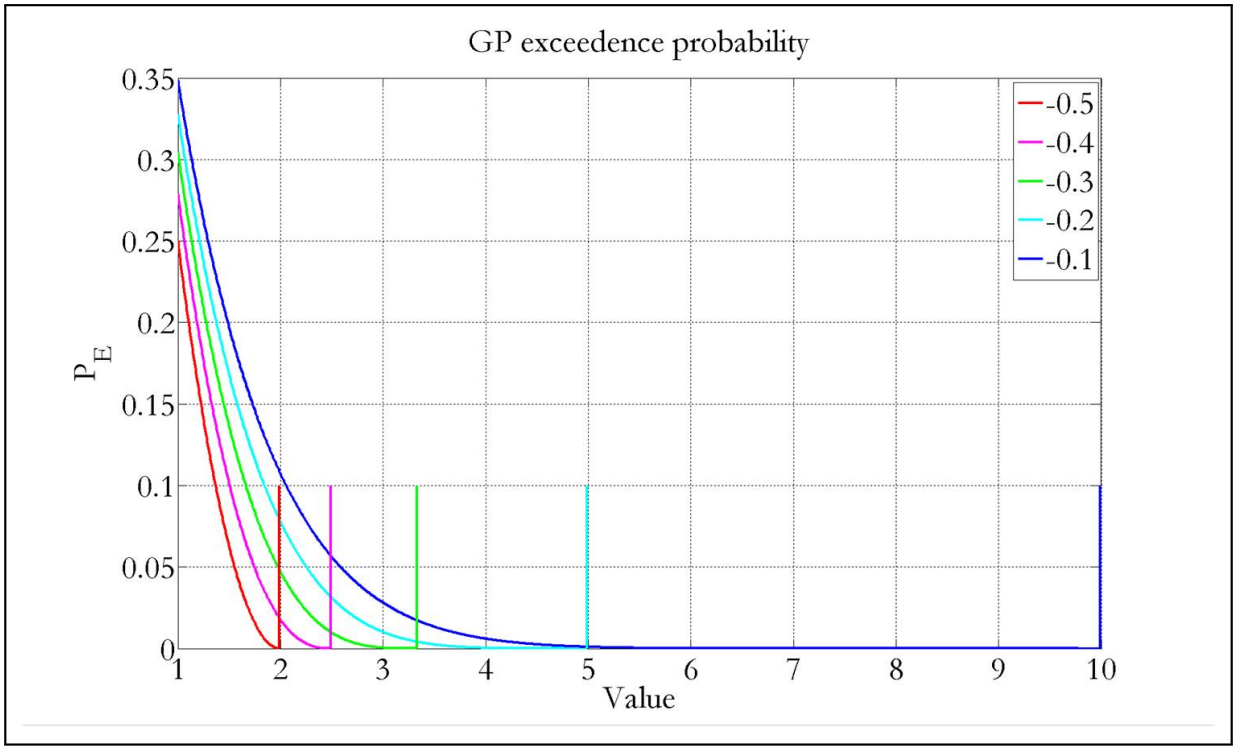
GP: effect of ξ

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

- For ocean waves, observation and physics suggests that x^* 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 $x^* = \infty$

GP: effect of $\xi < 0$



- $x^* = u - \sigma/\xi$ for $\xi < 0$
- Assuming $u = 0, \sigma = 1$ in the figure, and $\mathbb{P}_E = 1 - F_u$
- As $\xi \uparrow 0$, then $x^* \uparrow \infty$

GP: threshold stability of ξ

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

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

- Conditional tail is still GP
- ξ is unchanged, σ varies linearly with gradient ξ aafo threshold
- This is a useful fit diagnostic!

GP: return values

Distribution of annual maximum, A

- $X|(X > u) \sim F_u(x) = \text{GP}(\xi, \sigma, u) = 1 - (1 + \xi \frac{x-u}{\sigma})^{-\frac{1}{\xi}}$
- $F_{A;u}$ is the (conditional) distribution of the annual maximum event A

$$F_{A;u}(x) = \exp\{-\lambda(1 - F_u(x)\}$$

P-year return value, x_P

$$1 - \frac{1}{P} = \exp\{-\lambda(1 - F_u(x_P))\}$$

- λ is the expected number of exceedences $X > u$ per annum.

Distribution of P-year maximum, A_P

$$F_{A_P;u}(x) = \exp\{-P\lambda(1 - F_u(x)\}$$

Equivalent expressions for GEV

In practice: threshold choice

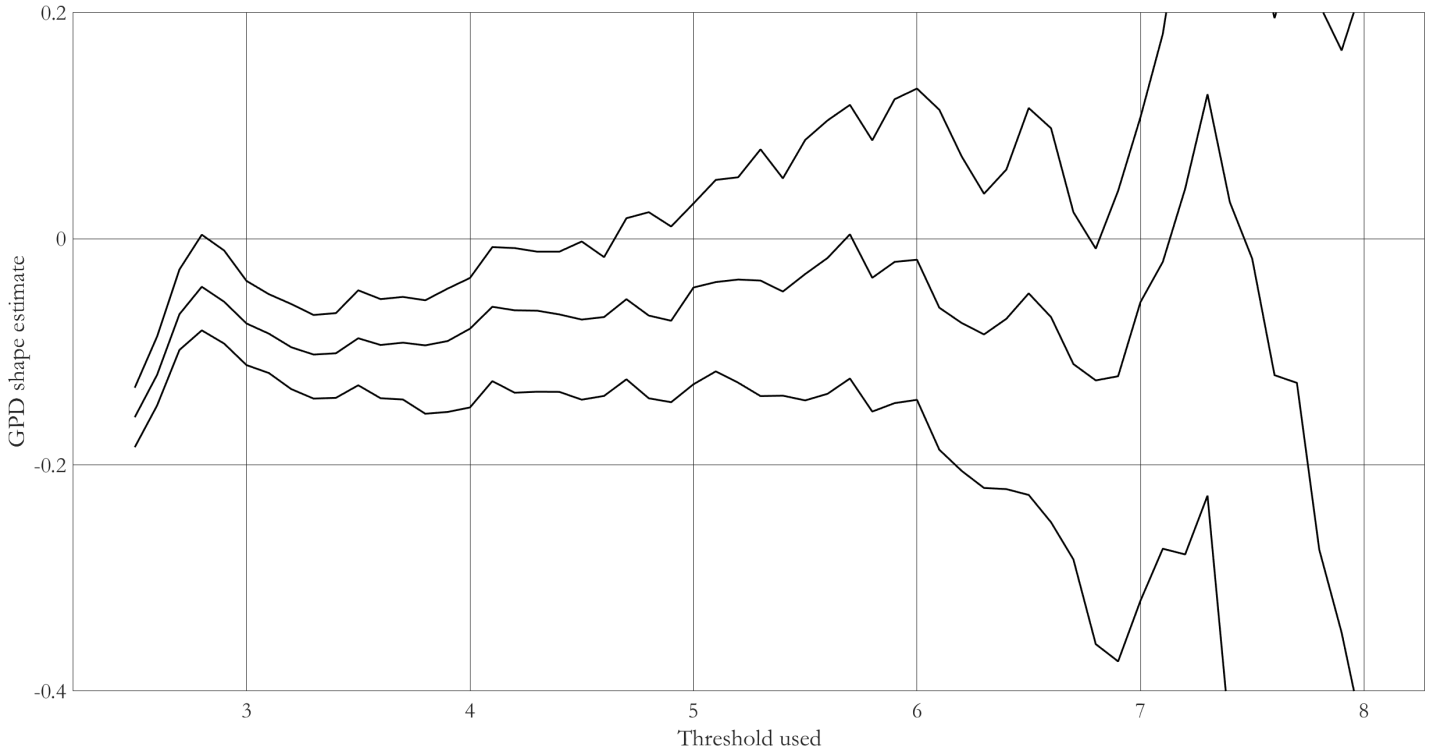
- We're fitting threshold exceedences using GP tail. What do we need to consider?
- Choice of threshold, u
  - If GP fits well, estimated ξ is constant as we vary u
  - Also check if estimated return value x_p is constant with u
  - Do we need to choose just one threshold? What about model averaging?
- PP and QQ plots (more later)
- Uncertainty in estimates
  - As sample size $\downarrow$, uncertainty in estimated parameters and $x_p \uparrow$
  - Often good to look at uncertainty of estimated x_p
  - Use bootstrap resampling

In practice: bootstrapping

- A data-driven way to estimate uncertainty, making minimal assumptions
- The procedure is:
  - Draw a new sample the same size as the original by resampling from the original with replacement
  - Re-do the analysis with the new sample
  - Repeat for a large number (≈ 250) of bootstrap resamples to estimate the uncertainty of the original estimates
- Standard practice in the statistics community

In practice: stability of shape estimate, $\hat{\xi}$

- North Sea data
- $\hat{\xi}$ should be stable with u if GP appropriate
- Plot shows $\hat{\xi}$ for actual sample (middle line)
- Also shows 95% bootstrap uncertainty band from resampling
- $u \in [3, 6]$ yields reasonably stable model



Stability of $\hat{\xi}$

In practice: cross-validation

- A data-driven way of assessing predictive performance
- Hence: a way of selecting the optimal value for a hyper-parameter
  - e.g. GP threshold u
- Or: a way of deciding which of competing models is best
  - e.g. GP or Weibull
- The procedure is:
  - Partition the sample into k groups
  - Estimate model using $k - 1$ of the groups, keeping the "withheld" group as a prediction (or "test") set
  - Assess predictive performance on the prediction set
  - Repeat until all k groups have been used exactly once for prediction
  - Estimate overall "k-fold" predictive performance for full sample
  - If necessary, choose the hyper-parameter or model that maximises predictive performance
- Standard practice in machine learning and computational statistics

In practice: model fitting

Maximum likelihood estimation

- $X|X > u \sim F_u(\xi, \sigma) = GP(\xi, \sigma)$, with density $f(x; \xi, \sigma)$
- $f(x; \xi, \sigma) = d(F_u(x))/dx$
- Observation x_i drawn from $GP(\xi, \sigma)$ has "likelihood" $f(x_i; \xi, \sigma)$
- Likelihood of whole sample $(x_1, x_2, x_3, ..., x_n)$ is $\mathcal{L}(\xi, \sigma)$

$$\mathcal{L}(\xi, \sigma) = \prod_{i=1}^{n} f(x; \xi, \sigma)$$

- Choose estimates of ξ and σ which maximise sample likelihood
- Usually have to estimate ξ and σ numerically

- Great coupled with cross-validation for hyper-parameter optimisation
- Equivalent to least squares for Gaussian samples
- Alternatives: method of moments for small samples

In practice: hyper-parameter and model selection

- We have a fiddle factor θ that we can't estimate easily. Why?
  - θ could be the GP threshold u
  - θ could switch between a GP tail ($\theta = 0$ say) and Weibull ($\theta = 1$ say)
- Use maximum likelihood estimation and cross-validation together

- Partition sample S into k groups $S_1, S_2, ..., S_k$
- Evaluate optimal predictive likelihood for hyper-parameter choice θ

$$\mathcal{L}^{\text{Opt}}_{\text{Pred}}(S;\theta) = \max_{\xi,\sigma} \prod_{j=1}^{k} \mathcal{L}_{\text{Pred}}(S_j;\xi,\sigma,S_{-j},\theta)$$

- $\mathcal{L}_{\text{Pred}}(S_j;\xi,\sigma,S_{-j},\theta)$ is the predictive likelihood of group S_j for a model estimated using S_{-j} (i.e. all data except S_j), parameters ξ and σ, and hyper-parameter θ

- Choose θ maximising $\mathcal{L}^{\text{Opt}}_{\text{Pred}}(S;\theta)$: $\theta^{\text{Opt}} = \arg\max_\theta \mathcal{L}^{\text{Opt}}_{\text{Pred}}(S;\theta)$
- A data-driven way to choose optimal threshold for GP, and even tail model form

In practice: penalised maximum likelihood

- What if parameters ξ and σ are subject to other constraints?
- Suppose we can write the "badness" of constraint violation as $C(\xi, \sigma)$, with $C = 0$ meaning that constraints are not violated
- We can perform maximum likelihood estimation with constraints

$$\begin{aligned} \mathcal{L}^*(\xi, \sigma) &= \mathcal{L}(\xi, \sigma) \exp\{-C(\xi, \sigma)\} \\ \text{or } \log(\mathcal{L}^*(\xi, \sigma)) &= \log(\mathcal{L}(\xi, \sigma)) - C(\xi, \sigma) \end{aligned}$$

- So we just maximise $\mathcal{L}^*$ in place of $\mathcal{L}$ over parameters
- Principled way of imposing constraints on estimates
- An intuitive Bayesian interpretation
- Reverts to maximum likelihood estimation if $C(\xi, \sigma) = 0$
- Key to fitting flexible covariate models (more later)

In practice: model averaging

- Can't decide which threshold to use, or whether we use GP or Weibull?
- Construct an ensemble of models, and use then all in decision making!
- Estimate m models $(\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_m)$
- Then we might estimate m return values $(x_{P;\mathcal{M}_1}, x_{P;\mathcal{M}_2}, ..., x_{P;\mathcal{M}_m})$, and use the full set of return value estimates for decision making, e.g.

$$x_P^* = \frac{1}{m} \sum_{j=1}^{m} x_{P;\mathcal{M}_j}$$

- Or $f^*(z) = \frac{1}{m} \sum_{j=1}^{m} f_{\mathcal{M}_j}(z)$ for "predictive" density of quantity z
- Principled model combination with intuitive Bayesian interpretation
- Weight models according to our beliefs about their relative performance
- Key to incorporating threshold and sampling uncertainty in covariate models (more later)

In practice: Bayesian inference

- Need a systematic way of incorporating prior knowledge?
- Specify a prior density $f(\theta)$ for parameters ($\theta = \xi, \sigma, ...$)
- Combine with sample likelihood $\mathcal{L}(\theta) \triangleq f(\text{Data}|\theta)$
- Estimate posterior density $f(\theta|\text{Data})$ for parameters

$$f(\theta|\text{Data}) = \frac{f(\text{Data}|\theta)f(\theta)}{\int_{\theta'} f(\text{Data}|\theta')f(\theta')d\theta'} \propto f(\text{Data}|\theta)f(\theta) = \text{Likelihood} \times \text{Prior}$$

- Calculate posterior predictive estimates for any quantity of interest
  - Posterior predictive estimate e.g. of return value $q(\theta)$
  - Posterior predictive density e.g. of sea state H_S or individual $H_{\max}$, $f(z|\theta)$

$$\begin{aligned} q_{\text{PP}} &= \int_\theta q(\theta)f(\theta|\text{Data})d\theta \\ f(z|\text{Data}) &= \int_\theta f(z|\theta)f(\theta|\text{Data})d\theta \end{aligned}$$

- Posterior predictive estimates are just "special" model averages
  - Weighted, with weights given by posterior density of parameters

Books worth getting

- Coles [2001]: Best readable introduction to extreme value analysis
- Pawitan [2001]: Readable introduction to applied statistical modelling
- Davison [2003]: Best coverage of applied statistical modelling, with extremes examples

References

S. Coles. *An introduction to statistical modelling of extreme values*. Springer, Berlin, 2001.

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

Y. Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Clarendon Press, Oxford, 2001.