Cross-validatory extreme value threshold selection and uncertainty with application to offshore engineering

Paul Northrop
University College London
p.northrop@ucl.ac.uk

Joint work with Nicolas Attalides and Philip Jonathan

Statistical Science Seminar, University of Exeter
2nd Feburary 2015

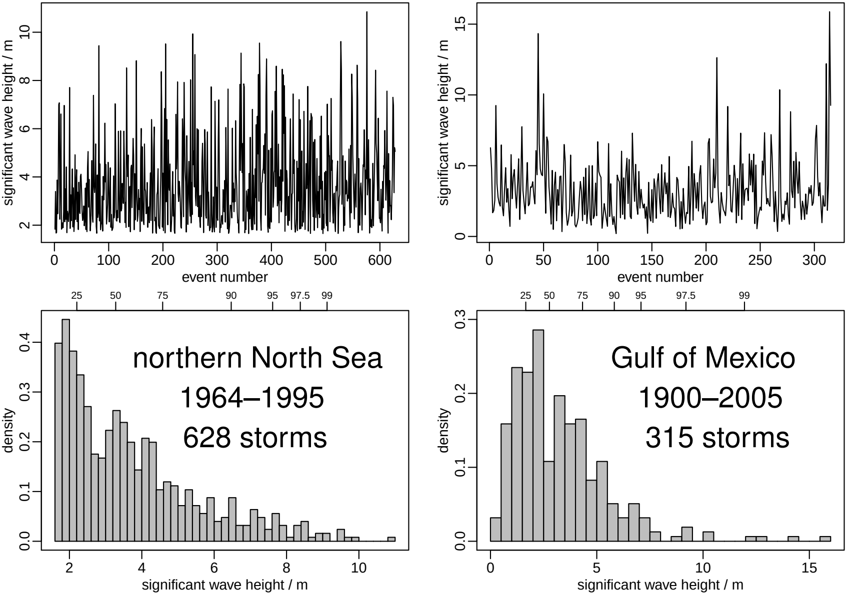
An oil platform

Want to avoid this . . .

Outline

- Significant wave height (H_S) datasets; extrapolation
- Threshold-based extreme value (EV) modelling
- Selection of a single threshold
- Averaging inferences over many thresholds
- Predictive inferences using Bayesian computation: the role of EV priors

Hindcast storm peak sig. wave heights



Extrapolation

A scenario for the Gulf of Mexico (105 years of data):

> *What level of storm peak H_S is exceeded with probability 0.05 in a 21-year period?*

Assuming stationarity and independence between years

- level occurs approx. once every $20 \times 21 = 420$ years
- 105 years: a sample of size 5 of quantity of interest (21-year maxima)
- w. p. $0.95^5 \approx 0.77$ this level is not attained in 105 years

We need to protect against conditions that are (probably) more severe than on record

Threshold-based EV models

…under very idealised assumptions, i.e.

$$X_1, X_2, \ldots, X_n \overset{\text{indep}}{\sim} \text{ with common CDF } H$$

- set a threshold u
- model motivated by considering the possible limiting distributions of (scaled) excesses of u as $u \to \infty$

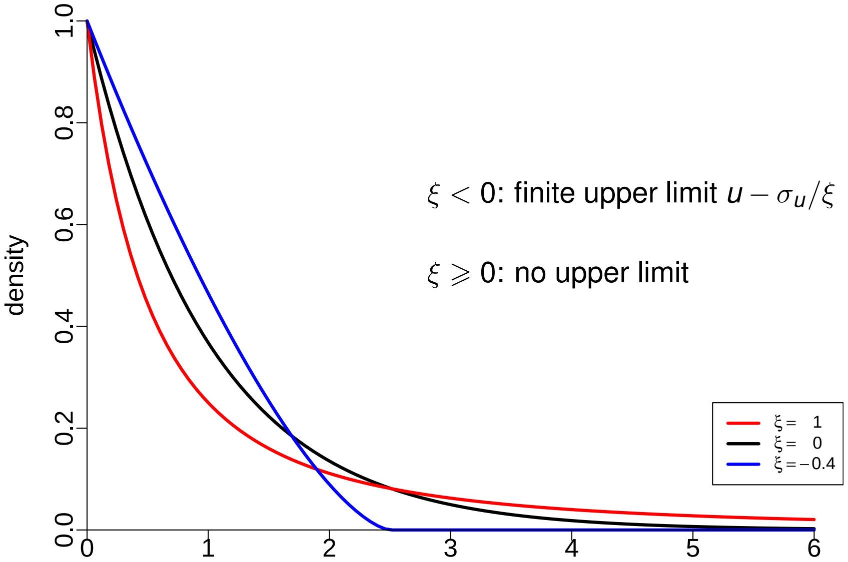
$$(X_i - u) \mid X_i > u \;\dot{\sim}\; GP(\sigma_u, \xi)$$

- Let $p_u = P(X_i > u)$ and N_u be the number of excesses of u

$$N_u \sim \text{binomial}(n, p_u)$$

Need u to be large enough that the bin-GP model(p_u, σ_u, ξ) might be useful

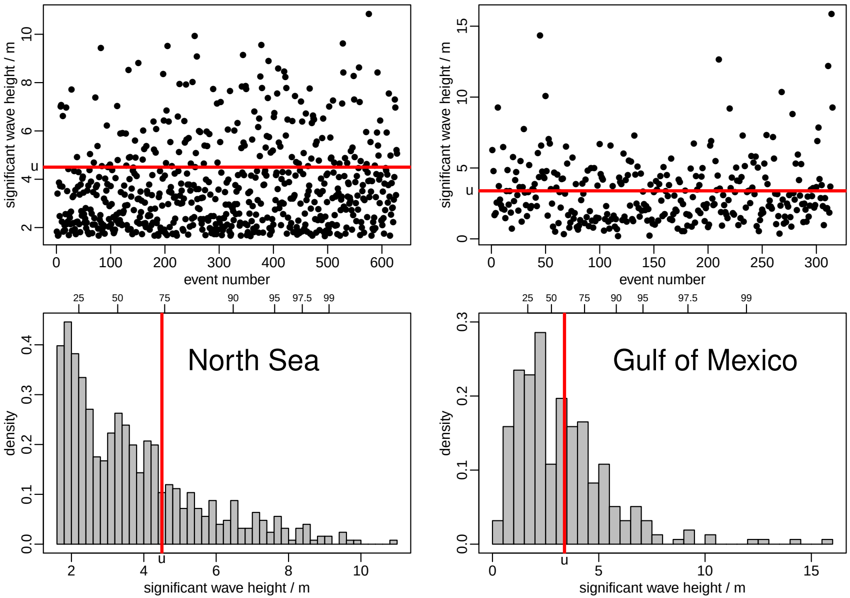
GP(1,ξ) densities



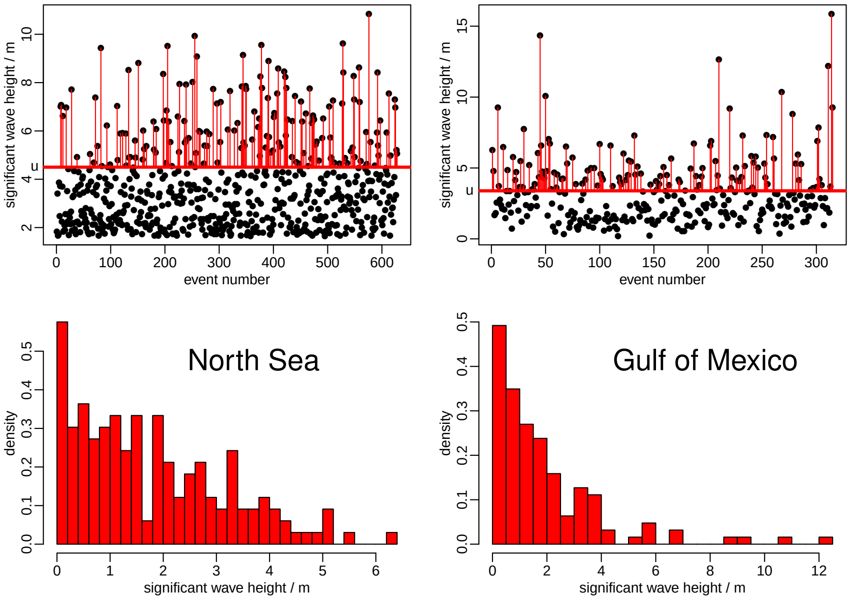
$\xi < 0$: finite upper limit $u - \sigma_u/\xi$

$\xi \geqslant 0$: no upper limit

Hindcast storm peak sig. wave heights



Threshold excesses



Threshold diagnostics

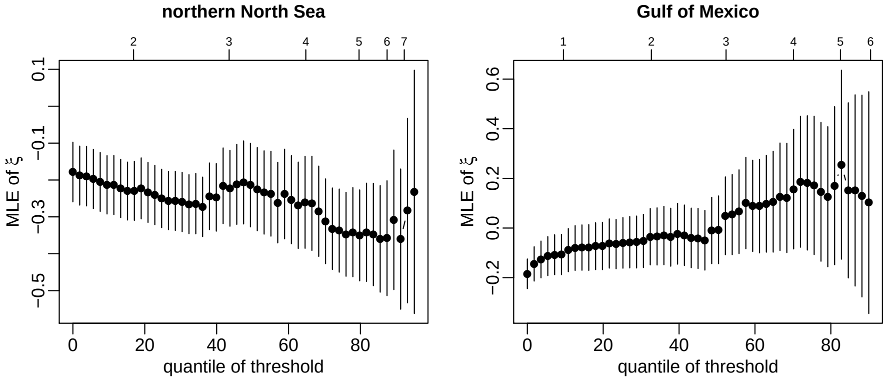
Bias-variance trade-off :

- u too low : GP model inappropriate $\rightarrow$ bias
- u too high : fewer excesses $\rightarrow$ unnecessary imprecision

Review paper: Scarrott & MacDonald (2012)

- Estimates of ξ stable above some level of threshold?
  [Drees et al. (2000), Wadsworth and Tawn (2012), Northrop and Coleman (2014)]
- Goodness-of-fit of GP distribution
  [Davison and Smith (1990), Dupuis (1998)]
- Minimize asymptotic MSE of estimates of ξ or extreme quantiles under assumptions about H
  [Ferreira, et al. (2003), Beirlant (2004)]
- Extend EV model below u and make u a model parameter
  [Wadsworth and Tawn (2012), MacDonald et al. (2011)]

Threshold stability plots



where to set threshold?

northern North Sea: MLEs of ξ are negative

Gulf of Mexico: MLEs of ξ become positive as u increases

Aims

- address bias-variance trade-off based on out-of-sample prediction (cross-validation)
- …using the bin-GP model
- simple graphical diagnostic for single threshold selection
- account for uncertainty in threshold
- develop method than can be generalized: e.g. to multivariate (MV) extremes

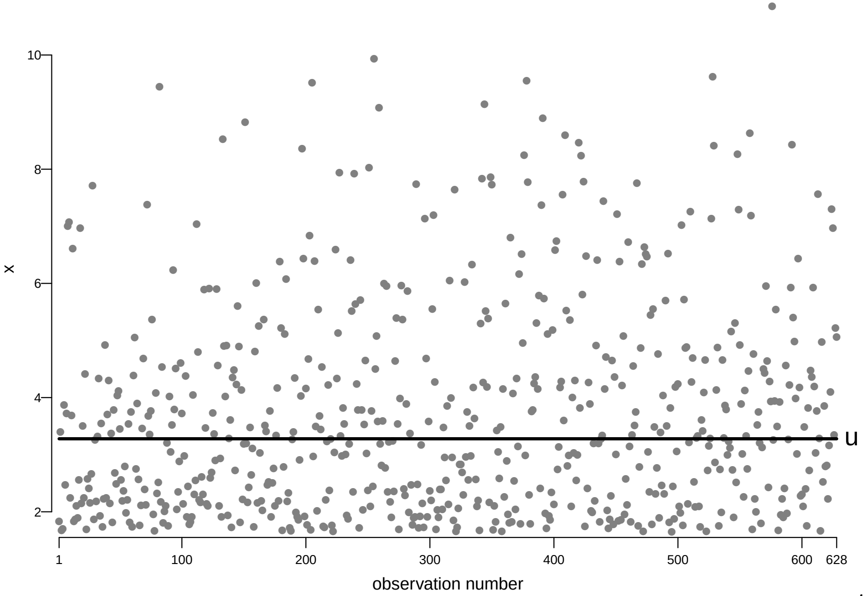
Extrapolation, threshold level and priors

- physical considerations: H_s has a finite upper limit
- unless $\xi \geqslant 0$ is ruled out there is a limit to how far we can extrapolate with realism

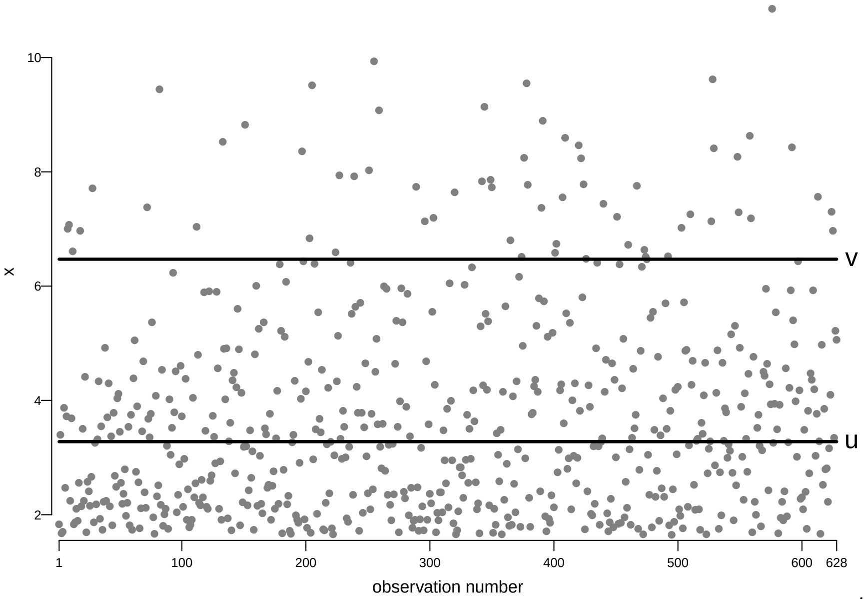
Types of prior

- 'informative', 'full subjective'
- 'regularizing','weakly-informative' (Gelman: "Keeping things unridiculous")
- formal rules: 'weakly-informative' (O'Hagan), 'reference'
  - expecting data to dominate prior
  - may not be the case for high u
  - high $u \rightarrow$ large uncertainty about $\xi \rightarrow$ high posterior probability on large positive $\xi \rightarrow$ greater chance of unrealistic inferences

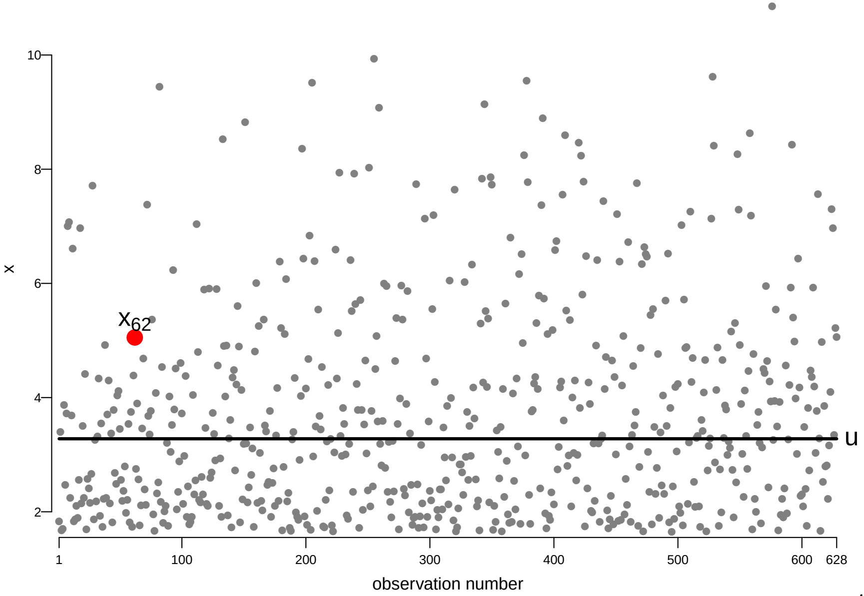
Training threshold u



Validation threshold $v \geqslant u$

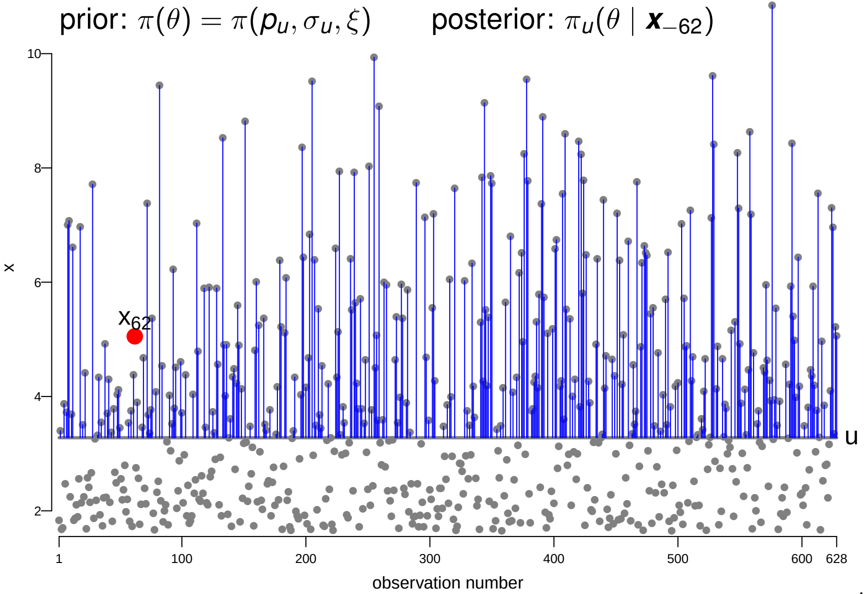


Leave-one-out cross-validation



Infer θ using $\boldsymbol{x}_{-62}$

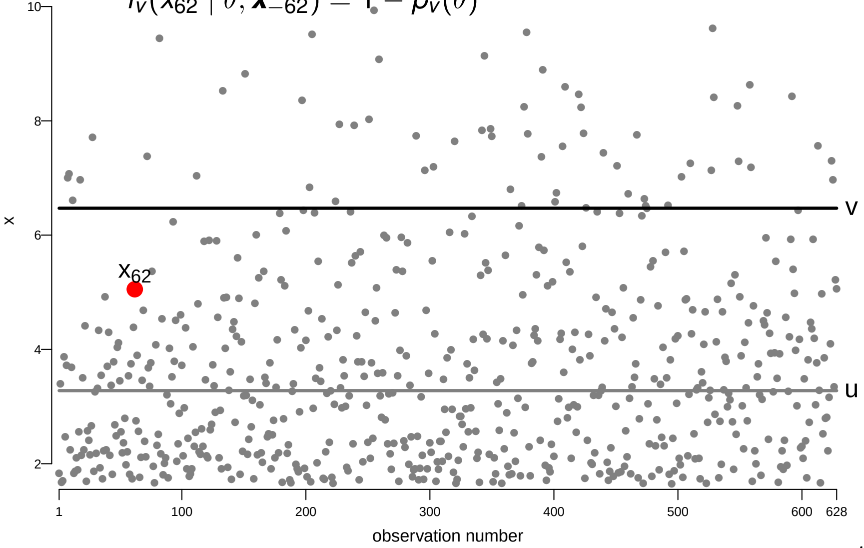
prior: $\pi(\theta) = \pi(p_u, \sigma_u, \xi)$ posterior: $\pi_u(\theta \mid \boldsymbol{x}_{-62})$



Prediction of non-exceedance of v

$$f_v(x_{62} \mid \boldsymbol{x}_{-62}, u) = \int f_v(x_{62} \mid \theta, \boldsymbol{x}_{-62})\, \pi_u(\theta \mid \boldsymbol{x}_{-62})\, \mathrm{d}\theta$$

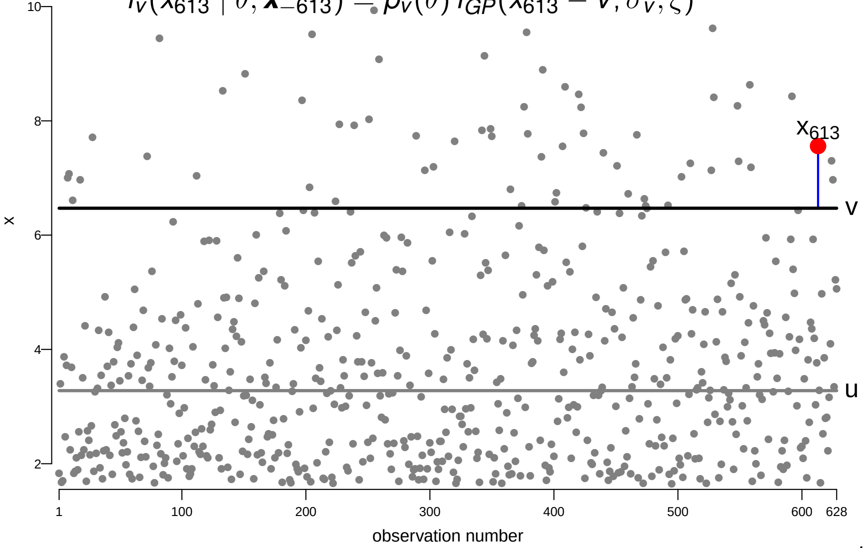
$$f_v(x_{62} \mid \theta, \boldsymbol{x}_{-62}) = 1 - p_v(\theta)$$



Prediction of exceedance of v

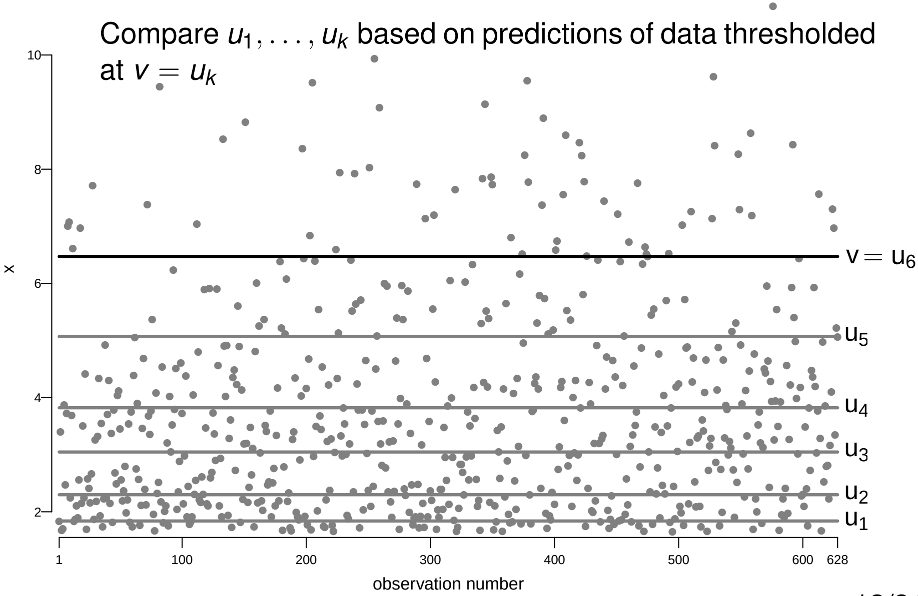
$$f_v(x_{613} \mid \boldsymbol{x}_{-613}, u) = \int f_v(x_{613} \mid \theta, \boldsymbol{x}_{-613}) \, \pi_u(\theta \mid \boldsymbol{x}_{-613}) \, \mathrm{d}\theta$$

$$f_v(x_{613} \mid \theta, \boldsymbol{x}_{-613}) = p_v(\theta) \, f_{GP}(x_{613} - v; \sigma_v, \xi)$$



Training and validation thresholds

Compare $u_1, \ldots, u_k$ based on predictions of data thresholded at $v = u_k$



Choice of $u_1, \ldots, u_k$ and v

Training thresholds $u_1, \ldots, u_k$

- needs to include range over which bias and variance compete
- perhaps the most crucial aspect is the choice of u_k
- rule-of-thumb: have no fewer than 50 excesses (Jonathan and Ewans, 2013)

Validation threshold v

- choose $v = u_k$
- if $v > u_k$ we
  - lose validation information: if $u_k < x \leqslant v$ then value of x is censored
  - ...and gain nothing: predictions of xs greater than v do not change

Comparing thresholds

Sample $\theta_1^{(r)}, \ldots, \theta_m^{(r)}$ from $\pi_u(\theta \mid \boldsymbol{x}_{-r})$ [R-o-U or MCMC]

$$\widehat{f}_v(x_r \mid \boldsymbol{x}_{-r}, u) = \frac{1}{m} \sum_{j=1}^{m} f_v(x_r \mid \theta_j^{(r)})$$

Measure of predictive performance at v when training at u

$$\widehat{T}_v(u) = \sum_{r=1}^{n} \log \widehat{f}_v(x_r \mid \boldsymbol{x}_{-r}, u)$$

Normalize over training thresholds $u_1, \ldots, u_k$

$$w_v(u_i) = \exp\{\widehat{T}_v(u_i)\} / \sum_{j=1}^{k} \exp\{\widehat{T}_v(u_j)\}$$

Threshold weights: $w_v(u_1), \ldots, w_k(u_k)$

Choose threshold with largest threshold weight

Importance sampling (IS)

- IS density $h(\theta)$
  [support of $h(\theta)$ must contain support of $\pi(\theta \mid \boldsymbol{x}_{-r})$]
- Let $q_r(\theta) = \pi_u(\theta \mid \boldsymbol{x}_{-r})/h(\theta)$

$$f_v(x_r \mid \boldsymbol{x}_{-r}, u) = \int f_v(x_r \mid \theta, \boldsymbol{x}_{-r})\, q_r(\theta)\, h(\theta)\, \mathrm{d}\theta, \quad r = 1, \ldots, n$$

IS ratio estimator, based on sample $\theta_1, \ldots, \theta_m$ from $h(\theta)$,

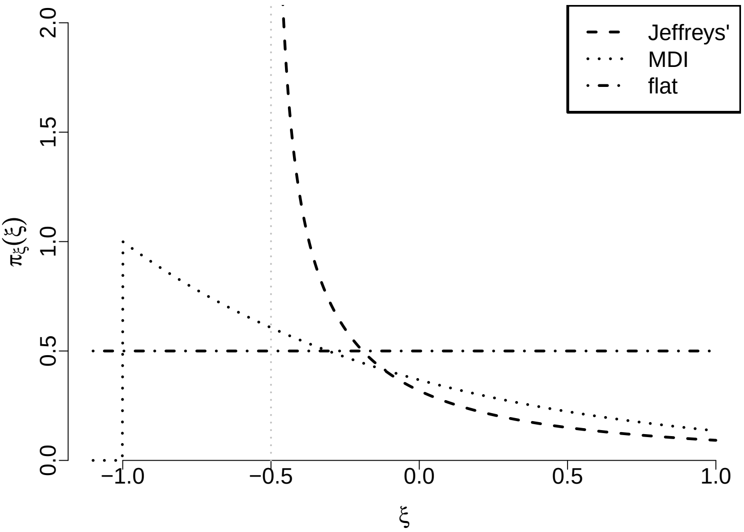
$$\widehat{f}_v(x_r \mid \boldsymbol{x}_{-r}, u) = \frac{\sum_{j=1}^m f_v(x_r \mid \theta_j)\, q_r(\theta_j)}{\sum_{j=1}^m q_r(\theta_j)}$$

Suppose that $x_1 < \cdots < x_n$. Use

$$h(\theta) = \begin{cases} \pi_u(\theta \mid \boldsymbol{x}) & \text{for } r = 1, \ldots, n-1 \\ \pi_u(\theta \mid \boldsymbol{x}_{-n}) & \text{for } r = n \end{cases}$$

. . . so only need to sample from two posteriors

GP priors: $\pi(\sigma_u, \xi) \propto \sigma_u{}^{-1}\pi_\xi(\xi)$



Simulation study with $p_u \in \{0.1, 0.5\}$ and $\xi \in \{-0.2, 0.1\}$

Comparing GP priors: simulation study

$M_N(\theta)$: largest value in N years under a bin-GP(θ)

- $P(M_N(\theta) \leqslant z) = F(z;\theta)^{n_y N}$
- $P(M_N \leqslant z \mid \boldsymbol{x}) = \int F(z;\theta)^{n_y N}\, \pi_u(\theta \mid \boldsymbol{x})\, \mathrm{d}\theta$

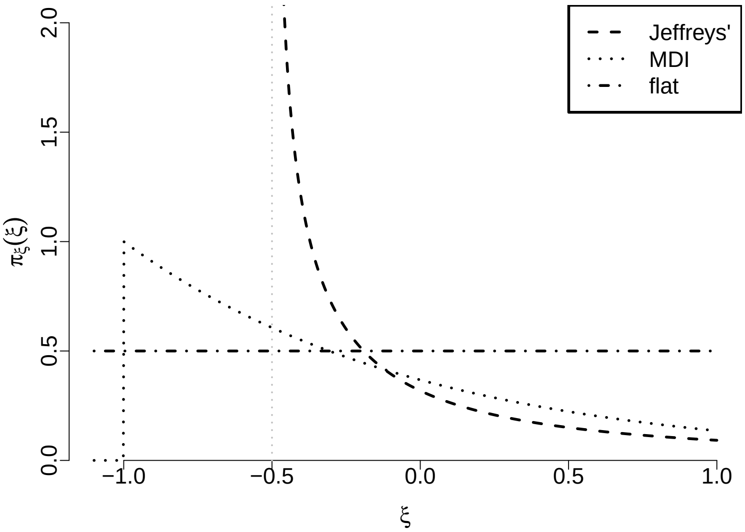
If $P(M_N \leqslant z \mid \boldsymbol{x}) = P(M_N(\theta) \leqslant z)$ then $P(M_N \leqslant M_N(\theta) \mid \boldsymbol{x}) \sim$ U(0,1)

1. simulate bin-GP(p_u, σ_u, ξ) sample $\boldsymbol{x}_{\mathrm{sim}}$, size 500:
   (50 years, 10 observations per year)
2. simulate $m_N(\theta)$ from $F(z;\theta)^{n_y N}$
3. calculate $\widehat{P}(M_N \leqslant m_N(\theta) \mid \boldsymbol{x}_{\mathrm{sim}})$

Repeat: putative sample of size 10,000 from a U$(0, 1)$

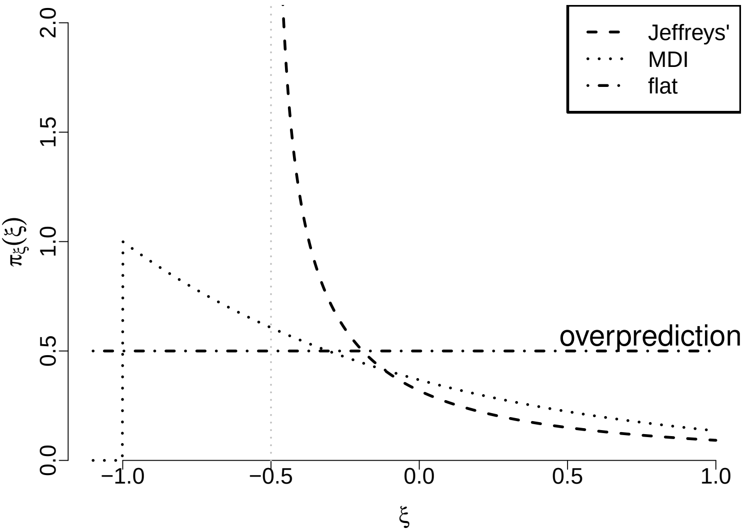
Use Jeffreys' prior, $p_u \sim$ beta$(1/2, 1/2)$.

GP priors: $\pi(\sigma_u, \xi) \propto {\sigma_u}^{-1}\pi_\xi(\xi)$



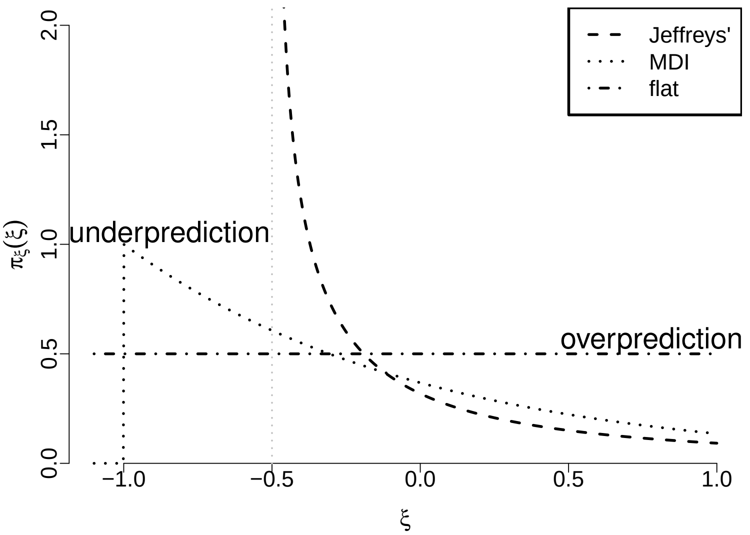
Simulation study with $p_u \in \{0.1, 0.5\}$ and $\xi \in \{-0.2, 0.1\}$

GP priors: $\pi(\sigma_u, \xi) \propto \sigma_u^{-1} \pi_\xi(\xi)$



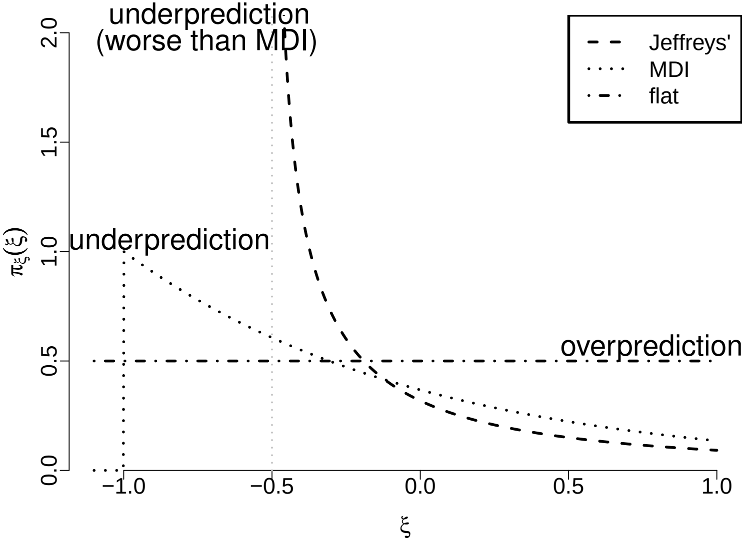
Simulation study with $p_u \in \{0.1, 0.5\}$ and $\xi \in \{-0.2, 0.1\}$

GP priors: $\pi(\sigma_u, \xi) \propto \sigma_u^{-1} \pi_\xi(\xi)$

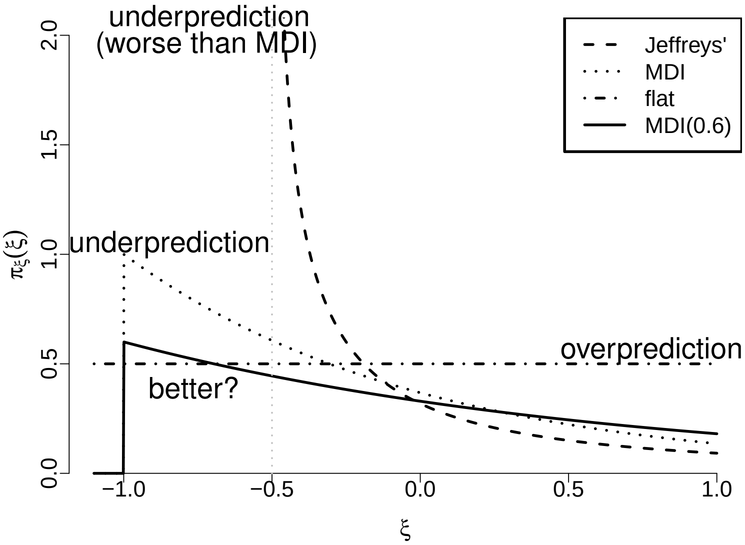


Simulation study with $p_u \in \{0.1, 0.5\}$ and $\xi \in \{-0.2, 0.1\}$

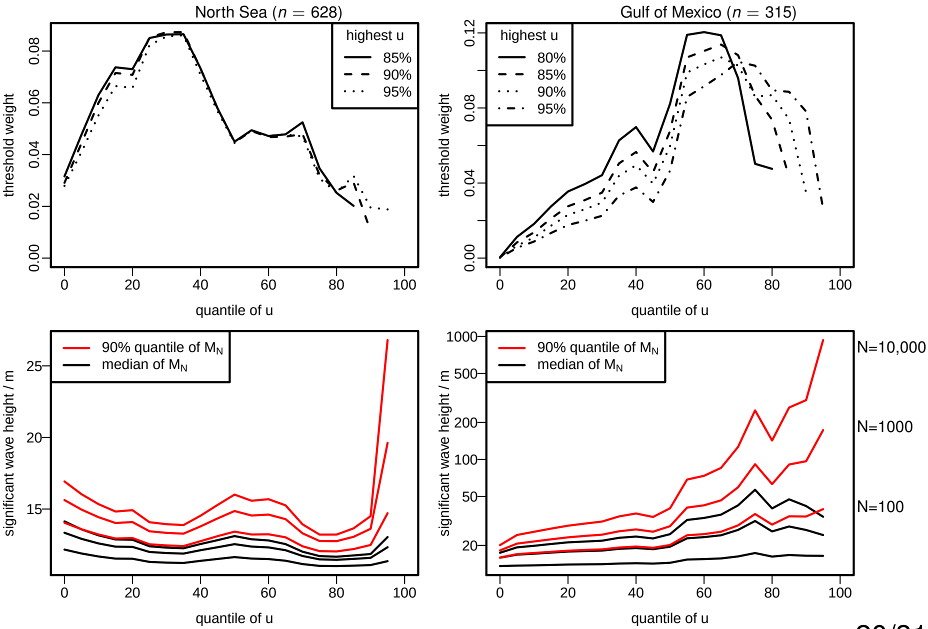
GP priors: $\pi(\sigma_u, \xi) \propto \sigma_u^{-1}\pi_\xi(\xi)$



Simulation study with $p_u \in \{0.1, 0.5\}$ and $\xi \in \{-0.2, 0.1\}$

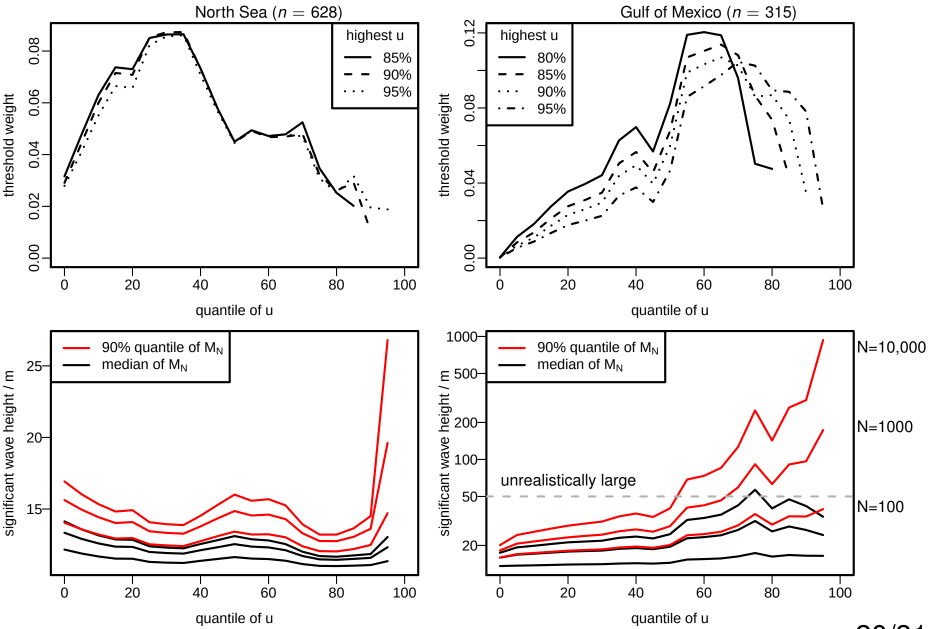


MDI(0.6): $\pi(\sigma_u, \xi) \propto \sigma_u^{-1} 0.6 \, \mathrm{e}^{-0.6(\xi+1)}, \sigma_u > 0, \xi \in \mathbb{R}$

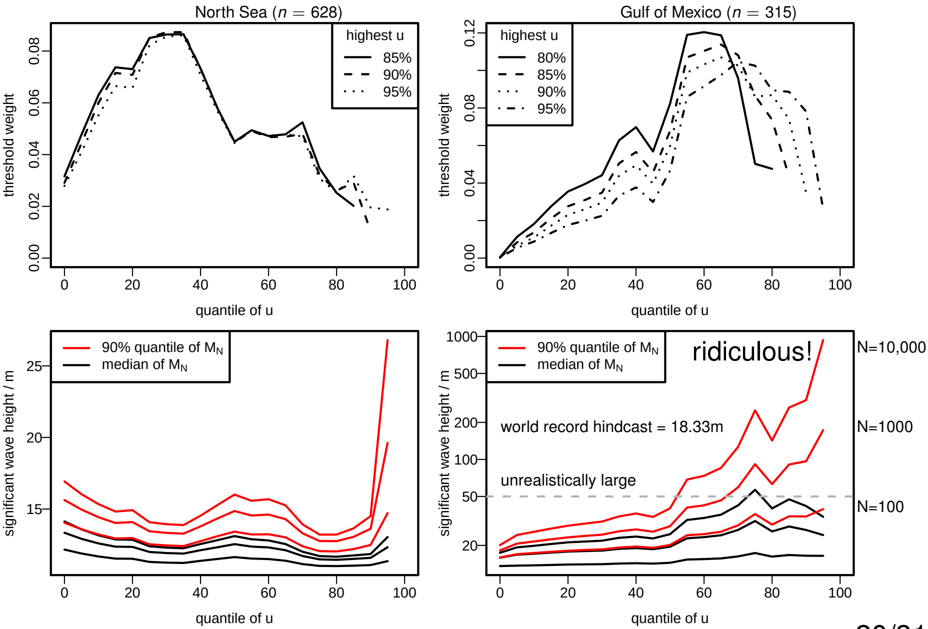


North Sea (n = 628)
highest u
85%
90%
95%
threshold weight
quantile of u
Gulf of Mexico (n = 315)
highest u
80%
85%
90%
95%
threshold weight
quantile of u
90% quantile of M_N
median of M_N
significant wave height / m
quantile of u
90% quantile of M_N
median of M_N
significant wave height / m
quantile of u
N=10,000
N=1000
N=100

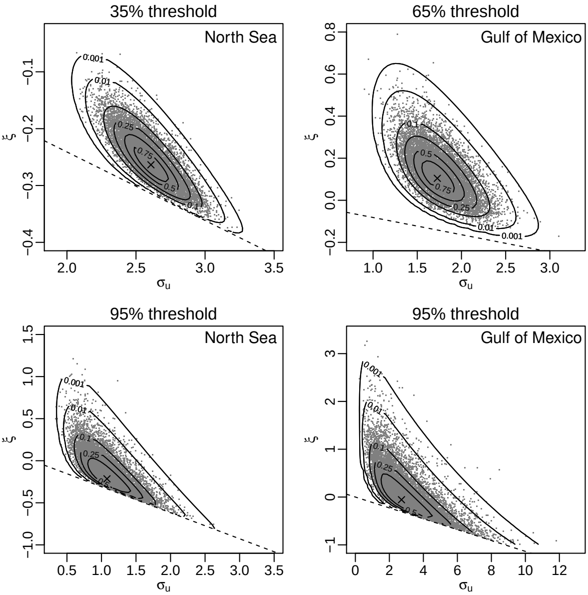
Threshold weights & predictive inference



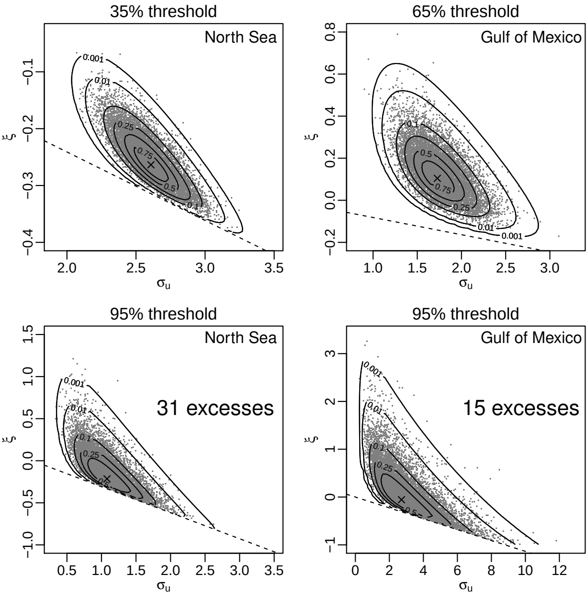
Threshold weights & predictive inference



GP posterior densities



GP posterior densities



Reference priors appropriate only when dominated by information in the data

Expect sig. wave heights to be bounded above ($\xi < 0$)

GoM: $P(\xi > 1/2 \mid \boldsymbol{x}) \approx 0.2$ and $P(\xi > 1 \mid \boldsymbol{x}) \approx 0.05$

Avoid small samples, get more data, give information in prior, don't extrapolate so far into the future

Bayesian model averaging (BMA)

Used by Sabourin et al. (2013) for MV EV models

View k thresholds $u_1, \ldots, u_k$ as defining k competing models

- Prior probabilities: $P(u_i) = 1/k, i = 1, \ldots, k$ [... or something else]
- $\theta_i = (p_i, \sigma_i, \xi_i)$ under model u_i, with prior $\pi(\theta_i \mid u_i)$

Posterior threshold weights:

$$P_v(u_i \mid \boldsymbol{x}) = \frac{f_v(\boldsymbol{x} \mid u_i)\, P(u_i)}{\sum_{i=1}^{k} f_v(\boldsymbol{x} \mid u_i)\, P(u_i)},$$

where

$$f_v(\boldsymbol{x} \mid u_i) = \int f_v(\boldsymbol{x} \mid \theta_i, u_i)\pi(\theta_i \mid u_i)\, \mathrm{d}\theta_i$$

$$\widehat{f}_v(\boldsymbol{x} \mid u_i) = \prod_{r=1}^{n} f_v(x_r \mid \boldsymbol{x}_{-r}, u_i) = \exp\{\widehat{T}_v(u_i)\}$$ [Geisser and Eddy (1979)]

$$\widehat{P}_v(u_i \mid \boldsymbol{x}) = \frac{\exp\{\widehat{T}_v(u_i)\}\, P(u_i)}{\sum_{j=1}^{k} \exp\{\widehat{T}_v(u_j)\}\, P(u_j)} \quad [= w_v(u_i)\,]$$

Simulation study: single u and BMA

- Sample size 500: 50 years, 10 observations per year
- Training thresholds: $(50, 55, \ldots, 90)\%$ sample quantiles
- Validation threshold: 90% sample quantile
- Compare median of predictive distribution of M_N with truth

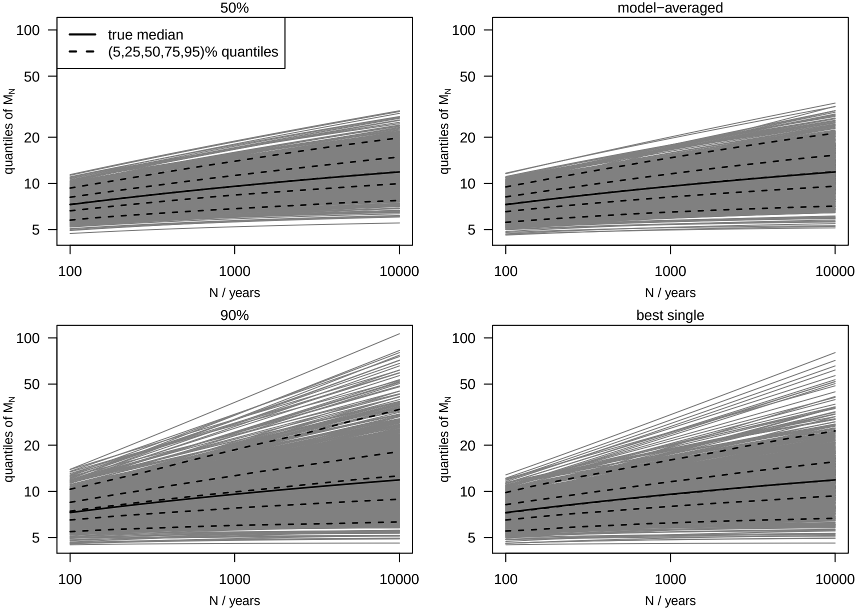
Strategies:

- Threshold known/expected to be good
- Threshold known/expected to be bad
- Threshold with best CV weight
- BMA (averaging inferences over all thresholds)

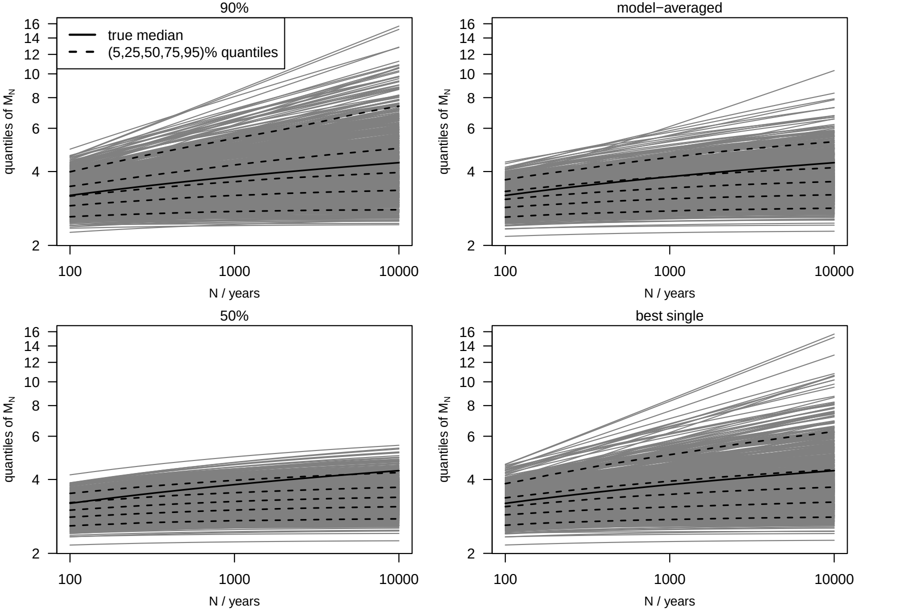
Distributions:

- **exp(1)**: GP(1,0) model holds for all thresholds
- **N(0,1)**: GP false for all u, GP approx. improves as $u \uparrow$
- **Uniform-GP hybrid**: GP holds for $u \geqslant 75\%$ quantile

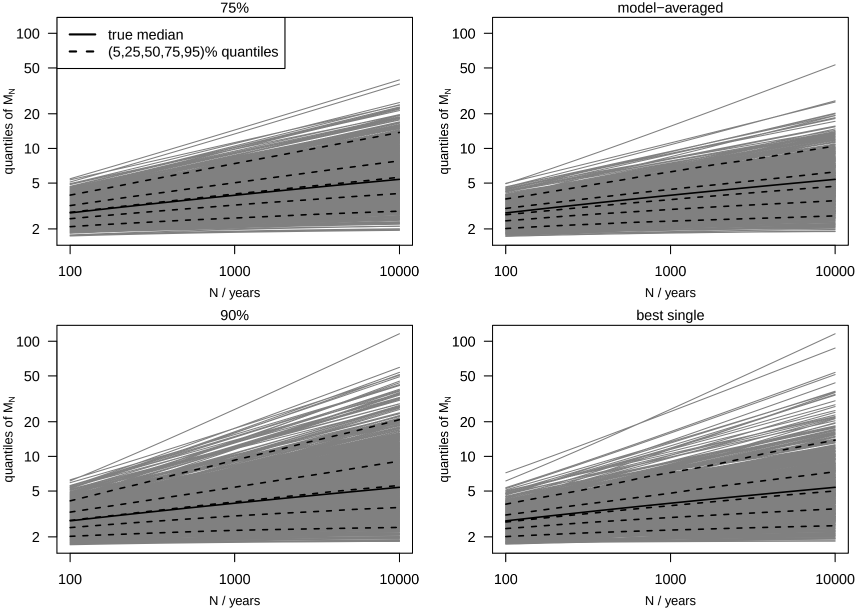
Exponential (1000 simulated datasets)



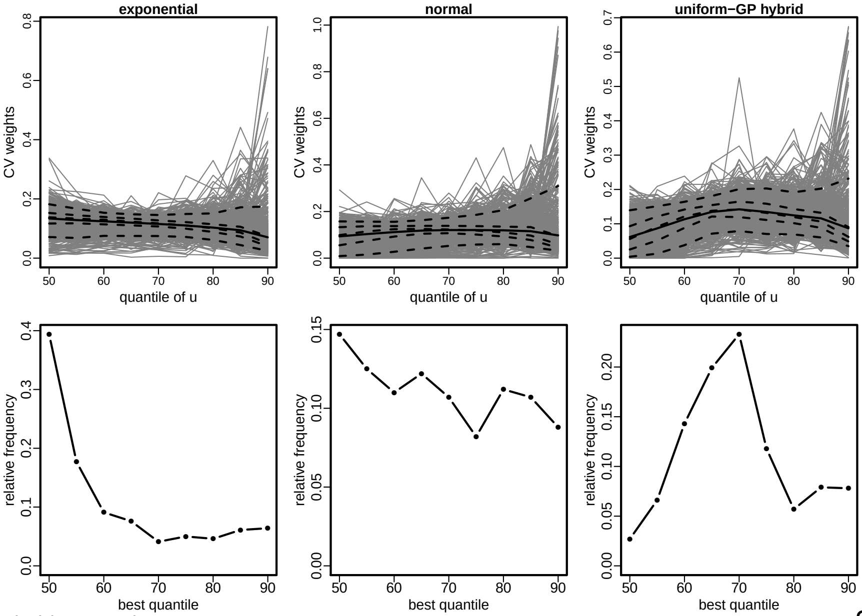
Normal


90%
model−averaged
50%
best single
true median
(5,25,50,75,95)% quantiles
quantiles of M_N
N / years
2
4
6
8
10
12
14
16
100
1000
10000

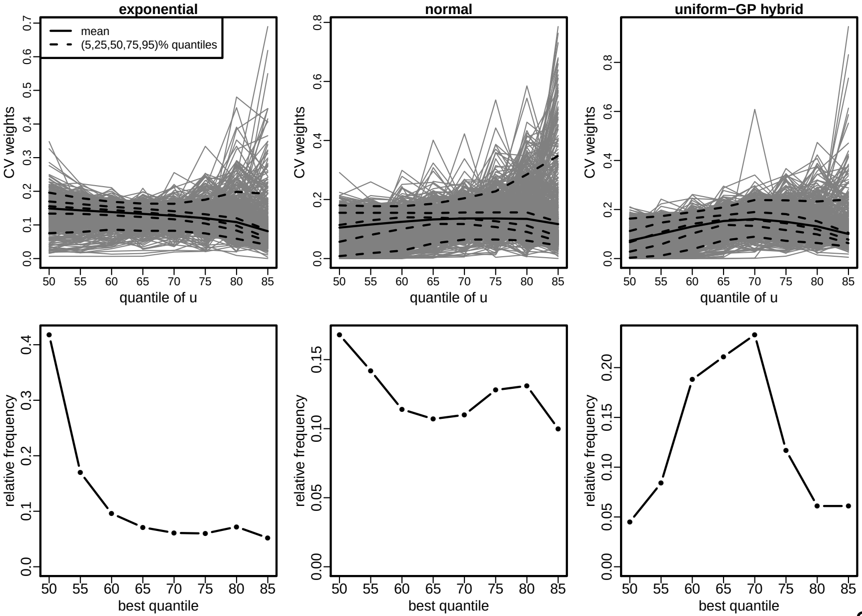
Uniform-GP(ξ=0.1) hybrid



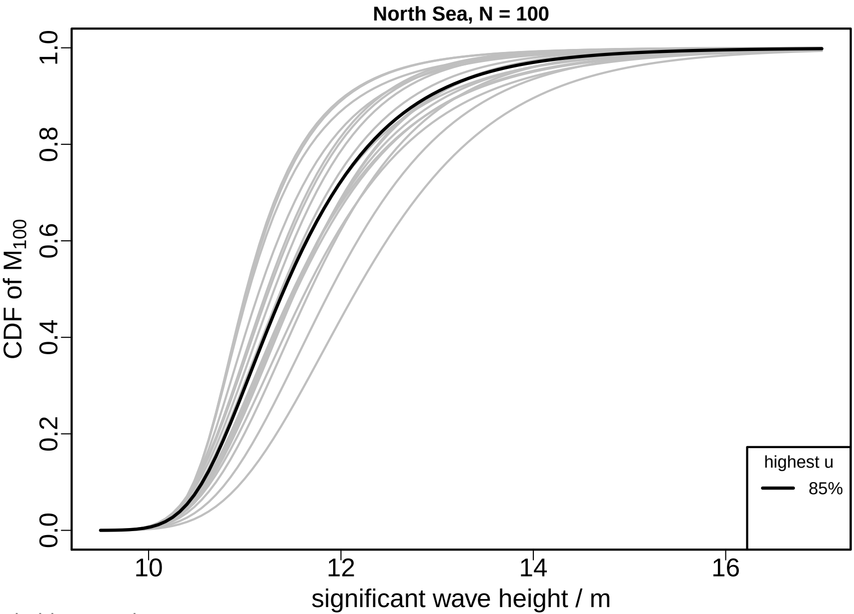
CV weights: $u_k = 90\%$ quantile (50 exc.)



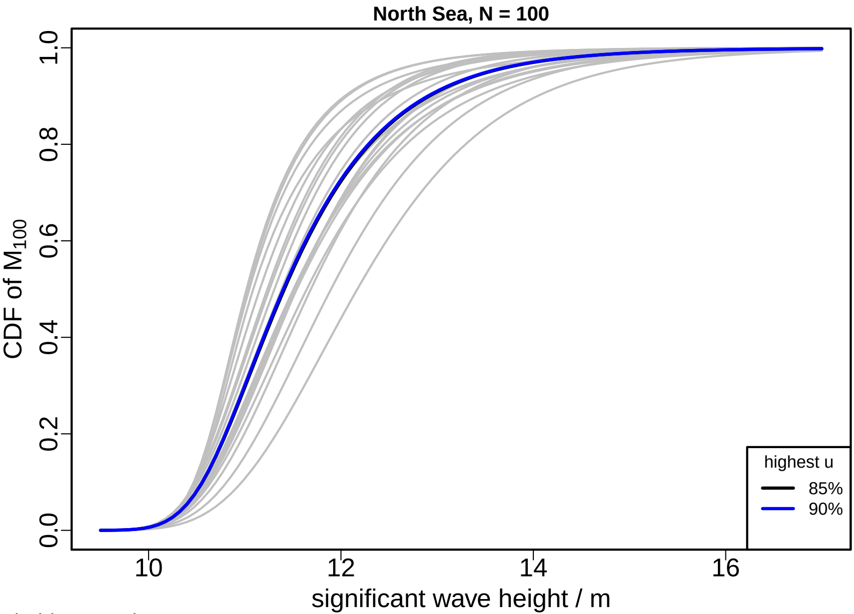
CV weights: $u_k = 85\%$ quantile (75 exc.)



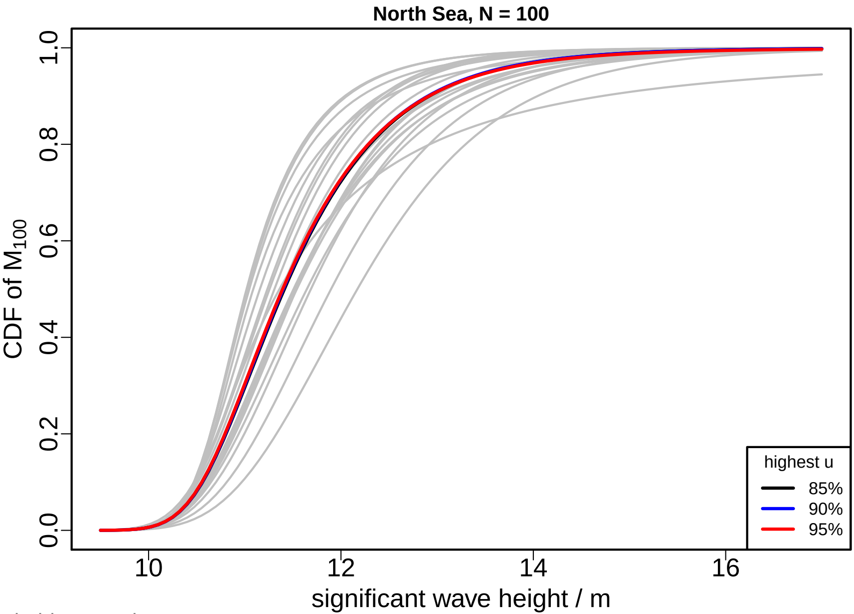
Averaging inferences over thresholds



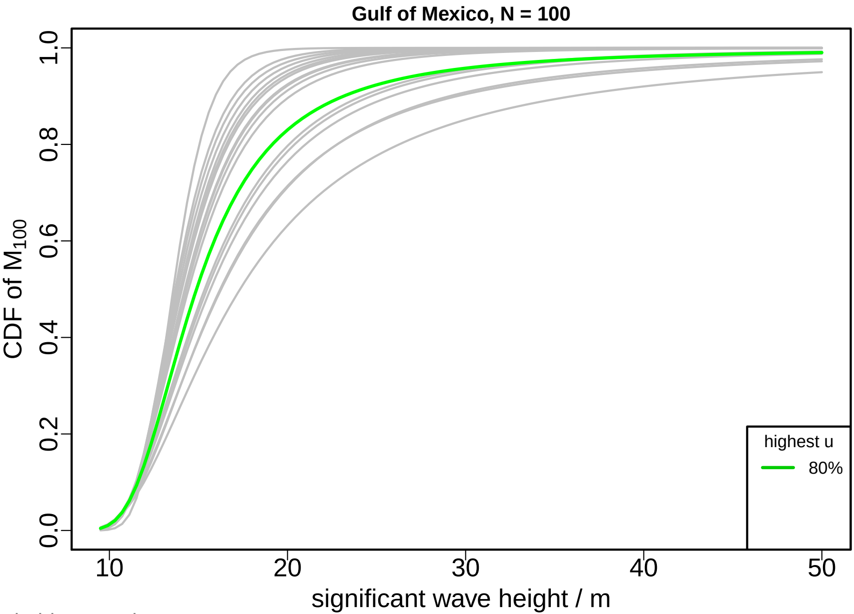
Averaging inferences over thresholds



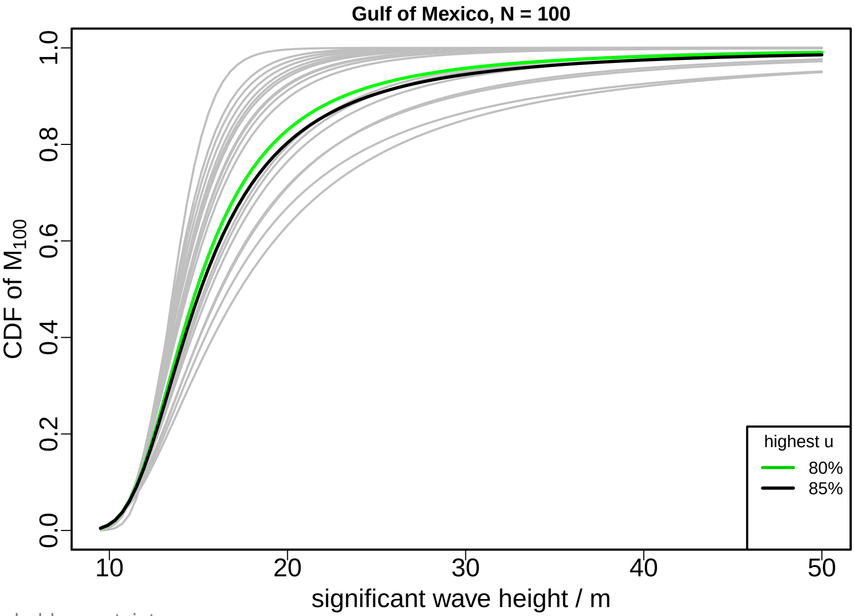
Averaging inferences over thresholds



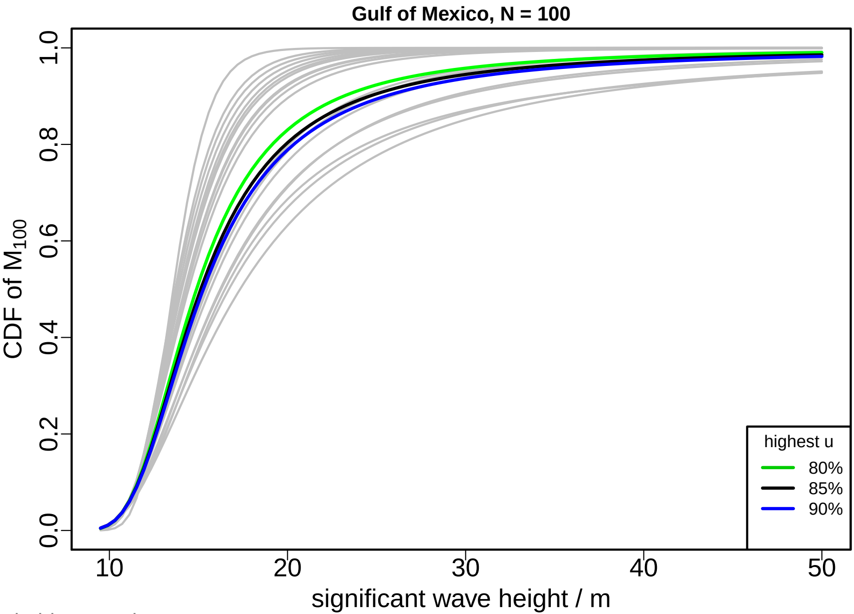
Averaging inferences over thresholds



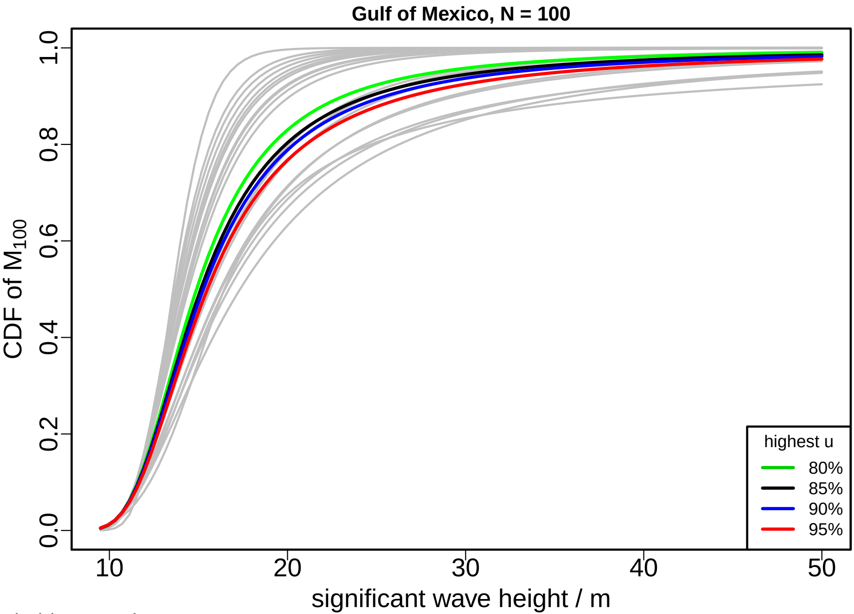
Averaging inferences over thresholds



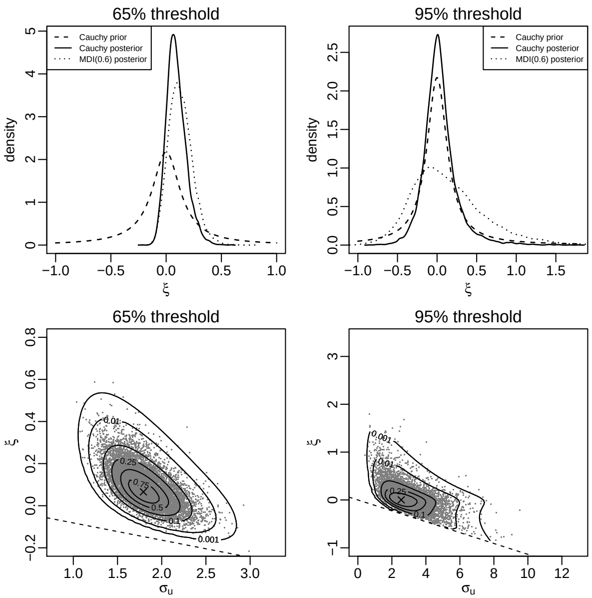
Averaging inferences over thresholds



Averaging inferences over thresholds



A weakly-informative (Cauchy) prior



Gulf of Mexico data

Downweight large values of ξ *a priori*, but give scope for data to contradict the prior

Cauchy: gentle slope in tails

$P(\xi > 1/2) = 0.05$ *a priori*

...based on expert opinion about ratio of 10,000 yr max to 100 yr max

65% threshold : change of prior has little impact

95% threshold : low posterior probability on large ξ

Concluding remarks

- Cross-validation used to address bias-variance trade-off
  - Could automate: pick ‘best’ threshold
- Threshold uncertainty : Bayesian model averaging
- Subjective inputs
  - Priors: reference, weakly-informative, informative
  - Training thresholds $u_1, \ldots, u_k$
- On-going ...
  - serial dependence
  - multivariate extremes
  - covariate effects
  - choice of measurement scale

Concluding remarks

- Cross-validation used to address bias-variance trade-off
  - Could automate: pick 'best' threshold
- Threshold uncertainty : Bayesian model averaging
- Subjective inputs
  - Priors: reference, weakly-informative, informative
  - Training thresholds $u_1, \ldots, u_k$
- On-going ...
  - serial dependence
  - multivariate extremes
  - covariate effects
  - choice of measurement scale

Thank you for your attention

References

Beirlant, J., Y. Goegebeur, J. Teugels, and J. Segers (2004). *Statistics of Extremes : Theory and Applications*. London: Oxford University Press.

Drees, H., L. de Haan, and S. Resnick (2000). How to make a Hill plot. *The Annals of Statistics* **28(1)**, 254–274.

Ferreira, A., L. de Haan, and L. Peng (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics* **37(5)**, 401–434.

Geisser, S. and W. F. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74(365)**, 153–160.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1 (3)**, 515–533.

Jonathan, P. and K. Ewans (2013). Statistical modelling of extreme ocean environments for marine design : a review. *Ocean Engineering*, **62**, 91–109

MacDonald, A., C. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell (2011). A flexible extreme value mixture model. *Comp. Statist. Data Anal.* **55**, 2137–2157.

Northrop, P. J. and Coleman, C. L. (2014) Improved threshold diagnostic plots for extreme value analyses. *Extremes* **17(2)**, 289–303.

O'Hagan, A. (2006). Science, subjectivity and software (comments on the articles by Berger and Goldstein). *Bayesian Analysis*, **1**, 445–450.

Sabourin, A., P. Naveau, and A.-L. Fougres (2013). Bayesian model averaging for multivariate extremes. *Extremes* **16(3)**, 325–350.

Scarrott C, MacDonald A (2012) A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT - Statistical Journal* **10(1)**, 33–60.

Wadsworth, J. and J. Tawn (2012). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *J. Royal Statist. Soc.* **B 74(3)**, 543–567.