Return values subject to uncertainty

Philip Jonathan

Lancaster University, Department of Mathematics & Statistics, UK.
Shell Research Ltd., London, UK.

ISSC-ITTC Workshop
(Slides at www.lancs.ac.uk/∼jonathan)

Acknowledgement

- Lancaster : Jon Tawn and Jenny Wadsworth
- Shell : David Randell

- Return values : Jonathan et al. [2021]
- A better approach : Towe et al. [2021]

Overview

- Return values
- Problem : incorporating estimation (epistemic) uncertainty
- Possible estimators
- Theoretical orderings of estimators
- Simulation study
- Conclusions and a better way

What is a return value?

- Random variable A represents the maximum value of some physical quantity X per annum
- The N-year return value x_N of X is then defined by the equation

$$F_A(x_N) = \Pr(A \leq x_N) = 1 - \frac{1}{N}$$

- Typically $N \in [10^2, 10^8]$ years

An alternative definition

- Random variable A_N represents the N-year maximum value of X
- The N-year return value x'_N of X can be found from F_{A_N} for large N, assuming independent annual maxima since

$$\begin{aligned} F_A(x_N) &= 1 - \frac{1}{N} \\ \Rightarrow F_{A_N}(x_N) &= \left(1 - \frac{1}{N}\right)^N \approx \exp(-1) \end{aligned}$$

- Use $F_{A_N}(x'_N) = \exp(-1)$ to define an alternative return value x'_N

Estimating a return value

- To estimate x_N, we need knowledge of the distribution function F_A of the annual maximum
- We might estimate F_A using extreme value analysis on a sample of independent observations of A
- Typically more efficient to estimate the distribution $F_{X|X>\psi}$ of threshold exceedances of X above some high threshold ψ using a sample of independent observations of X, and use this in turn to estimate F_A and x_N
- How is this done?

Estimating a return value

- Asymptotic theory suggests for high threshold $\psi \in (-\infty, \infty)$ that

$$F_{X|X>\psi}(x|\psi, \sigma, \xi) = 1 - \left(1 + \frac{\xi}{\sigma}(x - \psi)\right)_{+}^{-1/\xi}$$

  for $x > \psi$, shape $\xi \in (-\infty, \infty)$ and scale $\sigma \in (0, \infty)$
- The full distribution of X is $F_X(x) = \tau + (1 - \tau)F_{X|X>\psi}(x)$ where $\tau = \Pr(X \leq \psi)$
- Thus

$$F_A(x) = \Pr(A \leq x) = \sum_{k=0}^{\infty} f_C(k) F_X^k(x)$$

  where C is the number of occurrences of X per annum, with probability mass function f_C to be estimated (say with a Poisson model with parameter λ)

- So what's the problem?

Parameter uncertainty

- x_N can be estimated easily in the absence of uncertainty

- In reality, we estimate parameters λ, ψ, σ and ξ from a sample of data, and we cannot know their values exactly
- How does this epistemic uncertainty affect return value estimates?

- A number of different plausible estimators for return values under uncertainty
- Different estimators perform differently (bias and variance)
- Which estimators are likely to perform reasonably in fairly general circumstances?

- Is it even **sensible or desirable** to estimate return values?

Incorporating uncertainty

- If a distribution $F_{Y|\boldsymbol{Z}}$ of random variable Y is known conditional on random variables $\boldsymbol{Z}$, and the joint density $f_{\boldsymbol{Z}}$ of $\boldsymbol{Z}$ is also known, the unconditional predictive distribution $\tilde{F}_Y$ can be evaluated using

$$\tilde{F}_Y(y) = \int_\zeta F_{Y|\boldsymbol{Z}}(x|\zeta) f_{\boldsymbol{Z}}(\zeta)\, d\zeta$$

- Th expected value of deterministic function g of parameters $\boldsymbol{Z}$ given joint density $f_{\boldsymbol{Z}}$ is

$$E[g(\boldsymbol{Z})] = \int_\zeta g(\zeta) f_{\boldsymbol{Z}}(\zeta)\, d\zeta$$

- $\zeta = (\lambda, \psi, \sigma, \xi)$, $Y = A$ (or $Y = A_N$)

Return value estimated using expected values of parameters, $x_N(E[\boldsymbol{Z}])$

- Motivated by the widespread approach of ignoring uncertainty in parameters ζ for estimation of return values

$$x_{N1} = x_N(E[\boldsymbol{Z}])$$

- Simply plug in the mean parameter estimates $E[\boldsymbol{Z}] = \int_\zeta \zeta f_{\boldsymbol{Z}}(\zeta)\, d\zeta$

- A related estimator converging to x_{N1} with increasing N, would be $x'_N(E[\boldsymbol{Z}])$
- Similar choices of estimator could be based on mode$(\boldsymbol{Z})$, median$(\boldsymbol{Z})$, ...

Expected quantile of distribution of A with NEP $1 - 1/N$, $E[x_N(\mathbf{Z})]$

$$x_{N2} = E[x_N(\mathbf{Z})] = \int_{\zeta} x_N(\zeta) f_{\mathbf{Z}}(\zeta)\, d\zeta$$

- Solve for quantile $x_N(\zeta)$ of the distribution of A with NEP $1 - 1/N$ for a large number of parameter choices ζ, and then take the mean
- A related estimator $E[x'_N(\mathbf{Z})]$ is the expected quantile of distribution of A_N with NEP $\exp(-1)$ (converges to x_{N2} as N increases)

Quantile of predictive distribution of A with NEP $1-1/N$, $\tilde{Q}_A(1-1/N)$

- First calculated the predictive distribution $\tilde{F}_A$

$$\tilde{F}_A(x) = \int_\zeta F_{A|\mathbf{Z}}(x|\zeta) f_{\mathbf{Z}}(\zeta)\, d\zeta$$

  integrating over parameter uncertainty

- Then find the $1-\frac{1}{N}$ quantile of $\tilde{F}_A$

$$\tilde{F}_A(x_{N3}) = 1 - \frac{1}{N}$$

- Write briefly as $x_{N3} = \tilde{Q}_A(1-1/N)$, where $\tilde{Q}_A$ is the predictive quantile (or inverse) function corresponding to $\tilde{F}_A$
- This would be the “obvious go-to” Bayesian estimate

Quantile of predictive distribution of A_N with NEP $\exp(-1)$, $\tilde{Q}_{A_N}(\exp(-1))$

- First calculated the N-year predictive distribution F_{A_N}

$$\tilde{F}_{A_N}(x) = \int_{\zeta} F_{A_N|\mathbf{Z}}(x|\zeta) f_{\mathbf{Z}}(\zeta)\, d\zeta$$

integrating over parameter uncertainty

- Then find the $\exp(-1)$ quantile of $\tilde{F}_{A_N}$

$$\tilde{F}_{A_N}(x_{N4}) = \exp(-1)$$

- Write briefly as $x_{N4} = \tilde{Q}_{A_N}(\exp(-1))$, where $\tilde{Q}_{A_N}$ is the predictive quantile function corresponding to $\tilde{F}_{A_N}$

Summary

- Take average over uncertain parameters, and plug in to return value calculation
$$q_1 = x_{N1} = x_N(E[\boldsymbol{Z}])$$
- Calculate return value for all sets of estimates independently, then take average
$$q_2 = x_{N2} = E[x_N(\boldsymbol{Z})]$$
- Calculate "average" annual maximum distribution, then take $1 - 1/N$ quantile
$$q_3 = x_{N3} = \tilde{Q}_A(1 - 1/N)$$
- Calculate "average" N-year maximum distribution, then take $\exp(-1)$ quantile
$$q_4 = x_{N4} = \tilde{Q}_{A_N}(\exp(-1))$$

- Without parameter uncertainty, all these estimators are equivalent
- With parameter uncertainty, all these estimators are different

Theoretical inequalities

- We can show that the estimators have orderings, e.g.

	Inequality	Condition
I1	$q_3 \geq q_4$	Always
I2	$q_2 > q_0$	$\xi_1 > \max(\xi_0, 0)$
I3	$q_3 > q_2$	$1 > \xi_1 > \max(\xi_0, 0)$
I4	$q_1 > q_0$	$\xi_0, \xi_1 < 0, \sum_i \sigma_i / \sum_i(-\xi_i) > \sigma_0/(-\xi_0)$
I5	$q_2 > q_0$	$\xi_0, \xi_1 < 0, (1/m) \sum_i(\sigma_i/(-\xi_i)) > \sigma_0/(-\xi_0)$
I6	$q_3 > q_0$	$\xi_0, \xi_1 < 0, \max_{k \in (1,2,\ldots,m)}(\sigma_k/(-\xi_k)) > \sigma_0/(-\xi_0)$

- GP parameter set $\mathcal{Z} = \{\xi_i, \sigma_i\}_{i=1}^m$ ordered s.t. $\xi_1 = \mathrm{argmax}_i(\xi_i)$
- ξ_0 and σ_0 are the true underlying data-generating parameters
- Condition $N \to \infty$ applies to all these cases
- Not specific to maximum likelihood estimation of GP parameters

Simulation study

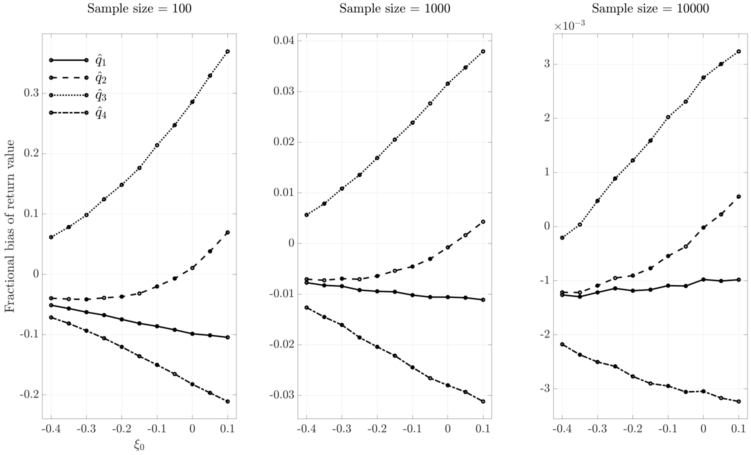
- Data from GP distribution, $\xi_0 \in -0.4, -0.35, ..., +0.1$, $\sigma_0 = 1$
- Sample sizes $n = 10^2, 10^3$ and 10^4, and $\lambda = 10^2$ events annually
- Return periods $N = 10^2$ and 10^4 years
- $m = 10^5$ sample realisations

- Fractional bias in return value
$$\frac{q_j}{q_0} - 1$$

- Fractional bias in exceedence probability
$$\Pr(A > q_j) - \frac{1}{N}$$

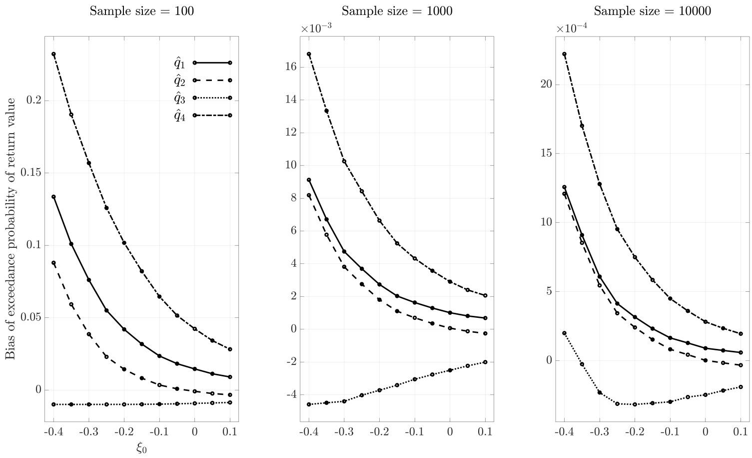
- Fractional bias in log exceedence probability
$$\log_{10}(\Pr(A > q_j)) - \log_{10}(\frac{1}{N})$$

Fractional bias in return value, $N = 100$ years, $\lambda = 100$ annually



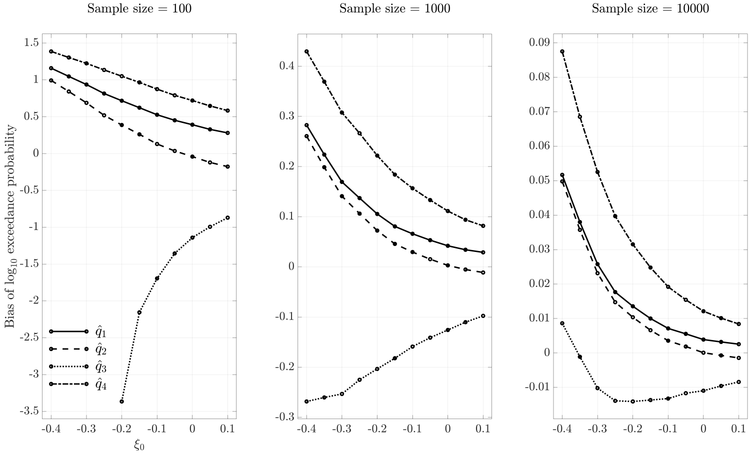
- $n/\lambda \Rightarrow 1, 10, 100$ years of data to predict 100 year return value
- $q_3 > q_2 > q_1 > q_4$, and q_2 shows lowest bias

Fractional bias in exceedance probability, $N = 100$, $\lambda = 100$



- q_3 underestimates exceedance probability
- q_2 shows lowest bias for $\xi_0 > -0.2$

Fractional bias in log exceedance probability, $N = 100$, $\lambda = 100$



- q_3 underestimates exceedance probability (huge for small n)
- q_2 generally shows lowest bias

Findings

- Different estimators $\Rightarrow$ different estimates and systematic bias
- Why? $E(g(\boldsymbol{Z})) \neq g(E(\boldsymbol{Z}))$ in general
- Worse for small sample size n, and $\to 0$ as $n \to \infty$

- $E[x_N(\boldsymbol{Z})]$ less biased, estimated from F_A or F_{A_N}
- $\tilde{Q}_A(1 - 1/N)$ is "obvious go-to", but poor performance
  $\Rightarrow$ Intuitively, better to take averages at end of calculation **only**
  $\Rightarrow$ Decision-theoretic approach better

- Maximum likelihood estimation used here; other inference schemes examined also; lots of other estimators possible!
- Uncertainties in return values are also large!
- Do safety factors elsewhere in the design process require return values with assumed characteristics?
- Discussion of differences in return values only makes sense when they have been calculated using the same approach

Decision-theoretic approach

- Consider structural loading R over some period for structural strength r_0
- Define a loss function, e.g. $L(R|r_0) = \mathbb{I}(R > r_0)$
- Estimate conditional distribution $F_{R|\mathbf{Z}}(r|\boldsymbol{\zeta})$ for uncertain environmental parameters $\boldsymbol{\zeta}$ (computationally challenging!)
- Calculate expected loss

$$E(L|r_0) = \int_{\zeta}\int_{r} L(r|r_0) f_{R|\mathbf{Z}}(r|\boldsymbol{\zeta}) f_{\mathbf{Z}}(\boldsymbol{\zeta})\, dr d\boldsymbol{\zeta}$$

- Adjust r_0 so that $E(L|r_0)$ is acceptably small

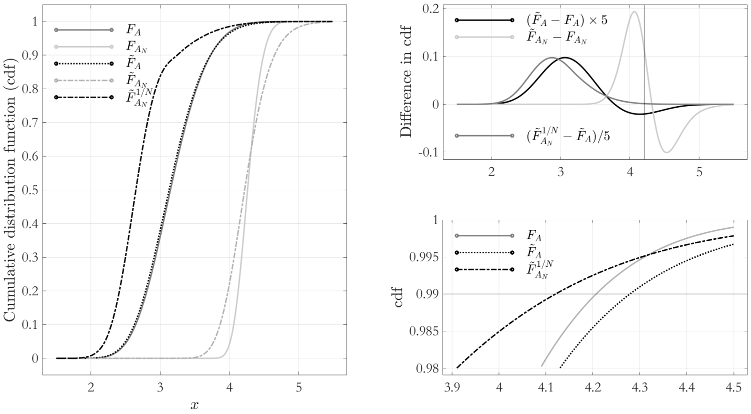
- Propagate uncertainty in full through design calculation, and integrate over uncertain parameters at the very end
- **No need for return values**

References

P. Jonathan, D. Randell, J. Wadsworth, and J.A. Tawn. Uncertainties in return values from extreme value analysis of peaks over threshold using the generalised Pareto distribution. *Ocean Eng.*, 220:107725, 2021.

R. Towe, E. Zanini, D. Randell, G. Feld, and P. Jonathan. Efficient estimation of distributional properties of extreme seas from a hierarchical description applied to calculation of un-manning and other weather-related operational windows. *Submitted to Ocean Engineering, draft at www.lancs.ac.uk/∼jonathan*, 2021.

q_3, q_4 and their uncertainties



- $\xi_0 = -0.2$, sample size $n = 10^3$ and $N = 100$ years