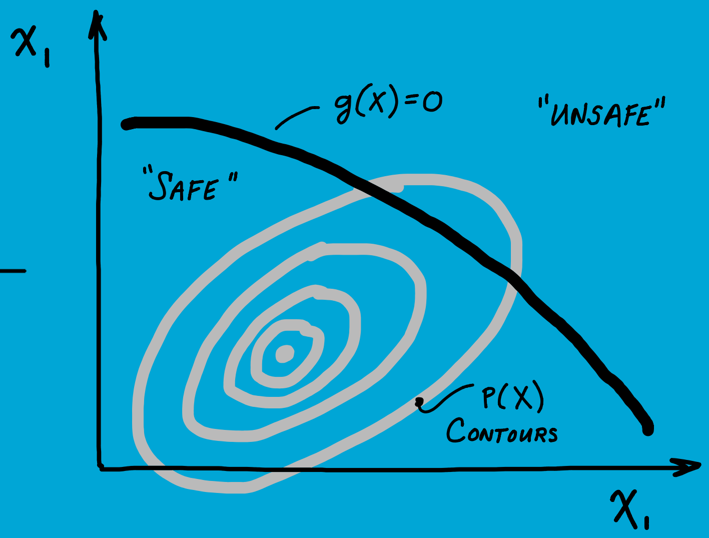
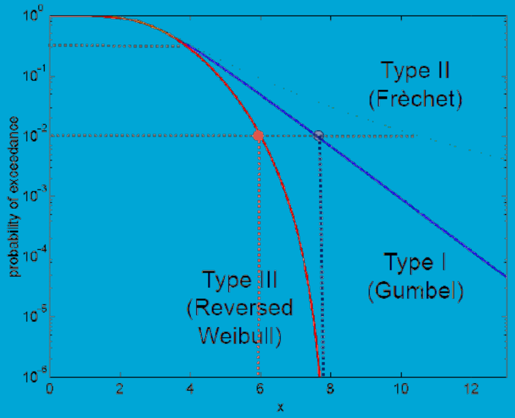
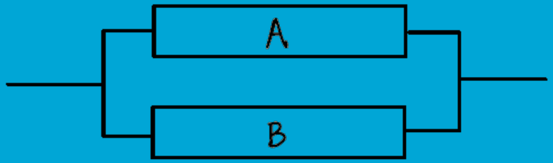
CIEM42X0 Probabilistic Design

Hydraulic and Offshore Structures (HOS) Track

Civil Engineering MSc Program

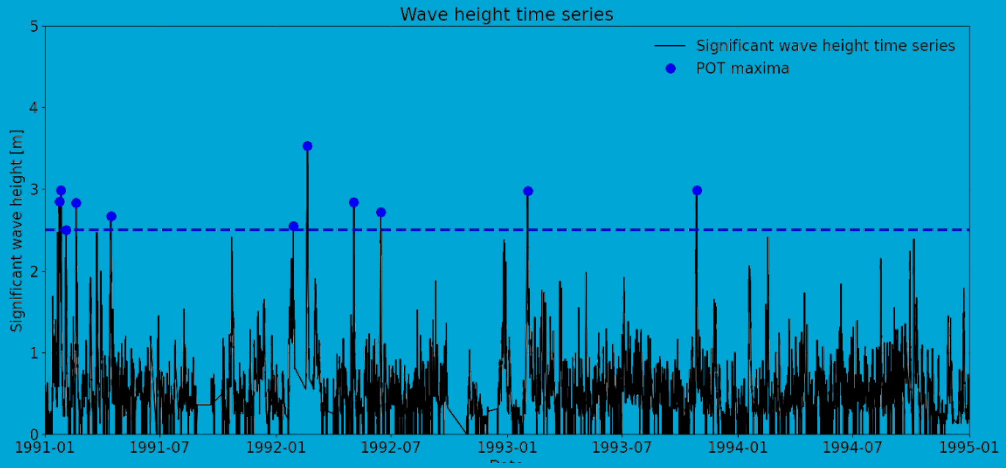






Extreme Value Analysis: basics

Patricia Mares Nasarre



TU Delft

What have you seen so far?

1. Identify what is an **extreme value** and apply it within the engineering context
2. Interpret and apply the concept of **return period and design life**
3. Apply **extreme value analysis** to datasets

Learning objectives

1. Identify what is an **extreme value** and apply it within the engineering context
2. Interpret and apply the concept of **return period and design life**
3. Apply **extreme value analysis** to datasets
4. Apply techniques to **support the threshold selection**

TU Delft

Join the Vevox session

Go to **vevox.app**

Enter the session ID: **125-461-830**

Or scan the QR code

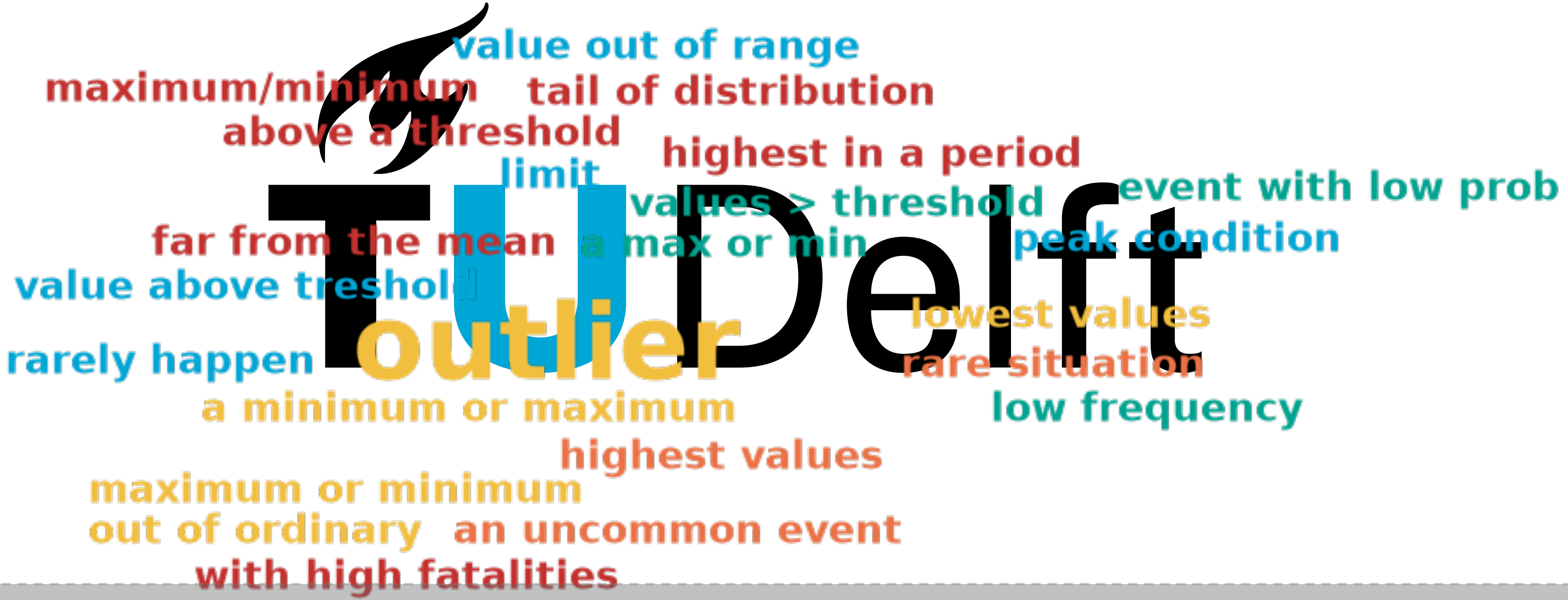
What is an extreme in probability theory?

20

Join at: **vevox.app** ID: **125-461-830**

Showing Results

What is an extreme in probability theory?



RESULTS SLIDE

Extremes and Extreme Value Analysis

An **extreme observation** is an observation that **deviates from the average observations.**

Infrastructures and systems are designed to **withstand extreme conditions (ULS)**.

- Breakwater → wave storm
- Flood defences → floods, droughts

To properly design and assess infrastructures and system **we need to characterize the uncertainty of the loads**.

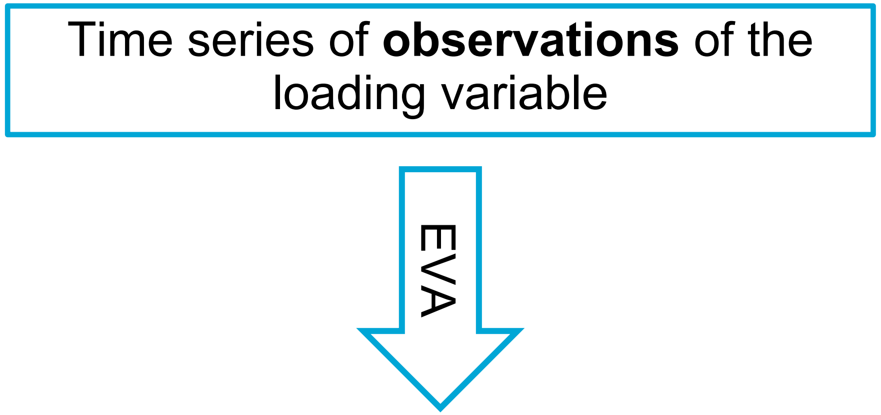
© Getty Images/C. Somodevilla

Extreme Value Analysis

Based on historical observed extremes (limited)…

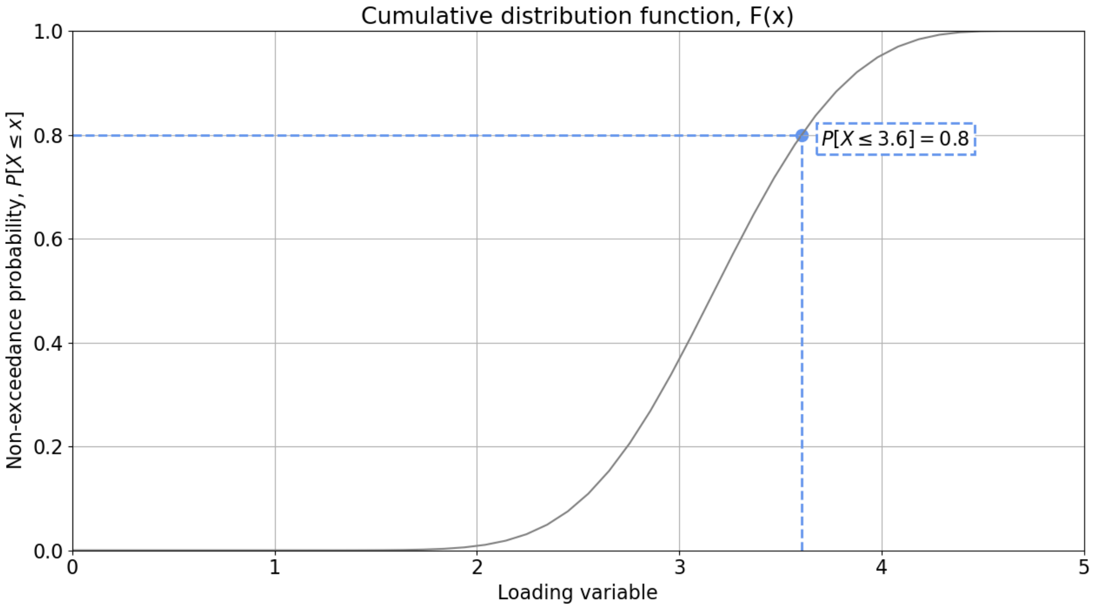
- Allows us to **model** the stochastic behaviour of extreme events

- Allows us to **infer** extremes we have not observed yet (extrapolation)



Time series of **observations** of the loading variable

EVA



Cumulative distribution function, F(x)

Imagine you are working with a continuous variable, such as the discharge in a river (Q). You want to use the cumulative distribution function (CDF) to compute the probability of observing a discharge Q=100 m3/s. Which probability would you obtain?

Exceedance probability, P[Q>100] 0%

Non-exceedance probability, P[Q<100] 0%

Probability of the event, P[Q=100] 0%

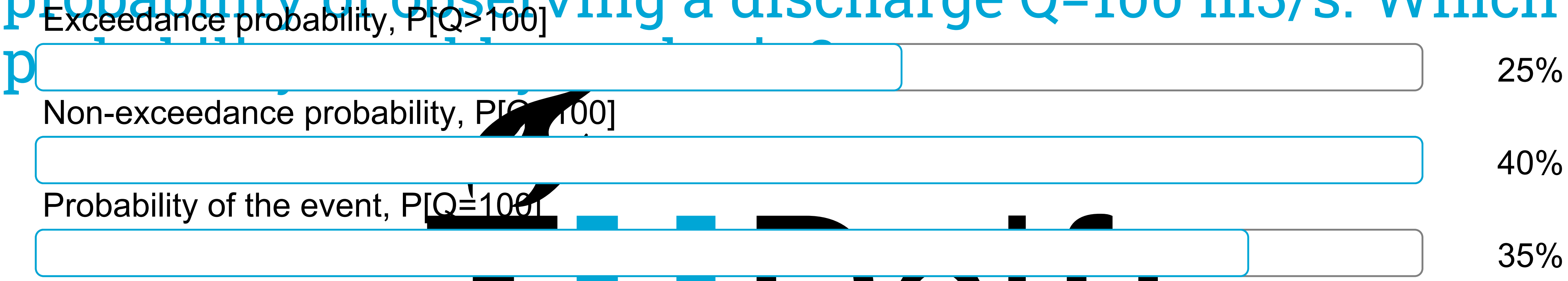
20

Join at: **vevox.app**

ID: **125-461-830**

Showing Results

Imagine you are working with a continuous variable, such as the discharge in a river (Q). You want to use the cumulative distribution function (CDF) to compute the probability of observing a discharge Q=100 m3/s. Which probability would you calculate?

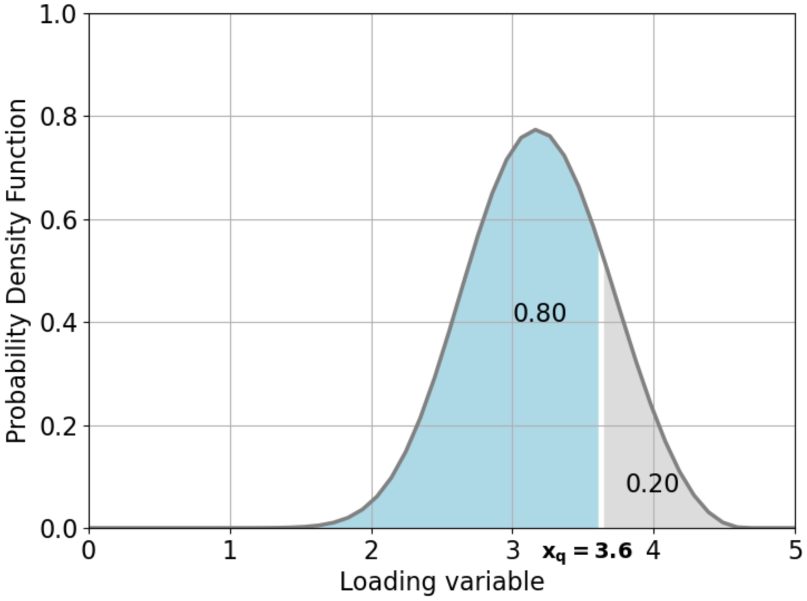


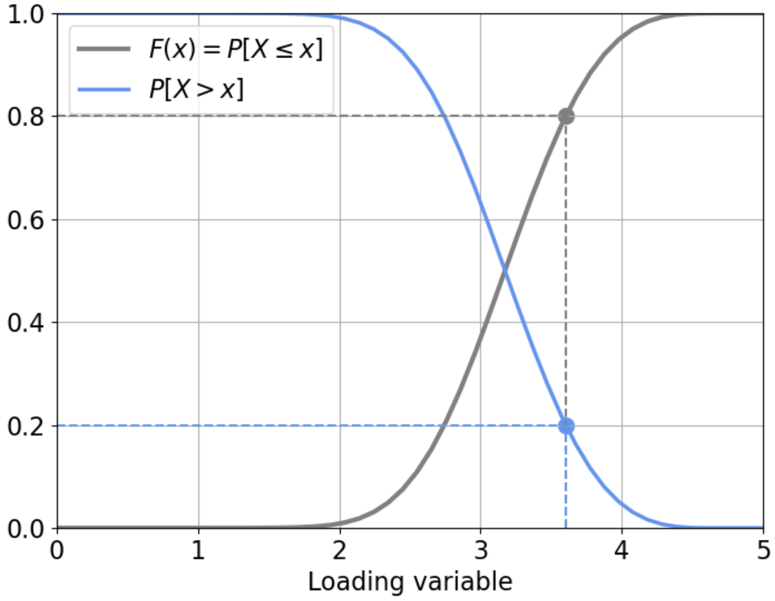
RESULTS SLIDE

Percentile and Exceedance Probability

Consider $\boldsymbol{x_q}$such that $\mathbf{Pr}\left(\boldsymbol{X} \leq \boldsymbol{x_q}\right) = \boldsymbol{F}\left(\boldsymbol{x_q}\right) = \boldsymbol{q}$

- $\boldsymbol{x_q}$ is the **q$^{\text{th}}$ – percentile**
- $\mathbf{Pr}\left(\boldsymbol{X} > \boldsymbol{x_q}\right) = 1 - F\left(x_q\right) = \mathbf{1} - \boldsymbol{q} = \boldsymbol{p}$ is the **exceedance probability**





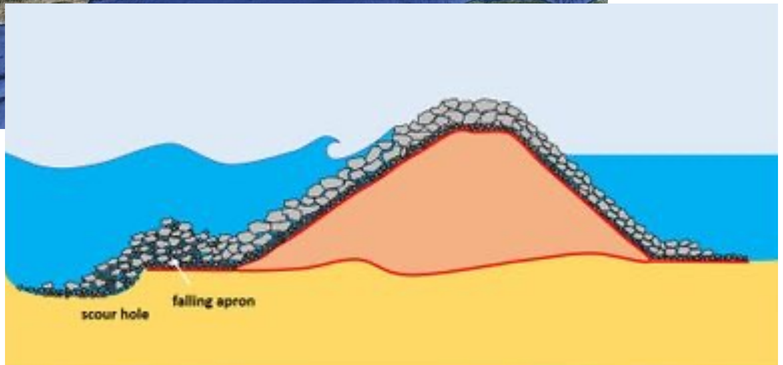
80$^{\text{th}}$-percentile: $x_q = 3.60$

$P\,r(X \leq 3.6) = 0.8$

Exceedance probability

$P\,r\left(X > x_q\right) = 0.20$

Example case: intervention in the Mediterranean coast



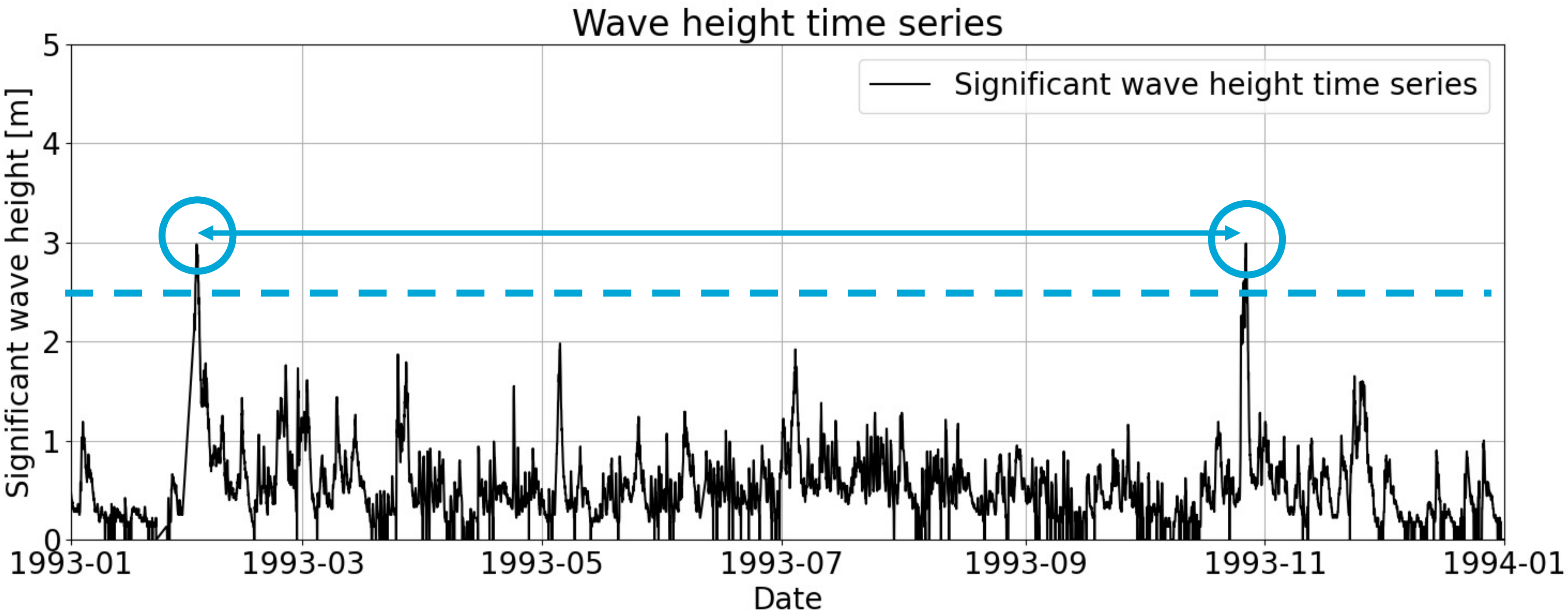
- It may be a coastal structure, a water intake, the restoration of a sandy beach, between others.
- Here: **design a mound breakwater**
- Mound breakwater must resist wave storms → H_s
- ***But which one?***

Return Period

The Return Period (T_R) is the expected time between exceedances. "In other words, we have to make, on average, *$1/p_{f,y}$* trials in order that the event happens once" (Gumbel) or **wait *$1/p_{f,y}$* years before the next occurrence**, being $p_{f,y}$ the exceedance probability.

Assumption of stationarity:
Every year the probability of the event being higher/lower than the threshold is always the same

$$\boldsymbol{T_R(t) = \frac{1}{p_{f,y}}}$$



Design requirements – Binomial distribution

$$\boxed{T_R = \frac{1}{\mathrm{p_{f,y}}} = \frac{1}{1-(1-\mathrm{p_{f,DL}})^{1/DL}}}$$

- DL = 20 years
- $\mathrm{p_{f,DL}}$ = 0.20

$$T_R = \frac{1}{\mathrm{p_{f,y}}} = \frac{1}{1-(1-0.2)^{1/20}} \approx 90\ years$$

$$\mathrm{p_{f,y}} \approx 0.011$$

Learning objectives

✓ 1. Identify what is an **extreme value** and apply it within the engineering context

✓ 2. Interpret and apply the concept of **return period and design life**

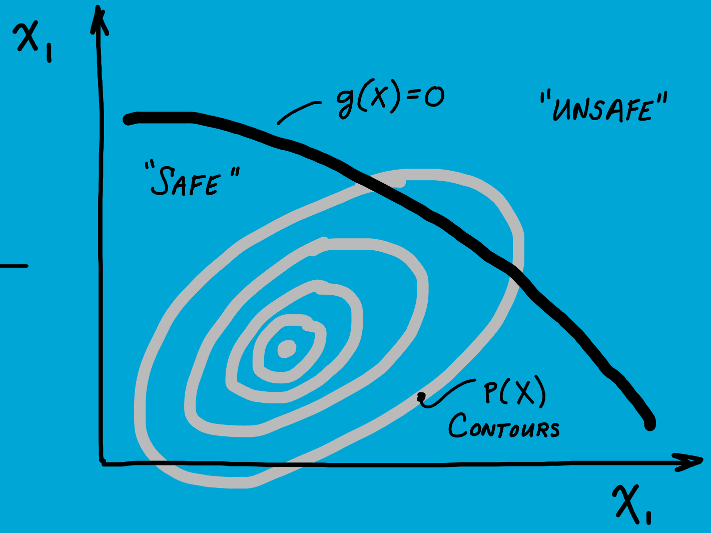
3. Apply **extreme value analysis** to datasets

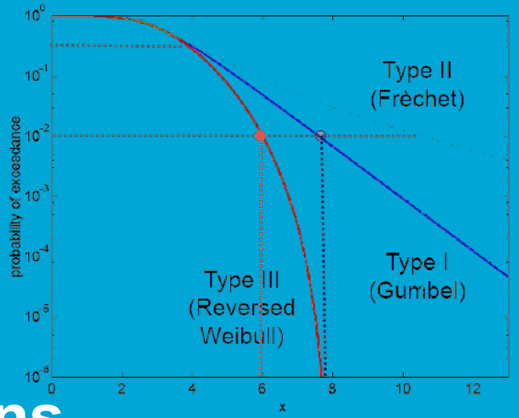
4. Apply techniques to **support the threshold selection**

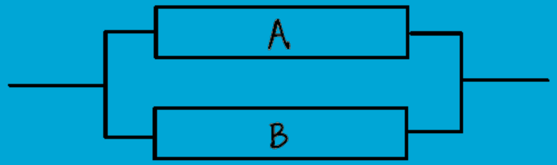
CIEM42X0 Probabilistic Design

Hydraulic and Offshore Structures (HOS) Track

Civil Engineering MSc Program

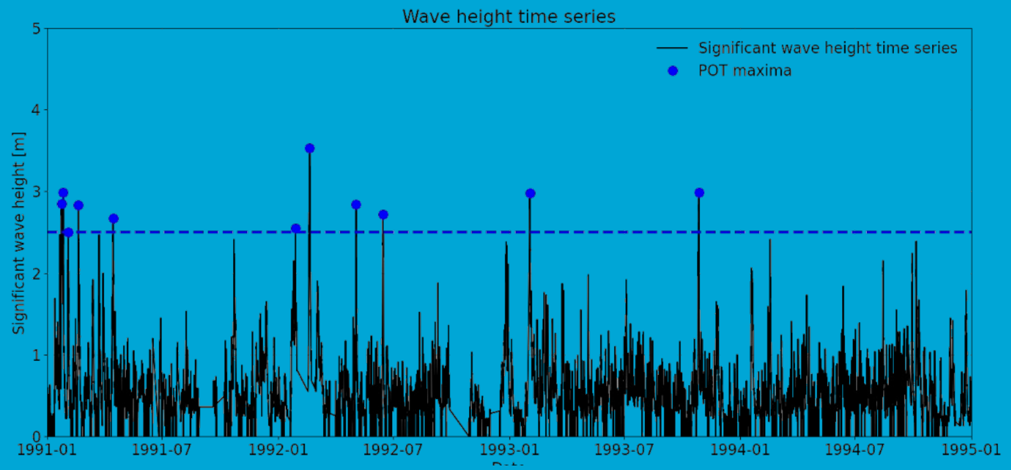






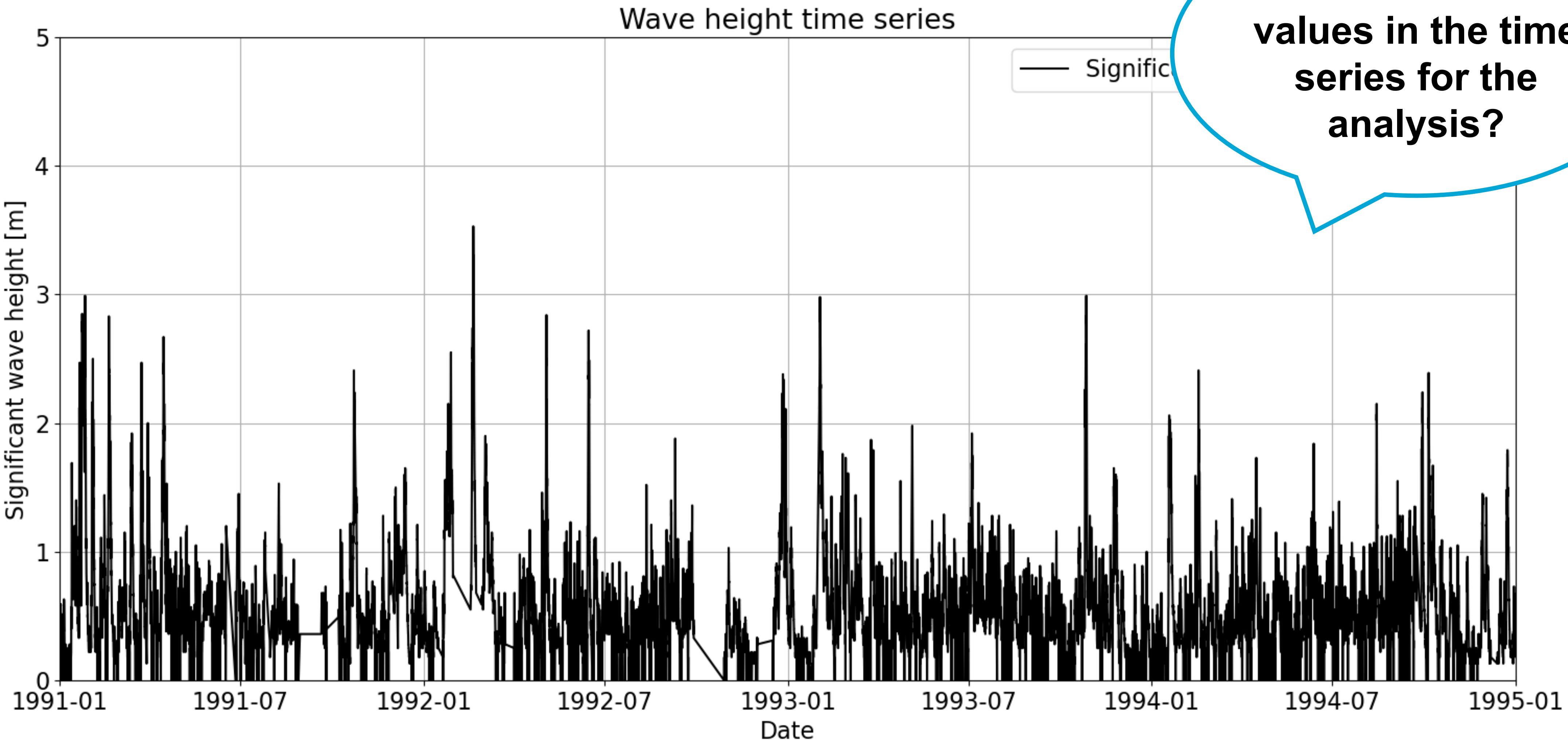
EVA: Sampling and distributions

Patricia Mares Nasarre



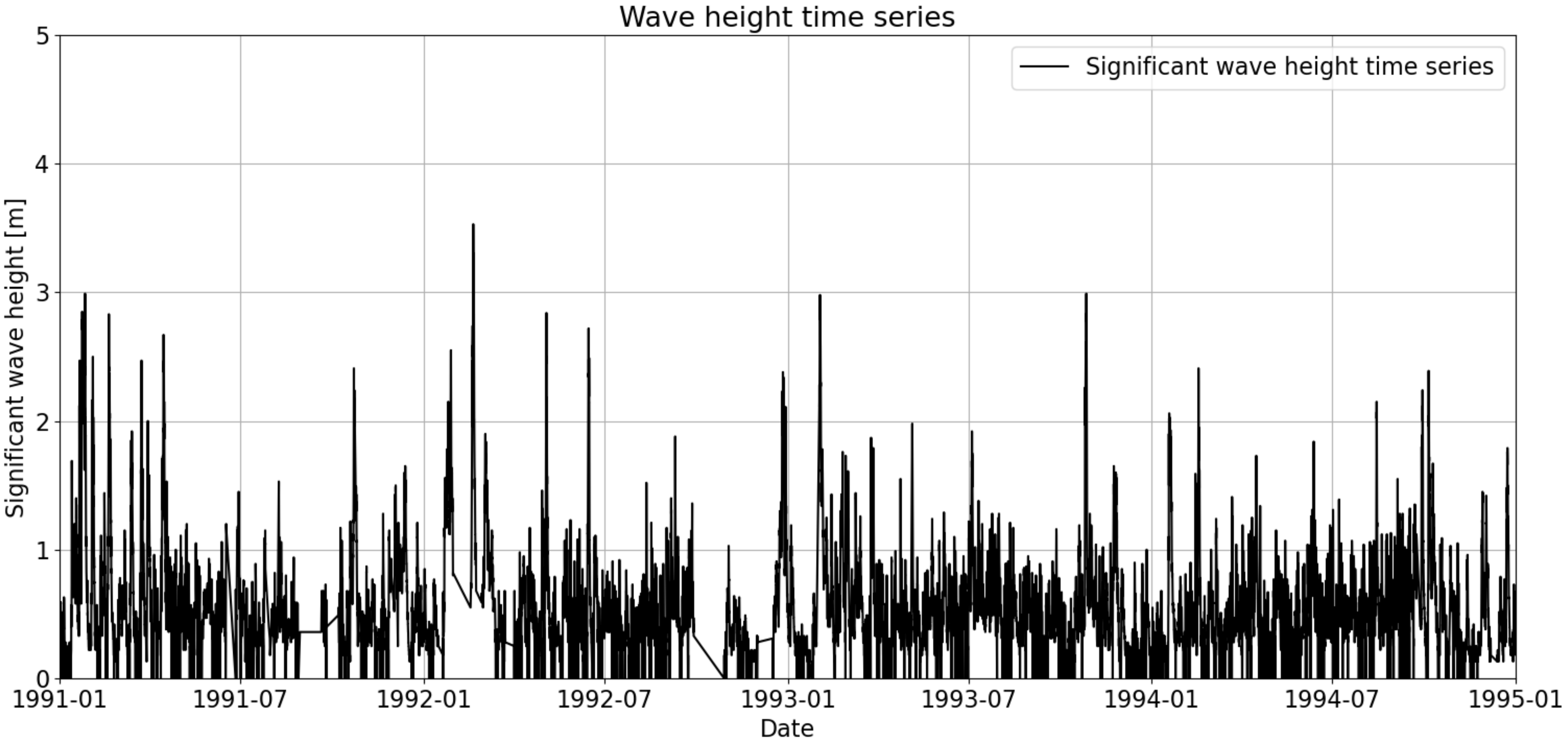
TU Delft

Time series



Can I use all the values in the time series for the analysis?

We need to sample extreme values!



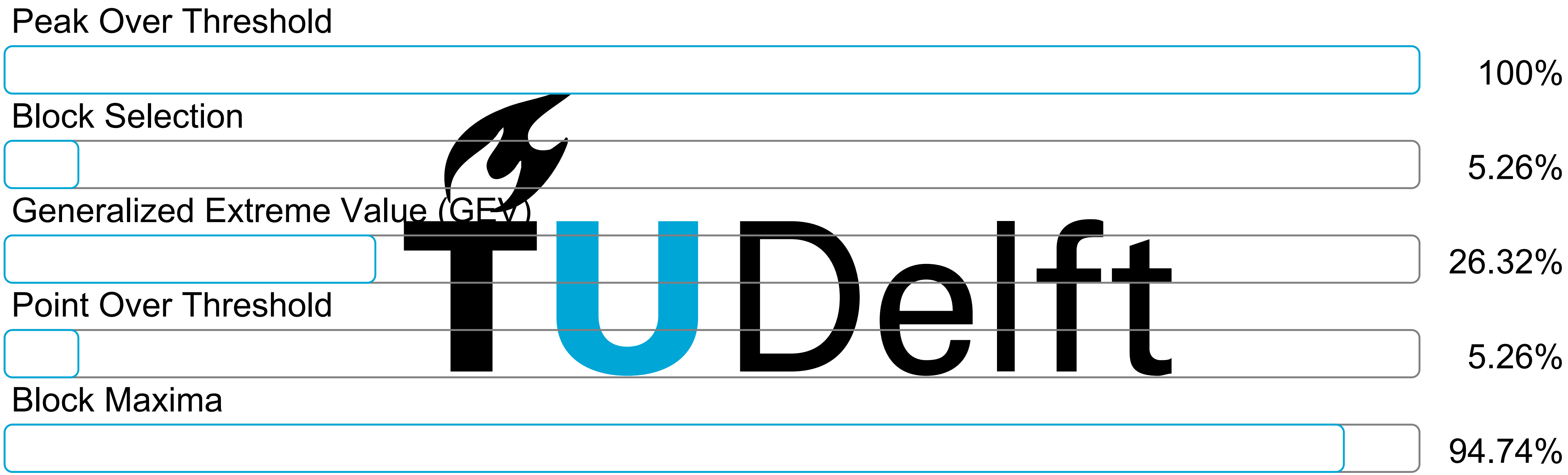
19/22 Join at: **vevox.app** ID: **125-461-830** Question slide

Which one of the following options is a sampling technique for extremes? You may select more than one option.

Peak Over Threshold	0%
Block Selection	0%
Generalized Extreme Value (GEV)	0%
Point Over Threshold	0%
Block Maxima	0%

19 Join at: **vevox.app** ID: **125-461-830** Showing Results

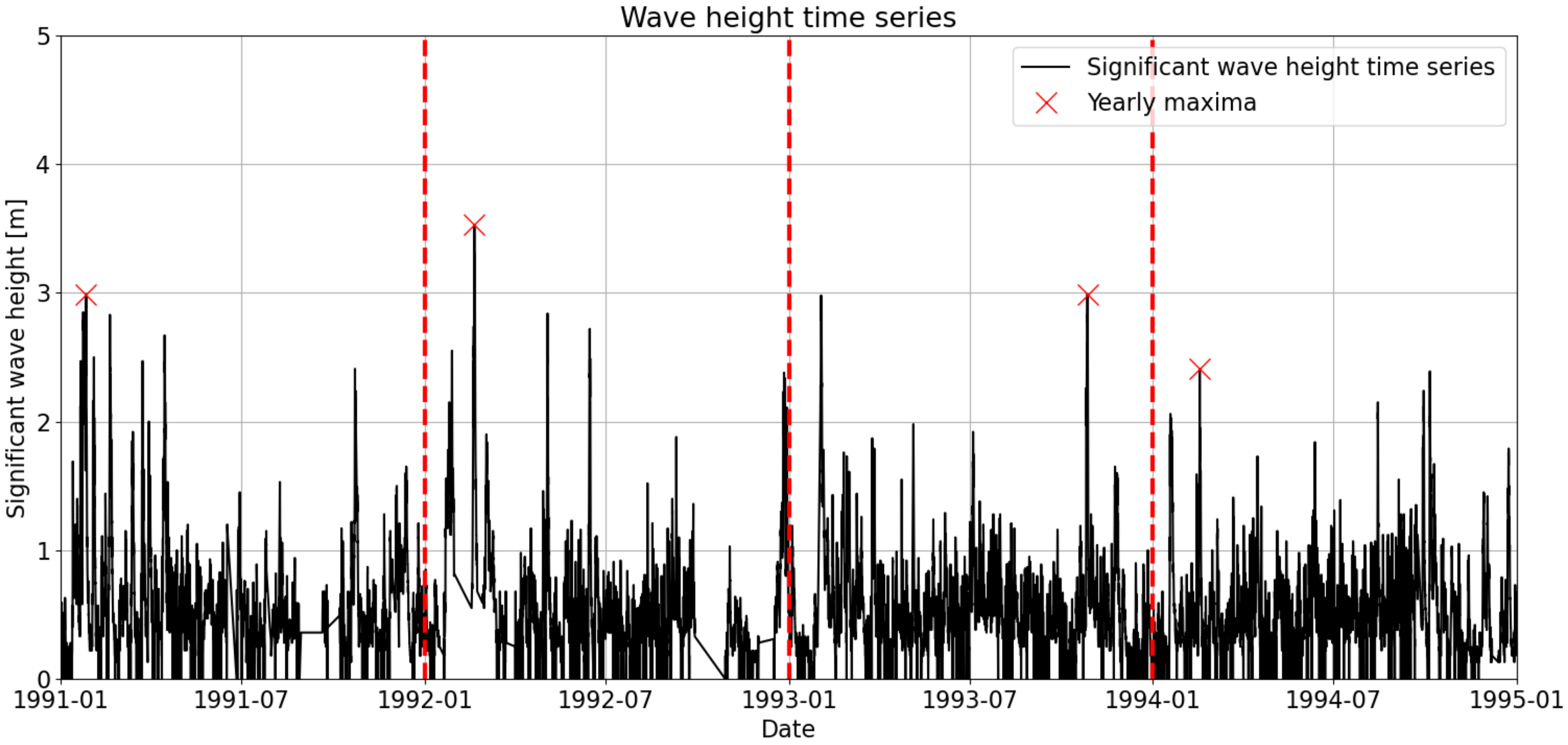
Which one of the following options is a sampling technique for extremes? You may select more than one option.



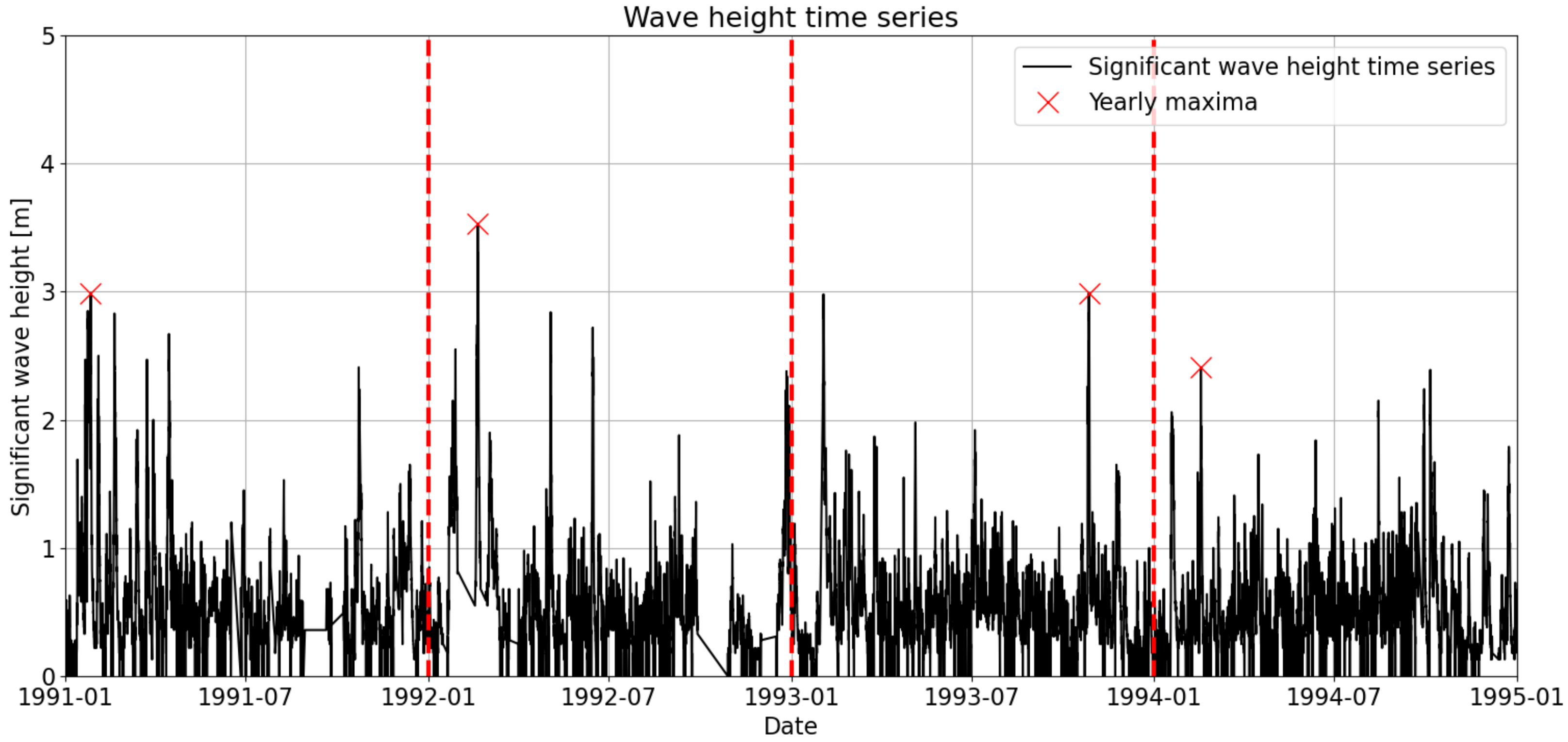
RESULTS SLIDE

Sampling extremes: Block Maxima

1. **Block Maxima**

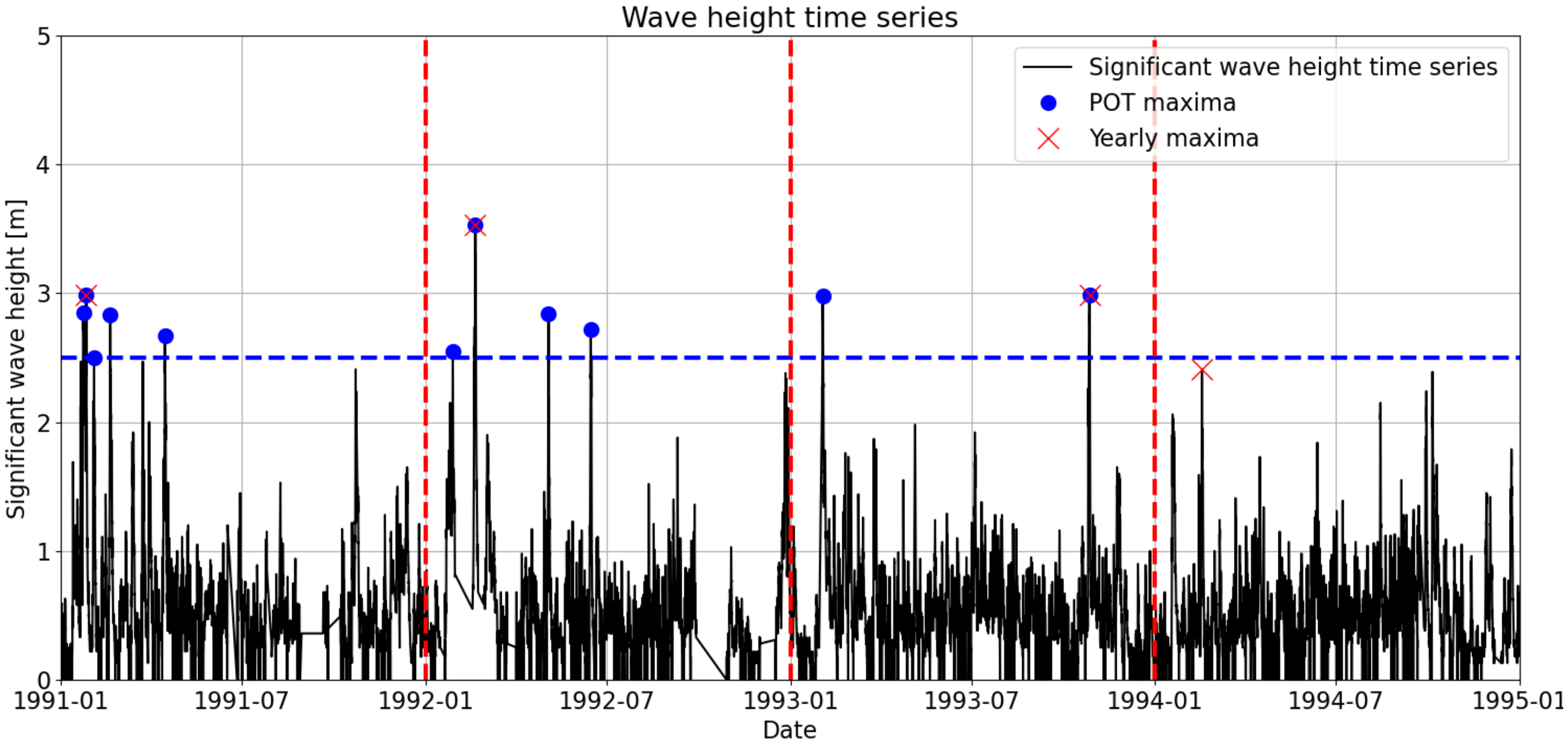


Sampling extremes: Block Maxima



1. **Block Maxima** (typically block=1year)
   - Maximum value within the block
   - Number of selected events=number of blocks recorded (e.g.: number of years)
   - Easy to implement

Sampling extremes: Peak Over Threshold (POT)



2. Peak Over Threshold (POT)

- Usually, higher number of extremes identified
- Additional parameters:
  - Threshold (*th*)
  - Declustering time (*dl*)

And what about the distributions?

19/20 Join at: **vevox.app** ID: **125-461-830** Question slide

Choose the right pairs of sampling technique with distribution function.

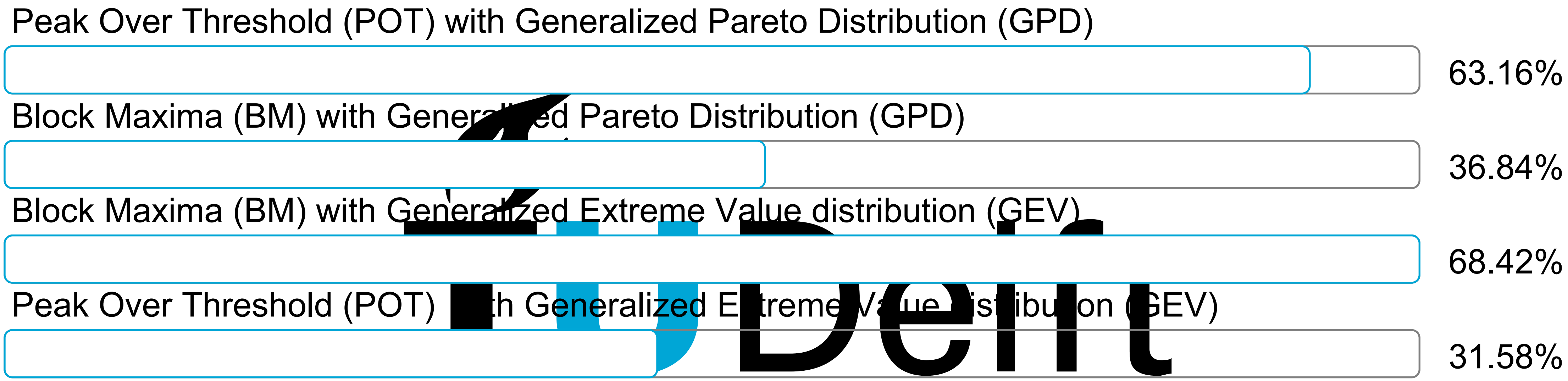
Peak Over Threshold (POT) with Generalized Pareto Distribution (GPD) 0%

Block Maxima (BM) with Generalized Pareto Distribution (GPD) 0%

Block Maxima (BM) with Generalized Extreme Value distribution (GEV) 0%

Peak Over Threshold (POT) with Generalized Extreme Value distribution (GEV) 0%

Choose the right pairs of sampling technique with distribution function.



RESULTS SLIDE

Block Maxima and Generalized Extreme Value Distribution

We are interested in modelling the maximum of the sequence $X = X_1, \dots, X_n$ of *iid* random variables, $M_n = \max(X_1, \dots, X_n)$, where *n* is the number of observations in a given block.

We can prove that for large *n*, **those maxima tend to the Generalized Extreme Value (GEV) family of distributions, regardless the distribution of *X*.**

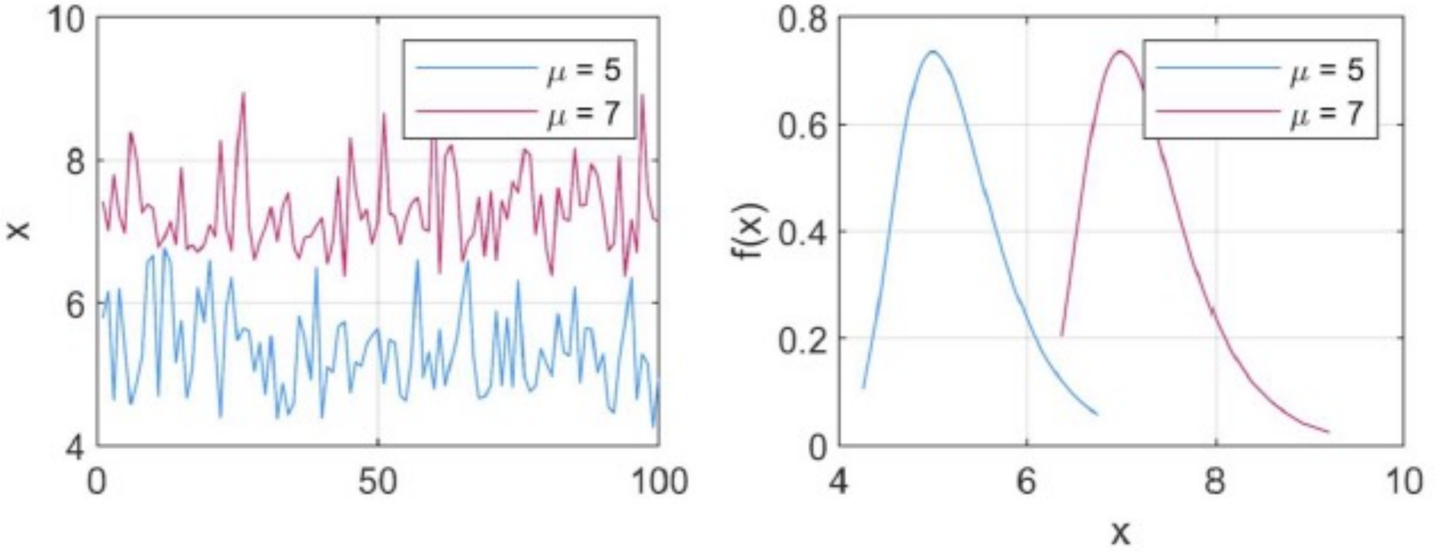
$$P[M_n \leq x] \rightarrow G(x)$$

Block Maxima and Generalized Extreme Value Distribution

Generalized Extreme Value is defined as

$$G(x) = exp{-}[1 + \xi \frac{x-\mu}{\sigma}]^{-1/\xi} \qquad (1 + \xi \frac{x-\mu}{\sigma}) > 0$$

With parameters location ($-\infty < \mu < \infty$), scale ($\sigma > 0$) and shape ($-\infty < \xi < \infty$).



Location parameter (μ)

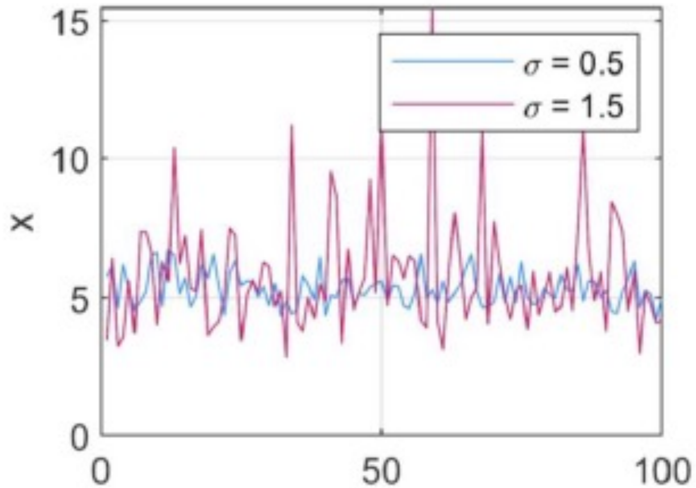
Higher μ, right displacement of the distribution, higher values.

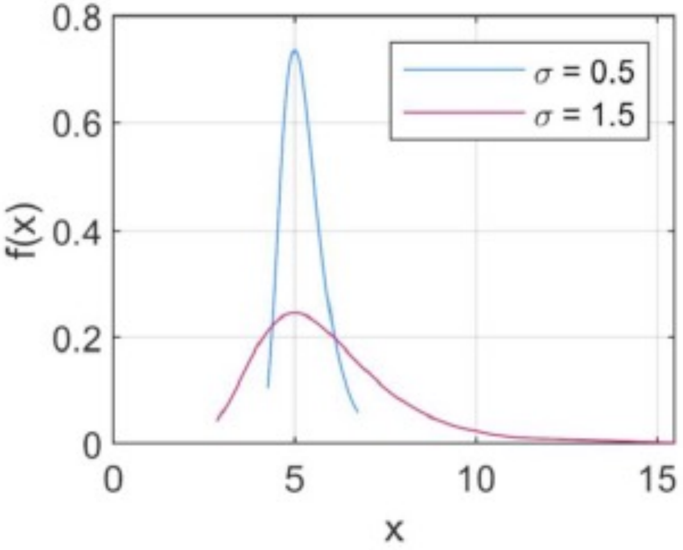
Block Maxima and Generalized Extreme Value Distribution

Generalized Extreme Value is defined as

$$G(x) = exp{-}[1+\xi\tfrac{x-\mu}{\sigma}]^{-1/\xi} \qquad (1+\xi\tfrac{x-\mu}{\sigma}) > 0$$

With parameters location ($-\infty < \mu < \infty$), scale ($\sigma > 0$) and shape ($-\infty < \xi < \infty$).





Scale parameter (σ)

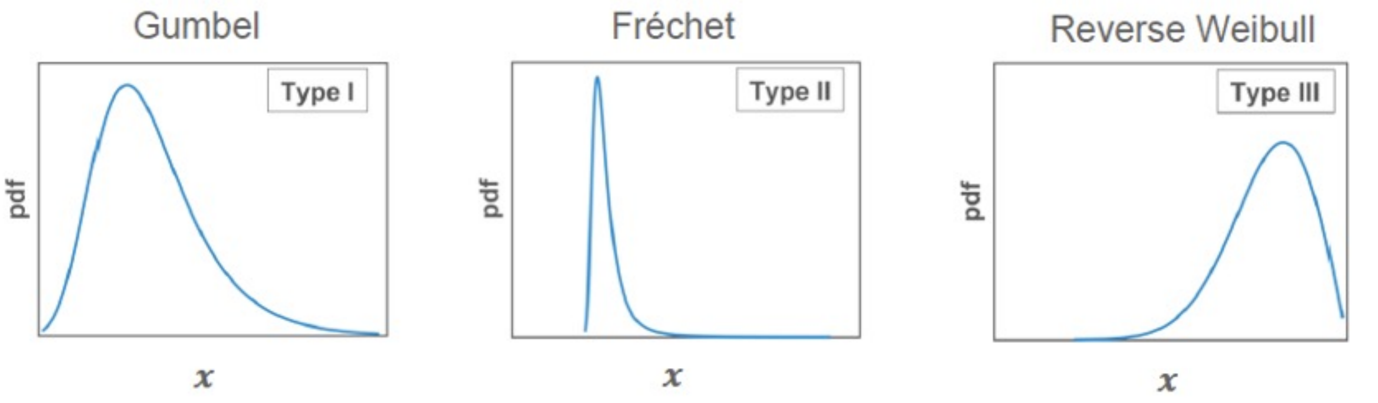
Higher $\boldsymbol{\sigma}$, wider distribution.

Block Maxima and Generalized Extreme Value Distribution

Generalized Extreme Value is defined as

$$G(x) = exp{-}[1 + \xi \frac{x-\mu}{\sigma}]^{-1/\xi} \qquad (1 + \xi \frac{x-\mu}{\sigma}) > 0$$

With parameters location ($-\infty < \mu < \infty$), scale ($\sigma > 0$) and shape ($-\infty < \xi < \infty$).



$\xi \to 0$
Exponential decay

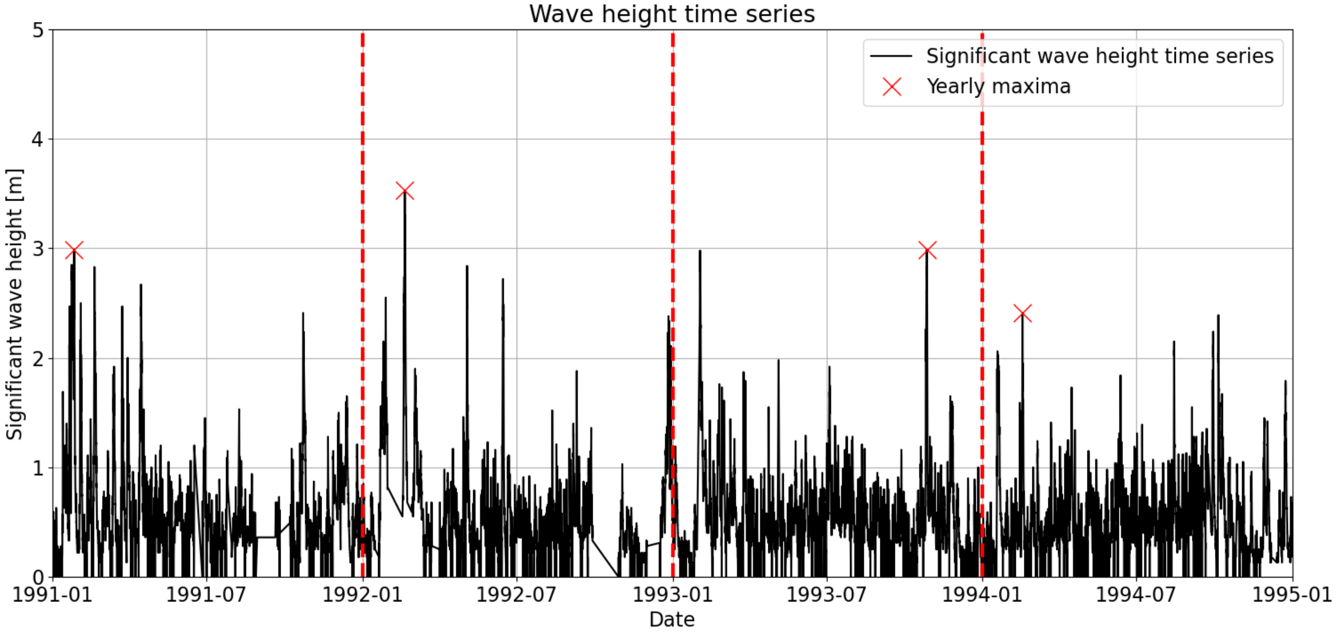
$\xi > 0$
Polynomial decay

$\xi < 0$
Upper bound $(\mu - {}^{\sigma}/_{\xi})$

Shape parameter (ξ)

Determines the tail of the distribution.

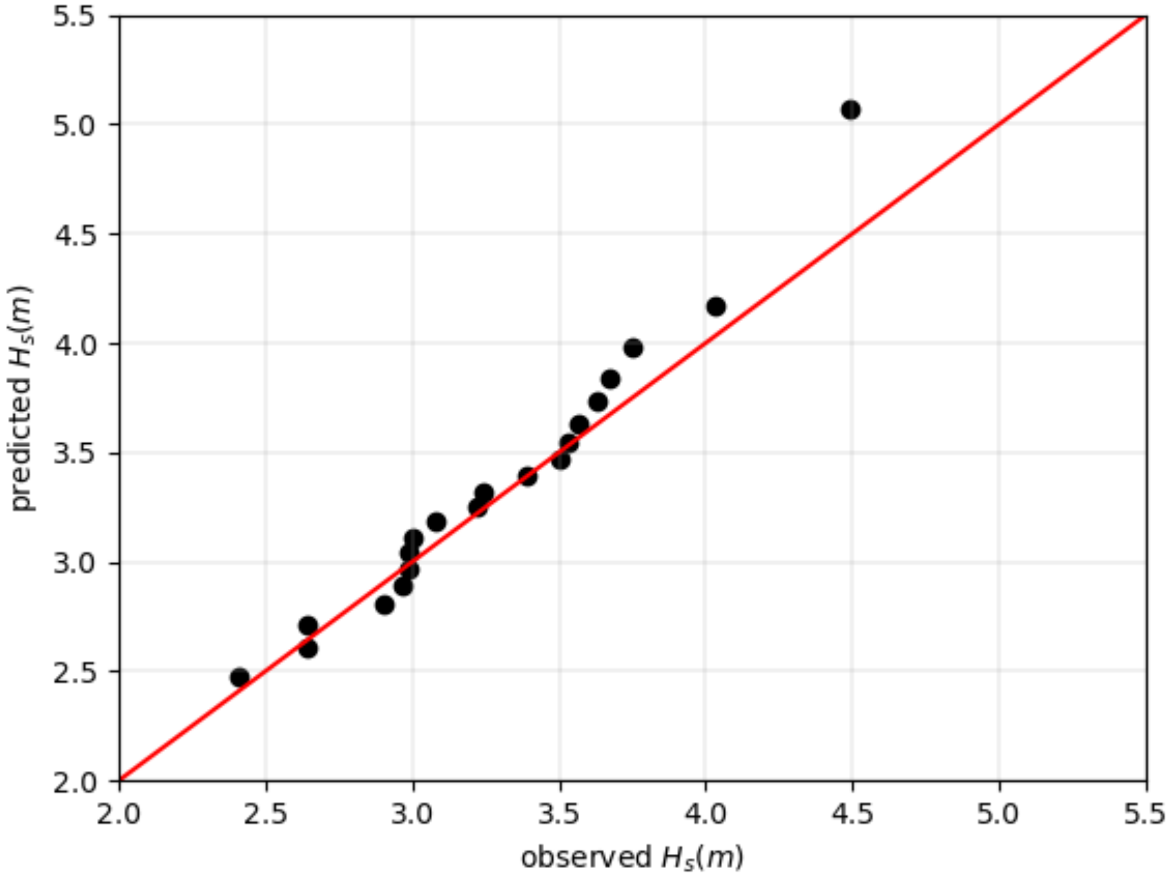
Let’s apply it



- **Load: significant wave height (T_R=90 years)**
- 20 years of hourly measurements → **20 yearly maxima samples**

```
read observations

for each year i:
    obs_max[i] = max(observations in year i)
end

fit GEV(obs_max)

check fit (e.g., QQ-plot or Kolmogorov-Smirnov test)

inverse GEV to determine the design event
```

Let’s apply it

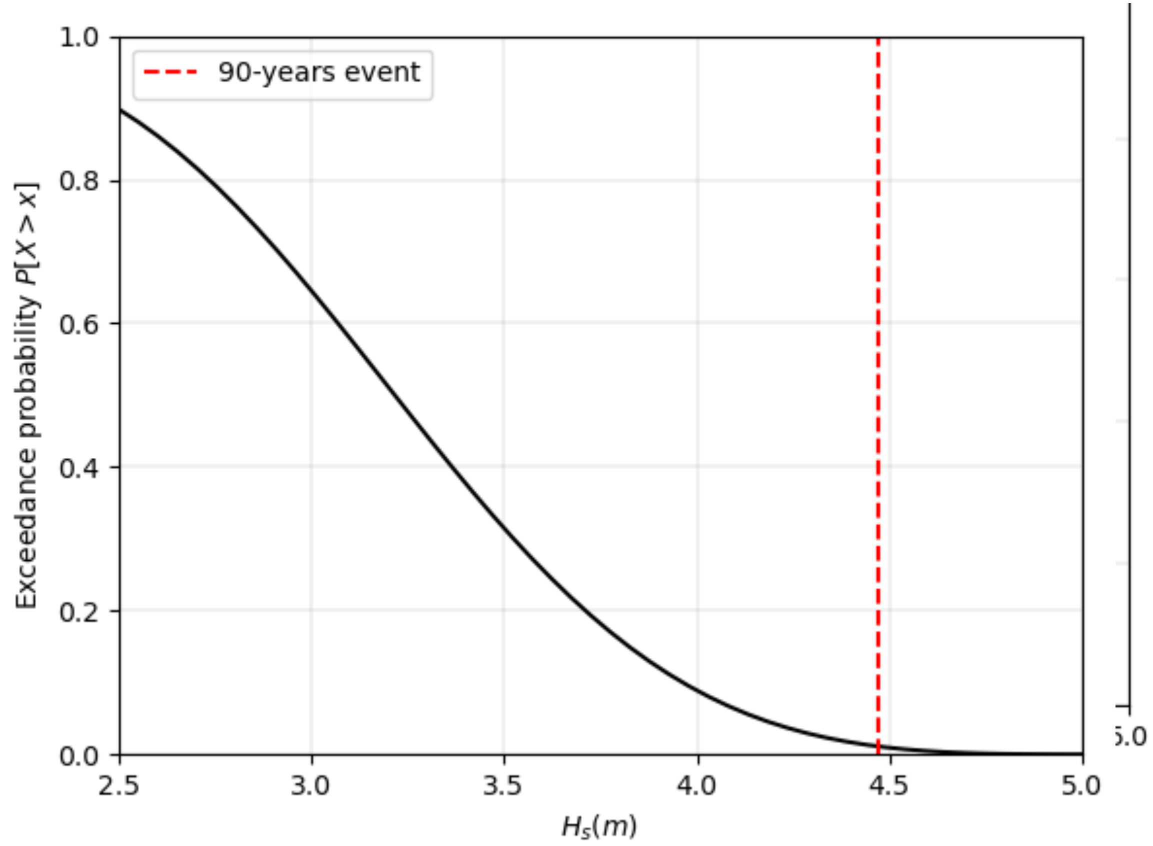


- **Load: significant wave height (T_R=90 years)**
- 20 years of hourly measurements → **20 yearly maxima samples**

```
read observations

for each year i:
obs_max[i] = max(observations in year i)
end

fit GEV(obs_max)

check fit (e.g., QQ-plot or Kolmogorov-Smirnov test)

inverse GEV to determine the design event
```

Let’s apply it

$$z_p = G^{-1}(1-p_{f,y}) = \begin{cases} \mu - \frac{\sigma}{\xi}[1-\{-log(1-p_{f,y})\}^{-\xi}] & for\ \xi \neq 0 \\ \mu - \sigma log\{1-p_{f,y}\} & for\ \xi = 0 \end{cases}$$



- **Load: significant wave height (T_R=90 years)**
- 20 years of hourly measurements → **20 yearly maxima samples**

```
read observations

for each year i:
obs_max[i] = max(observations in year i)
end

fit GEV(obs_max)

check fit (e.g., QQ-plot or Kolmogorov-Smirnov test)

inverse GEV to determine the design event
```

Common mistakes - Let’s talk about the units

Daily maxima of discharges Q is performed on the observations which last for 5 years. We have then 365x5=1,825 extremes. A GEV is fitted.

We want to compute the discharge associated with a **return period of 100 years**.

$$z_p = G^{-1}(1-p_{f,y}) = \begin{cases} \mu - \frac{\sigma}{\xi}[1-\{-log(1-\overset{??}{\boxed{p_{f,y}}})\}^{-\xi}] & for\ \xi \neq 0 \\ \mu - \sigma log\{1-p_{f,y}\} & for\ \xi = 0 \end{cases}$$

Common mistakes - Let’s talk about the units

Daily maxima: ‘units’ of the probabilities in the GEV distribution?

Empirical CDF

Let’s do it slowly!

Length = 5 Days!

x	Sort(x)	Rank	Rank/length + 1
3.2	2	1	1/6 = 0.17
4.5	3.2	2	2/6 = 0.33
3.8	3.8	3	3/6 = 0.5
7.5	4.5	4	4/6 = 0.67
2	7.5	5	5/6 = 0.83

```
>> read observations

>> x = sort observations in ascending order

>> length = the number of observations

>> probability of not exceeding = (range of integer values from 1 to length) / length + 1

>> Plot x versus probability of not exceeding
```

Common mistakes - Let’s talk about the units

Daily maxima: ‘units’ of the probabilities in the GEV distribution $\frac{1}{days}$

Return period: 100 years

$$T_R = \frac{1}{\mathrm{p_{f,y}}} \rightarrow \mathrm{p_{f,y}} = \frac{1}{T_R} = \frac{1}{100\ years}$$

$$T_R = \frac{1}{\mathrm{p_{f,y}}} \rightarrow \mathrm{p_{f,y}} = \frac{1}{T_R} = \frac{1}{100\ years}\frac{1\ year}{365\ days} = 2.7 \cdot 10^{-5}\ \text{1/days}$$

$$z_p = G^{-1}(1-p_{f,y}) = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - \{-log(1 - \boxed{p_{f,y})}\}^{-\xi}] & for\ \xi \neq 0 \\ \mu - \sigma log\{1 - p_{f,y}\} & for\ \xi = 0 \end{cases}$$

POT and Generalized Pareto Distribution

The maximum of the sequence $X = X_1, \ldots, X_n$ of *iid* random variables, $M_n = \max(X_1, \ldots, X_n)$, where *n* is the number of observations in a given block, follows **the Generalized Extreme Value (GEV) family of distributions, regardless the distribution of *X*** for large *n*.

$$P[M_n \leq x] \rightarrow G(x)$$

If that is true, **the distribution of the excesses can be approximated by a Generalized Pareto distribution**.

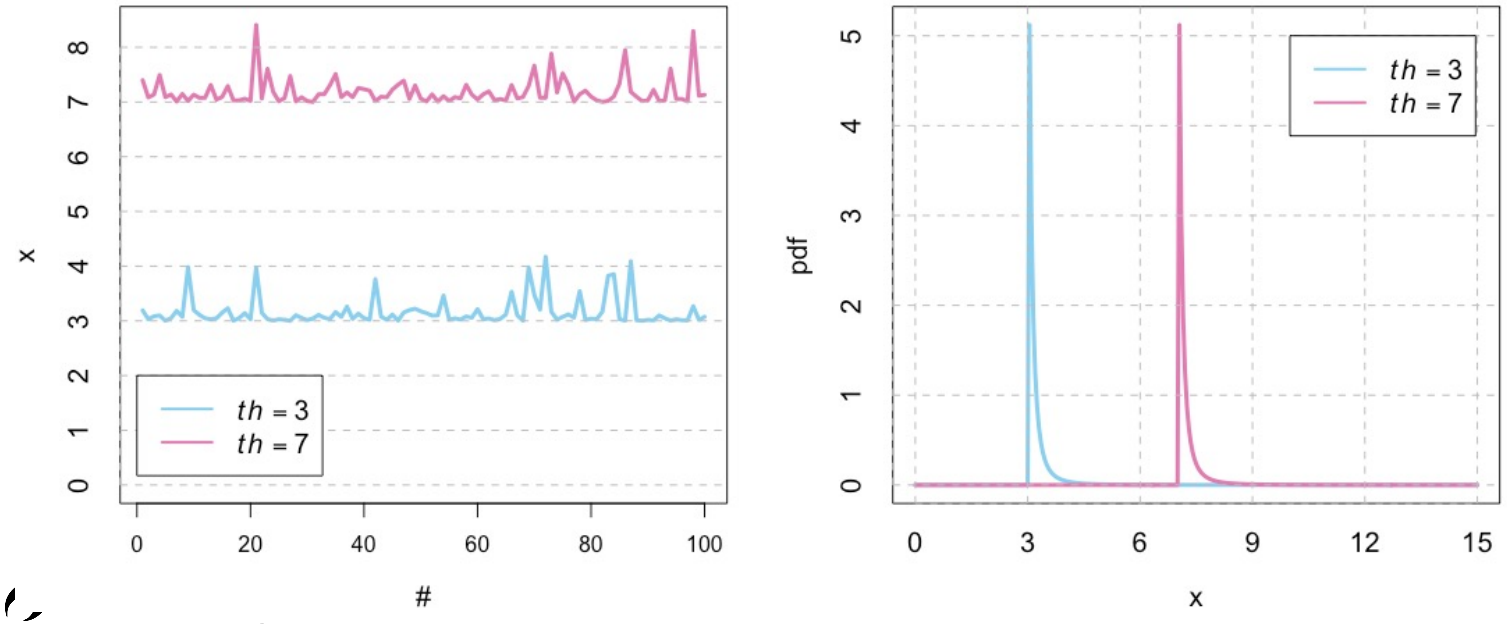
$$F_{th} = P[X - th \leq x | X > th] \rightarrow H(y)$$

where the excesses are defined as *Y*=*X*−*th* for *X*>*th*

POT and Generalized Pareto Distribution

$$P[X < x|X > th] = \begin{cases} 1 - \left(1 + \frac{\xi(x-th)}{\sigma_{th}}\right)^{-1/\xi} & for\ \xi \neq 0 \\ 1 - exp\left(-\frac{x-th}{\sigma_{th}}\right) & for\ \xi = 0 \end{cases}$$

With parameters threshold (*th*>0), pareto's scale ($\sigma_{th} > 0$) and shape ($-\infty < \xi < \infty$).



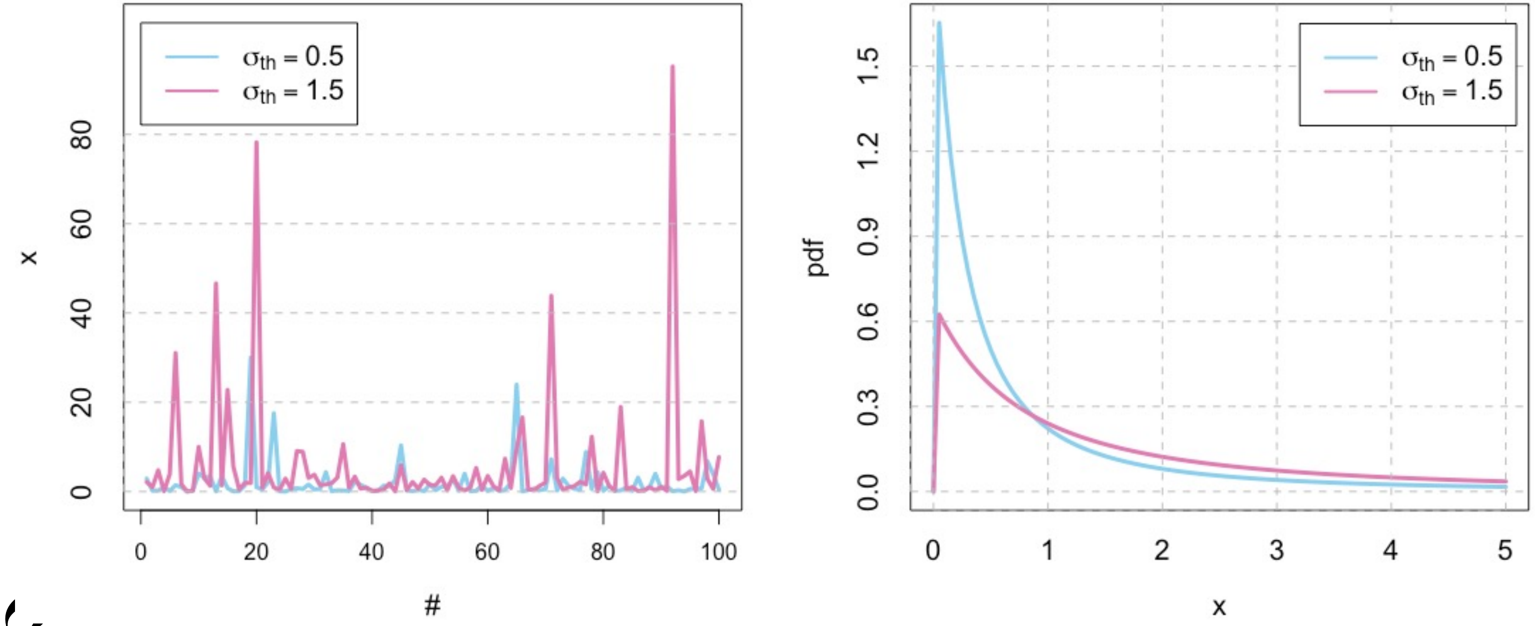
Threshold (*th*)

Acts like a location parameter.

POT and Generalized Pareto Distribution

$$P[X < x|X > th] = \begin{cases} 1 - \left(1 + \frac{\xi(x-th)}{\sigma_{th}}\right)^{-1/\xi} & for\ \xi \neq 0 \\ 1 - exp\left(-\frac{x-th}{\sigma_{th}}\right) & for\ \xi = 0 \end{cases}$$

With parameters threshold (*th*>0), pareto's scale ($\sigma_{th} > 0$) and shape ($-\infty < \xi < \infty$).



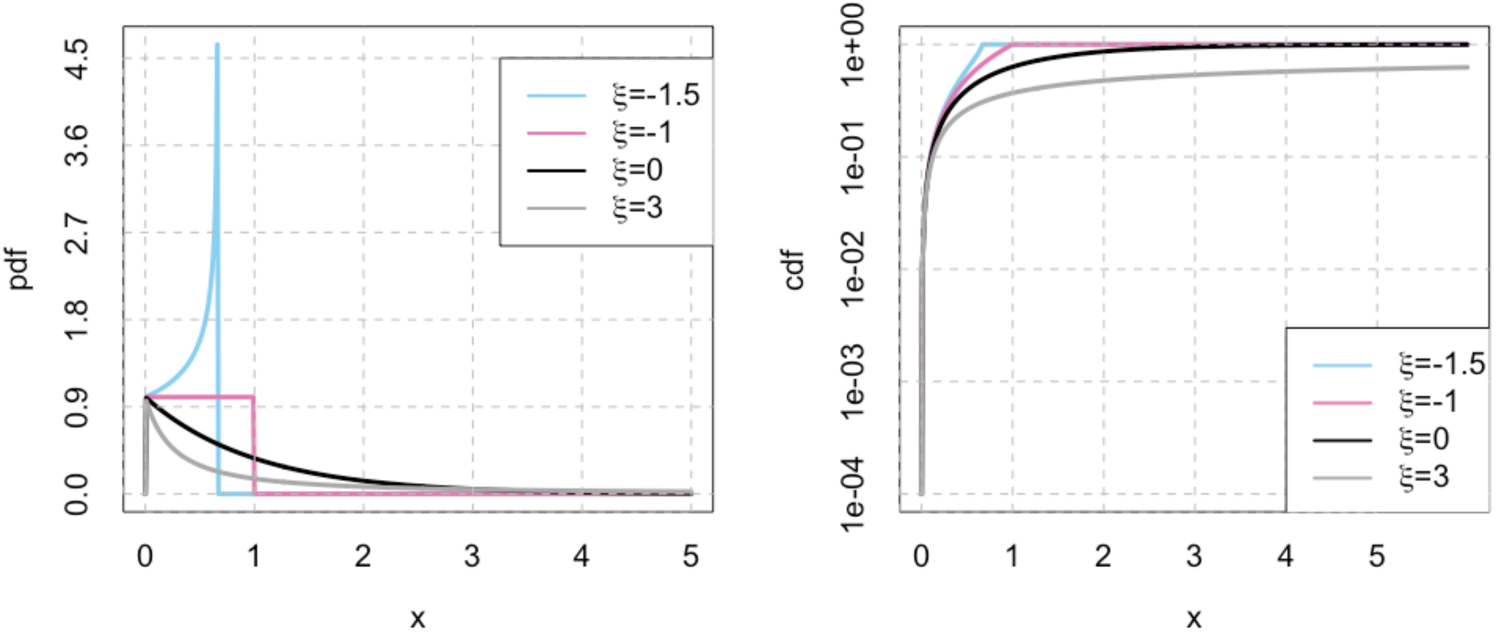
Scale parameter (σ_{th})

Higher σ_{th}, wider distribution.

POT and Generalized Pareto Distribution

$$P[X < x|X > th] = \begin{cases} 1 - \left(1 + \frac{\xi(x-th)}{\sigma_{th}}\right)^{-1/\xi} & for\ \xi \neq 0 \\ 1 - exp\left(-\frac{x-th}{\sigma_{th}}\right) & for\ \xi = 0 \end{cases}$$

With parameters threshold (*th*>0), pareto's scale ($\sigma_{th} > 0$) and shape ($-\infty < \xi < \infty$).



Shape parameter (ξ)

ξ<0: upper bound

ξ>0: heavy tail

ξ=0 & *th* = 0: Exponential

ξ=-1: Uniform

Let’s talk about the units again…

POT of discharges Q is performed on the observations which last for 5 years. A GPD is fitted to the observations.

We want to compute the discharge associated with a **return period of 100 years**.

Let’s talk about the units again…

POT: units of the probabilities in the GPD?

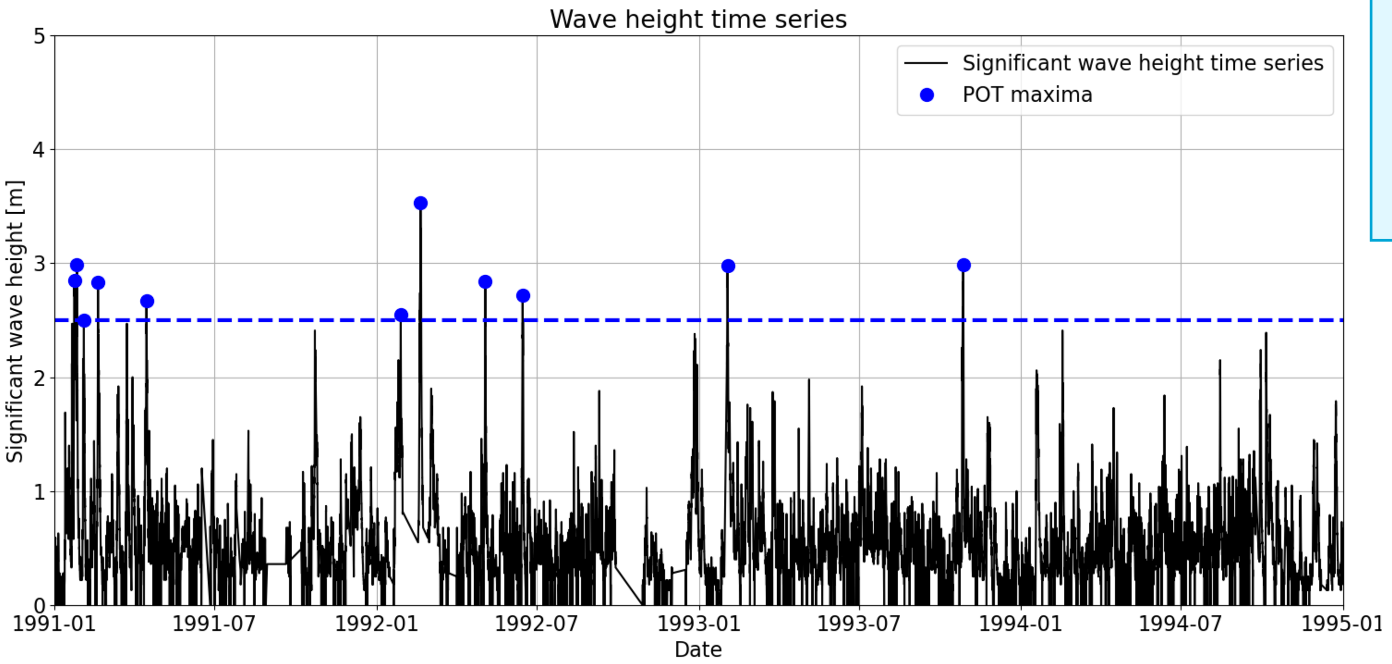
Event-wise probabilities: **not a fixed number in a time block**

We use the average number of exceedances per year

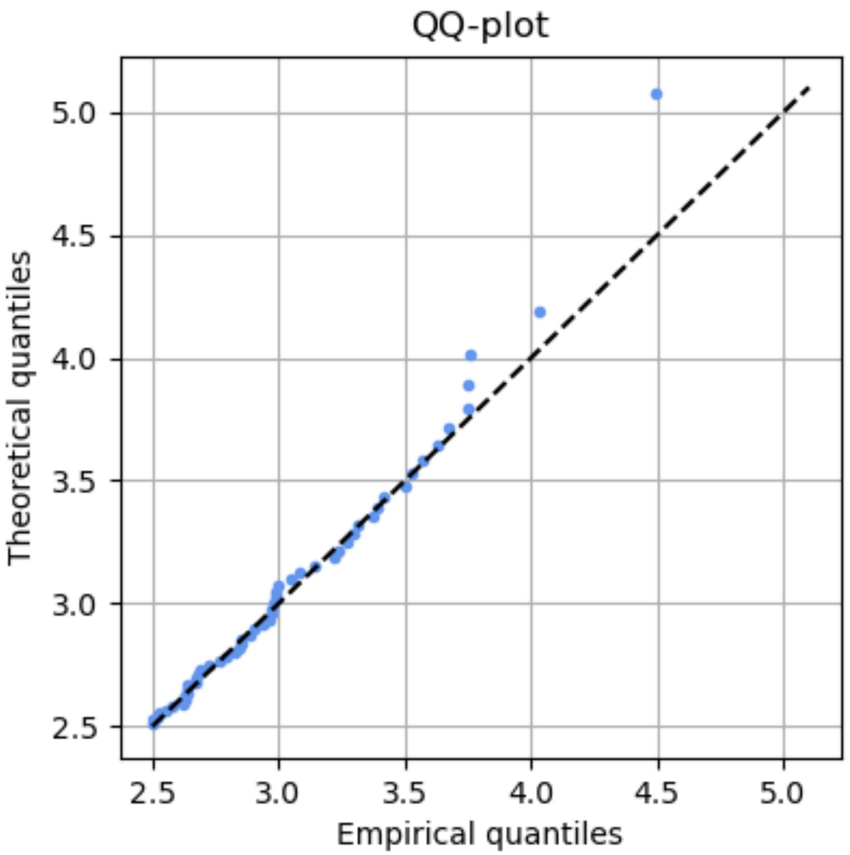
Let's apply it

- **Load: significant wave height (T_R=90 years)**

```
read observations

th = 2.5
dl = 48 #in hours
excesses = find_peaks(observations, threshold = th, distance = dl) – th

fit GPD(excesses)

check fit (e.g., QQ-plot or Kolmogorov-Smirnov test)

determine lambda

inverse GPD to determine the design event
```



Let’s apply it



- **Load: significant wave height (T_R=90 years)**

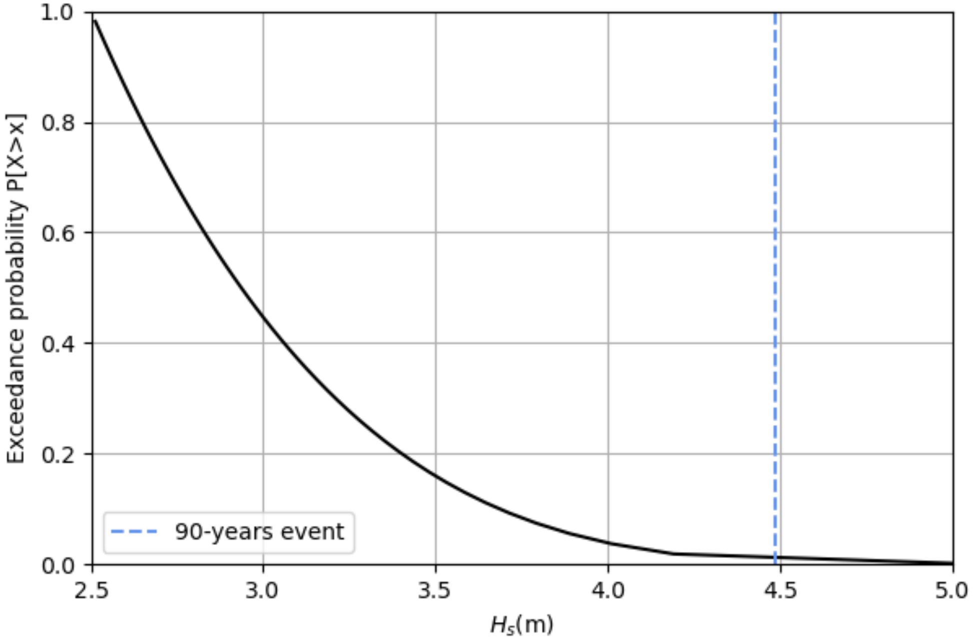
```
read observations

th = 2.5
dl = 48 #in hours
excesses = find_peaks(observations, threshold = th, distance = dl) – th

fit GPD(excesses)

check fit (e.g., QQ-plot or Kolmogorov-Smirnov test)

determine lambda

inverse GPD to determine the design event
```

Let's apply it

$$x_N = \begin{cases} th + \frac{\sigma_{th}}{\xi}[(\lambda N)^{\xi} - 1] & for\ \xi \neq 0 \\ th + \sigma_{th} log(\lambda N) & for\ \xi = 0 \end{cases}$$

T_R=90 years
M = 20 years
n_{th} = 54 events

⇨ $\hat{\lambda} = \frac{54}{20} = 2.7$



- **Load: significant wave height (T_R=90 years)**

```
read observations

th = 2.5
dl = 48 #in hours
excesses = find_peaks(observations, threshold = th, distance = dl) – th

fit GPD(excesses)

check fit (e.g., QQ-plot or Kolmogorov-Smirnov test)

determine lambda

inverse GPD to determine the design event
```

Learning objectives

✓ 1. Identify what is an **extreme value** and apply it within the engineering context

✓ 2. Interpret and apply the concept of **return period and design life**

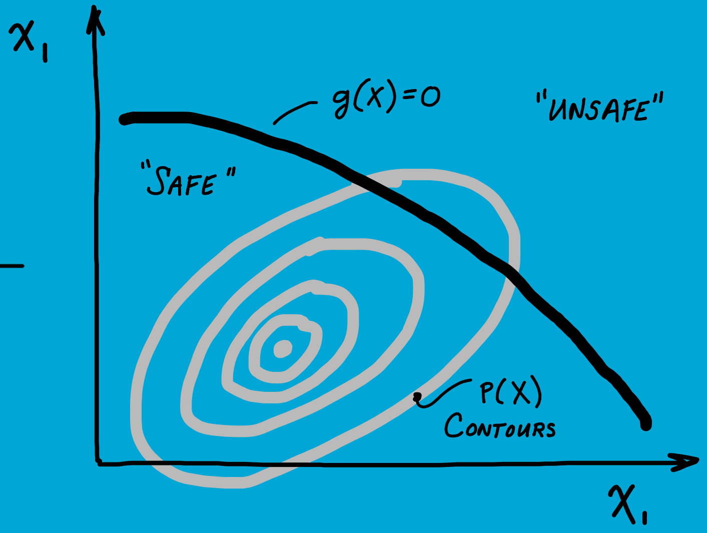
✓ 3. Apply **extreme value analysis** to datasets

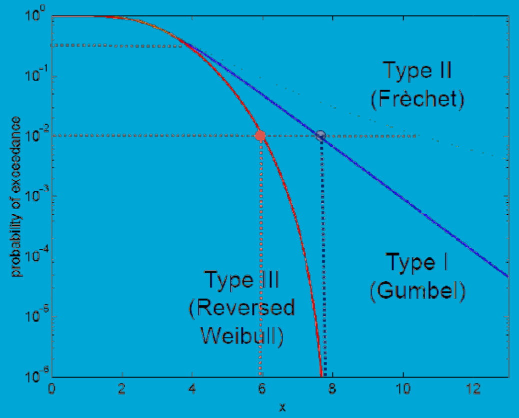
4. Apply techniques to **support the threshold selection**

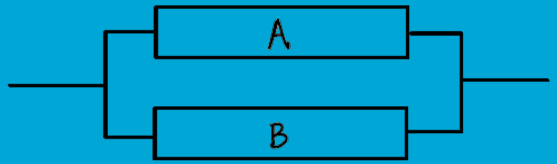
CIEM42X0 Probabilistic Design

Hydraulic and Offshore Structures (HOS) Track

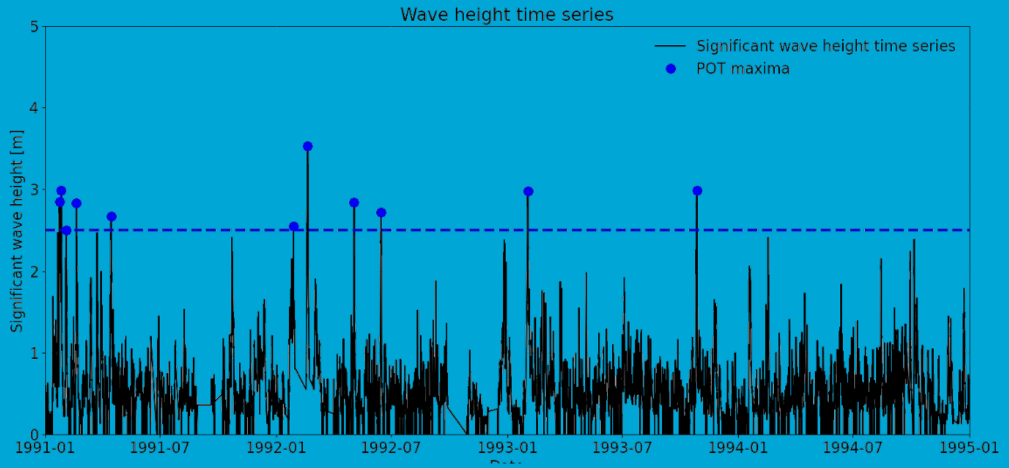
Civil Engineering MSc Program







EVA: Threshold and declustering time selection.

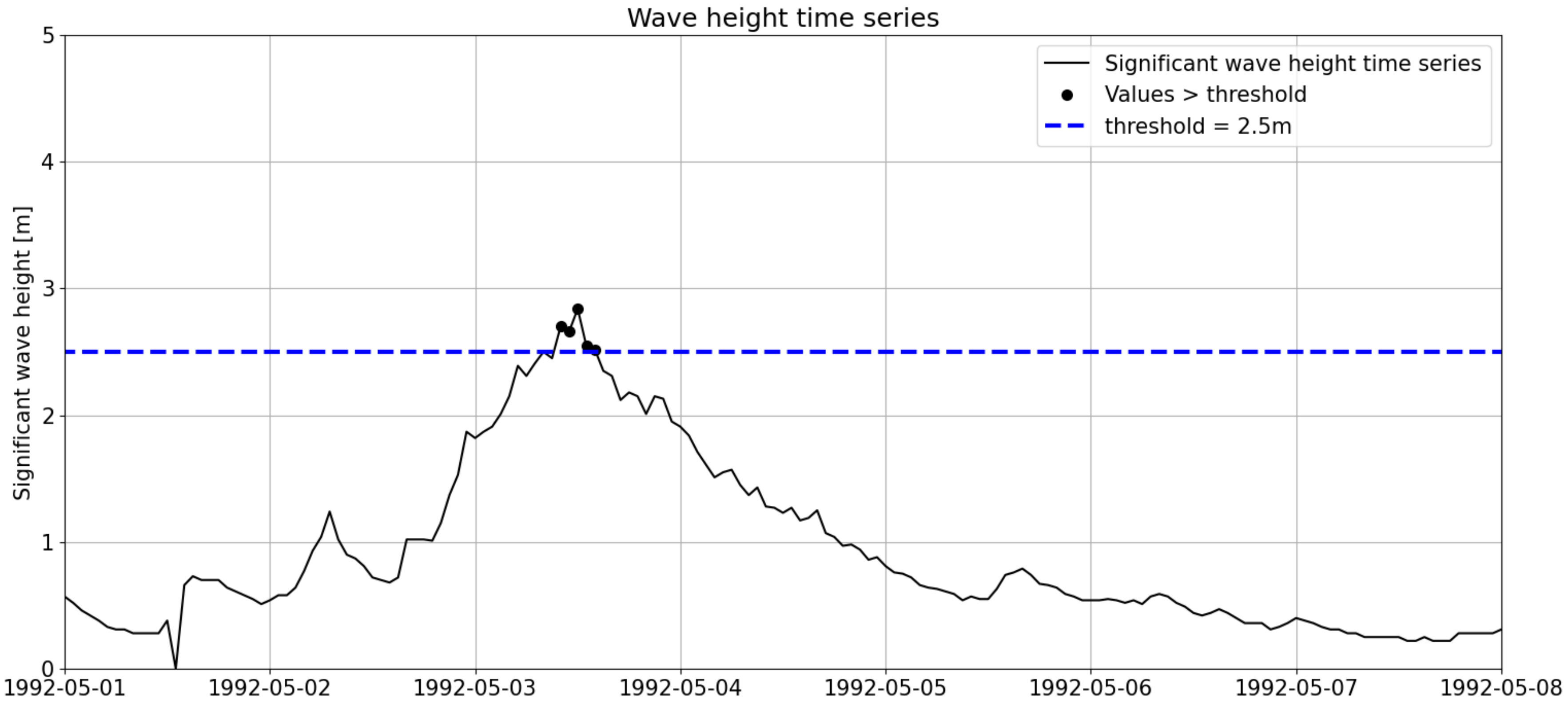


Patricia Mares Nasarre

TUDelft

Choosing POT parameters

Basic assumption of EVA: extremes are *iid* ⟹ *th* and *dl* should be chosen so the identified extreme events are independent.

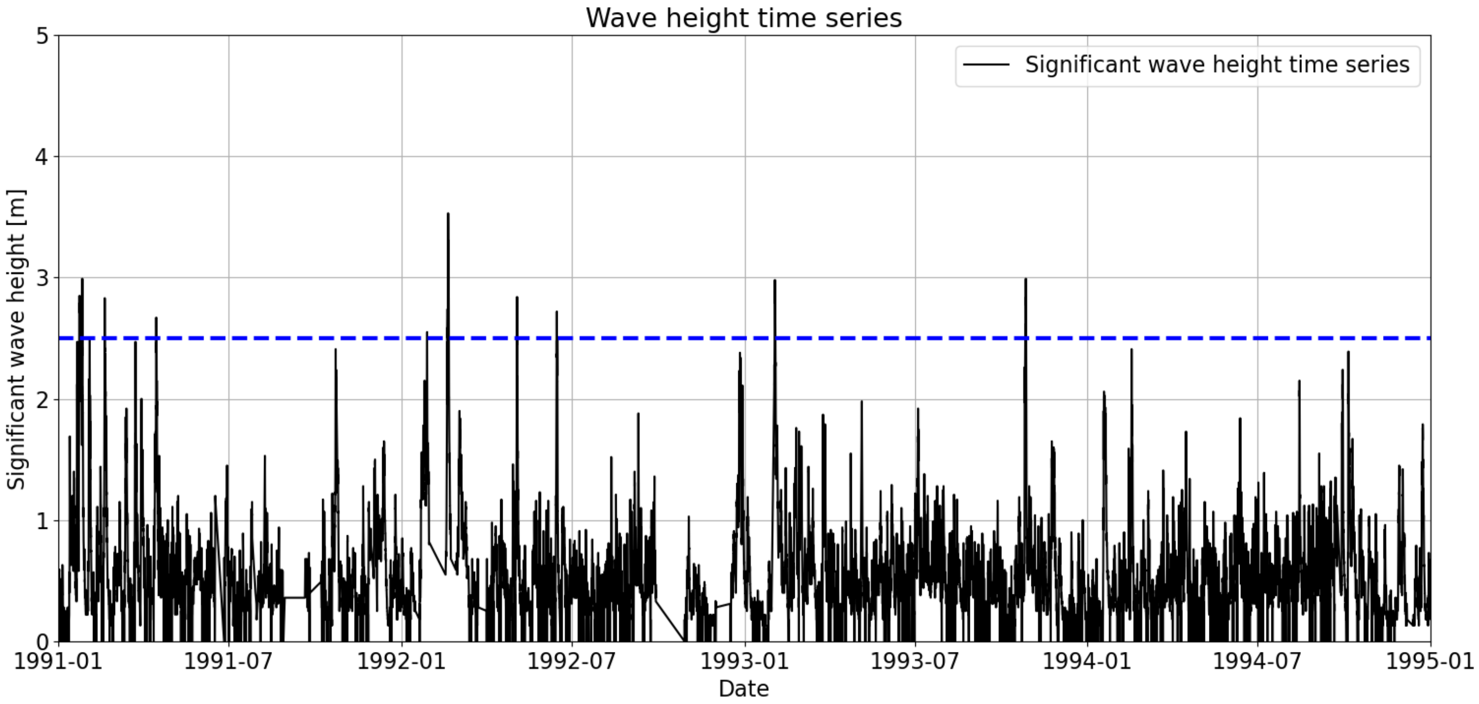


Extremes cluster in time!

If *dl* is big enough, we ensure that extremes do not belong to the same storm.

dl → *th,* physical phenomena (local conditions)

POT and Poisson

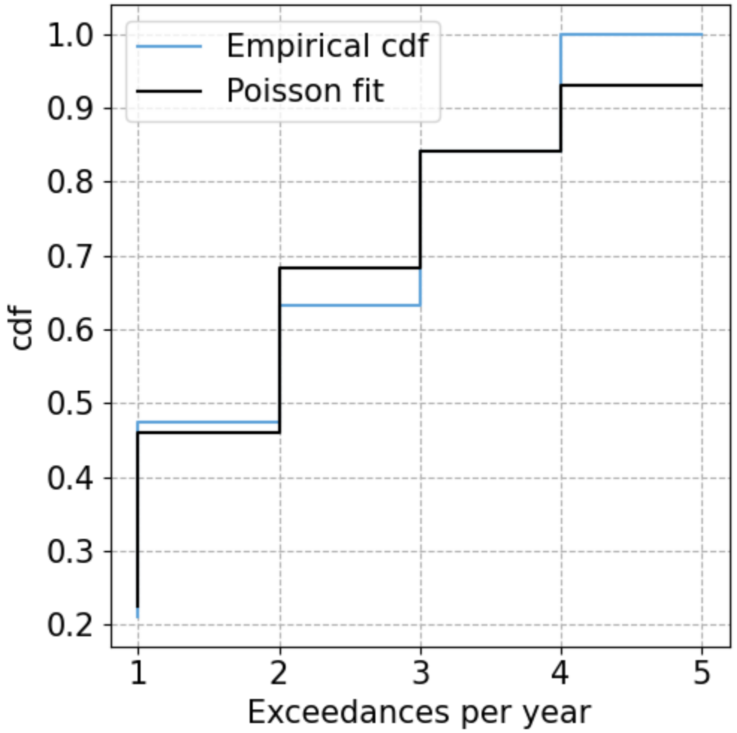


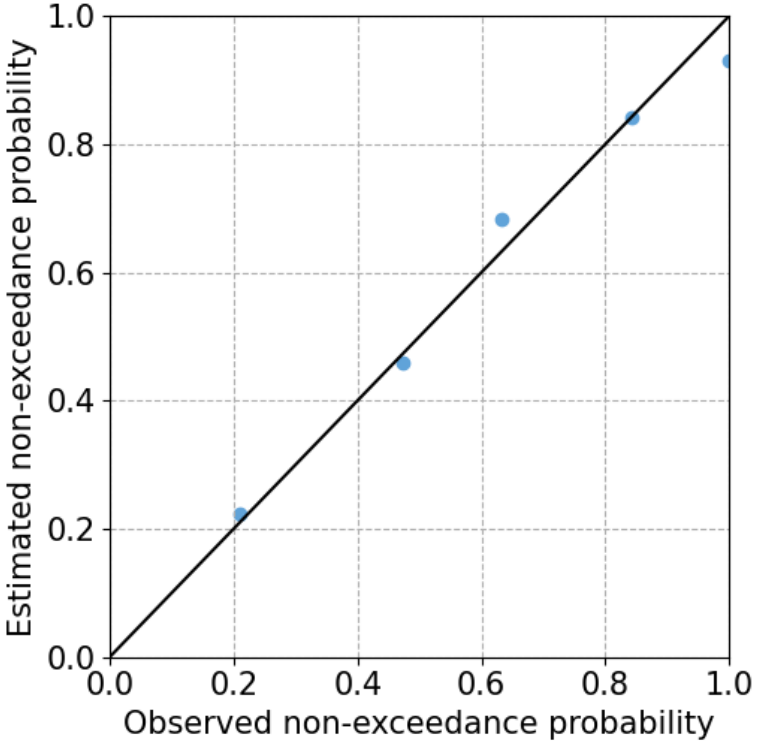
- Each hour is a trial ($n \rightarrow \infty$)
- Over or below the threshold?
- p_{above} is very small (tail of the distribution)
- Block = 1 year
- Number of excesses over the threshold ~ Poisson

Almost all the techniques to formally select the threshold and declustering time for POT are based on the assumption that the sampled extremes should follow a Poisson distribution.

Samples: Poisson

If the number of excesses per year follows a Poisson distribution ⇨ Sampled maxima are independent ✓





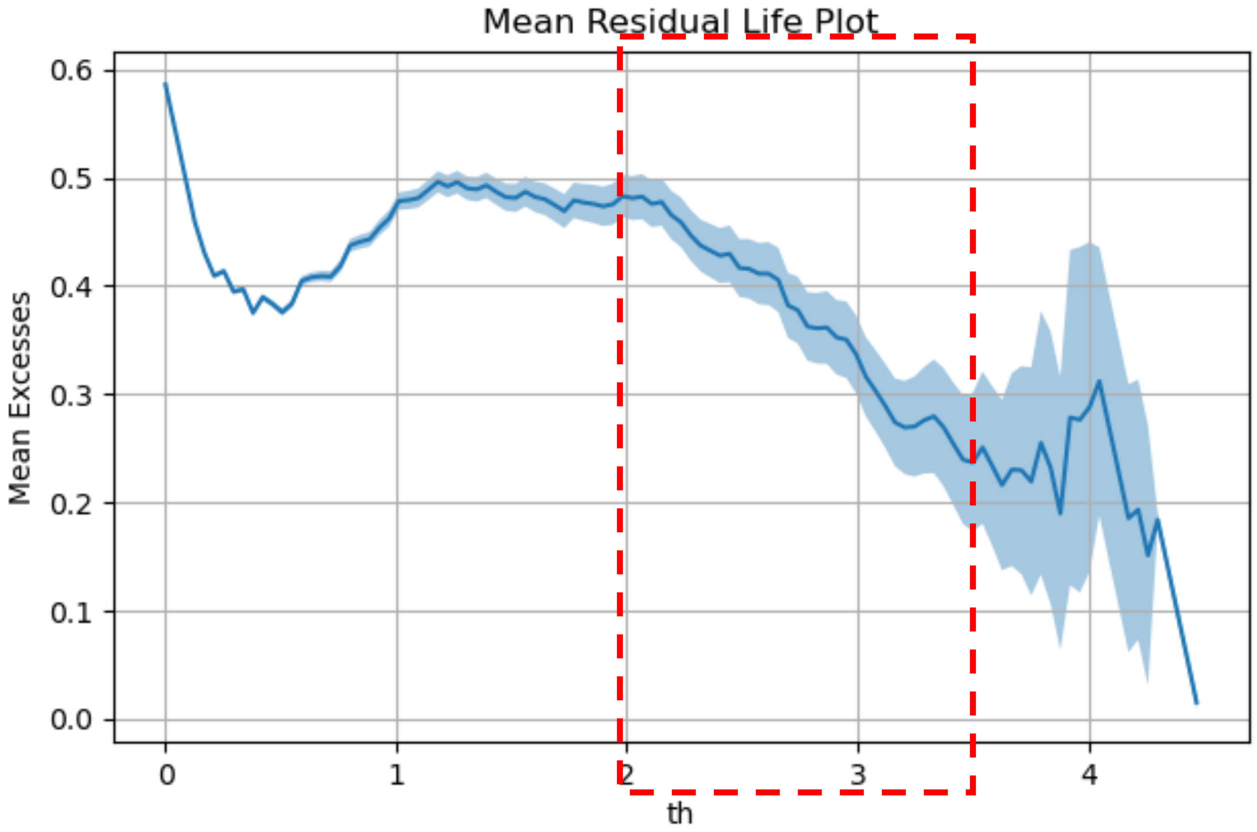
- Compute the number of excesses per year
- Empirical pmf and cdf
- Fit Poisson distribution using Moments

$$E[X] = Var[X] = \lambda$$

- Check the fit
  - Graphically
  - Chi-squared test

Mean Residual Life (MRL) plot

MRL plot presents in the x-axis different values of *th* and, in the y-axis, the mean excess for that value of the *th*. The range of **appropriate threshold** would be that where the **mean excesses follows a linear trend**.

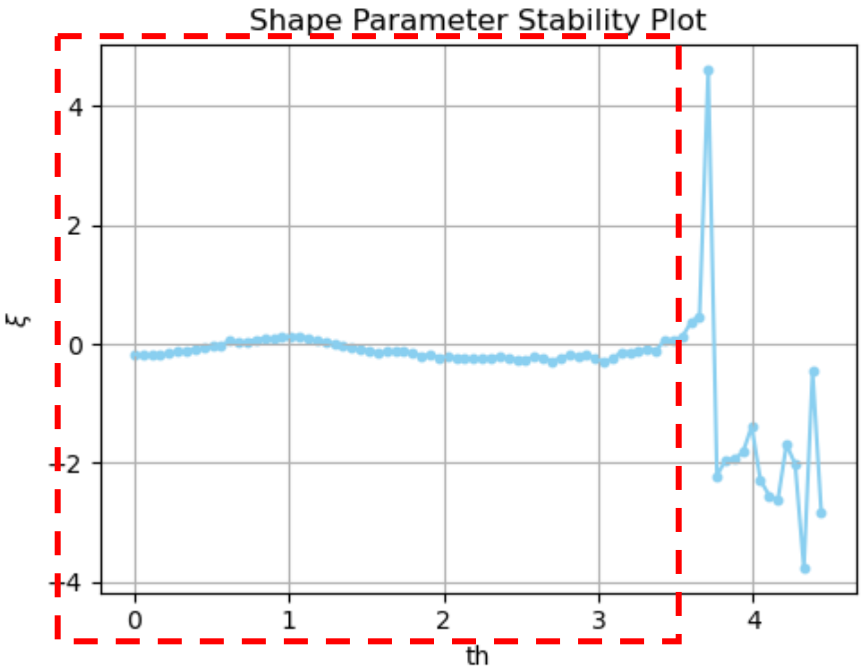


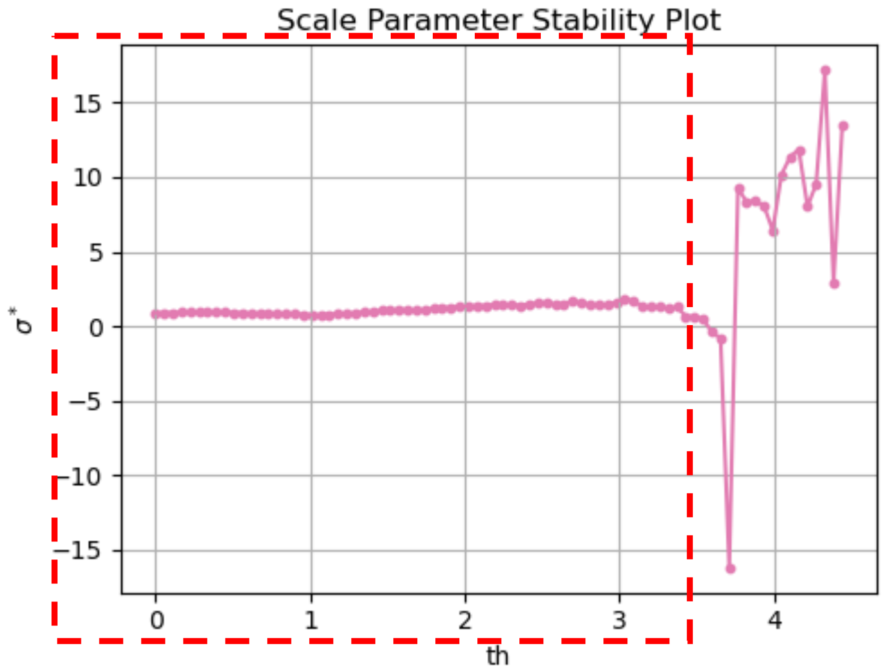
GPD parameter stability plot

GPD distribution is “threshold stable”

If the exceedances over a high threshold (*th0*) a GPD with parameters ξ and σ_{th0}, then for any other threshold (*th>th0)*, the exceedances will also follow a GPD with the same ξ and

$$\sigma_{th} = \sigma_{th0} + \xi(th - th0) \Longrightarrow \sigma^* = \sigma_{th} - \xi th \Longrightarrow \sigma^* = \xi\, th0$$





Dispersion Index (DI)

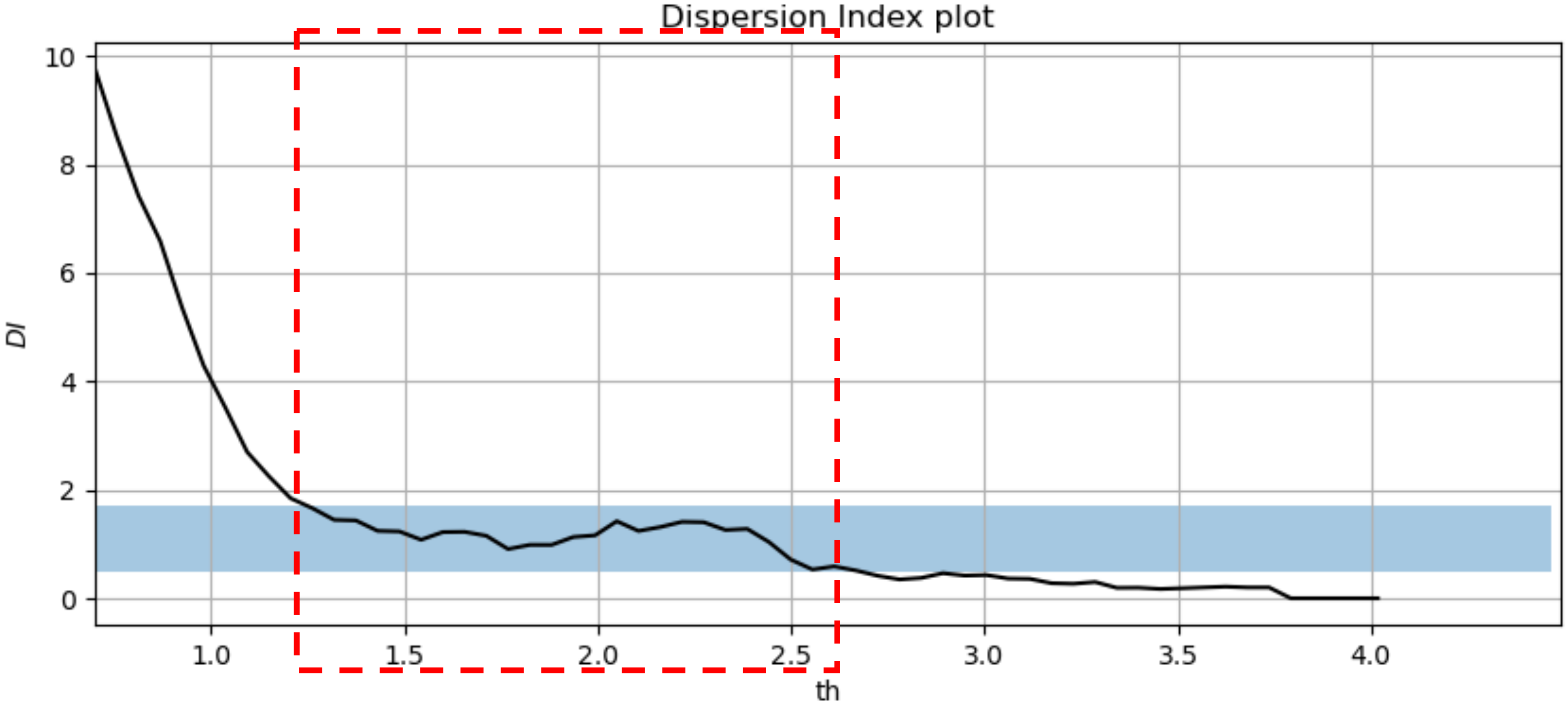
Based on Poisson process

Property of Poisson distribution: $E[X] = Var[X] = \lambda$

Dispersion Index: $DI = \frac{\sigma^2}{\mu} \approx 1$

Confidence interval for DI:

$$\left(\frac{\chi^2_{\alpha/2,M-1}}{(M/1)}, \frac{\chi^2_{1-\alpha/2,M-1}}{(M/1)} \right)$$



Learning objectives

✓ 1. Identify what is an **extreme value** and apply it within the engineering context

✓ 2. Interpret and apply the concept of **return period and design life**

✓ 3. Apply **extreme value analysis** to datasets

✓ 4. Apply techniques to **support the threshold selection**

Any questions?

Patricia Mares Nasarre

p.maresnasarre@tudelft.nl

TUDelft