Modelling, Uncertainty and Data for Engineers (MUDE)

Week 2.7 : Extreme Value Analysis

Patricia Mares Nasarre

TU Delft

Welcome to...

Modelling, Uncertainty, and Data for Engineers

Probability is back!!

WEEK 7!

Today's session – Extreme Value Analysis (EVA)

1. Refresher (mainly week 7) and motivation for EVA
2. Extremes, design conditions and return period
3. Block Maxima & GEV
4. POT & GPD

Bernoulli process
Binomial distribution
Poisson distribution

1. Refresher and motivation

TU Delft

Recap of Q1 – Week 1

Deterministic vs Stochastic

Deterministic models are those which for some given inputs, always provide the same output. For instance, a equation which gives the average concentration of CO_2 in a city as function of the traffic. For a certain value of traffic, the model will always provide the same concentration of CO_2. Therefore, these models that there is no uncertainty. On the contrary, stochastic models are those which embrace the uncertainty. This is stochastic models will produce different outputs for a given input. In fact, the inputs and outputs of stochastic models are probabilistic distributions (you will learn more about this later!), which relate the values of the variable with the probability of observing it.

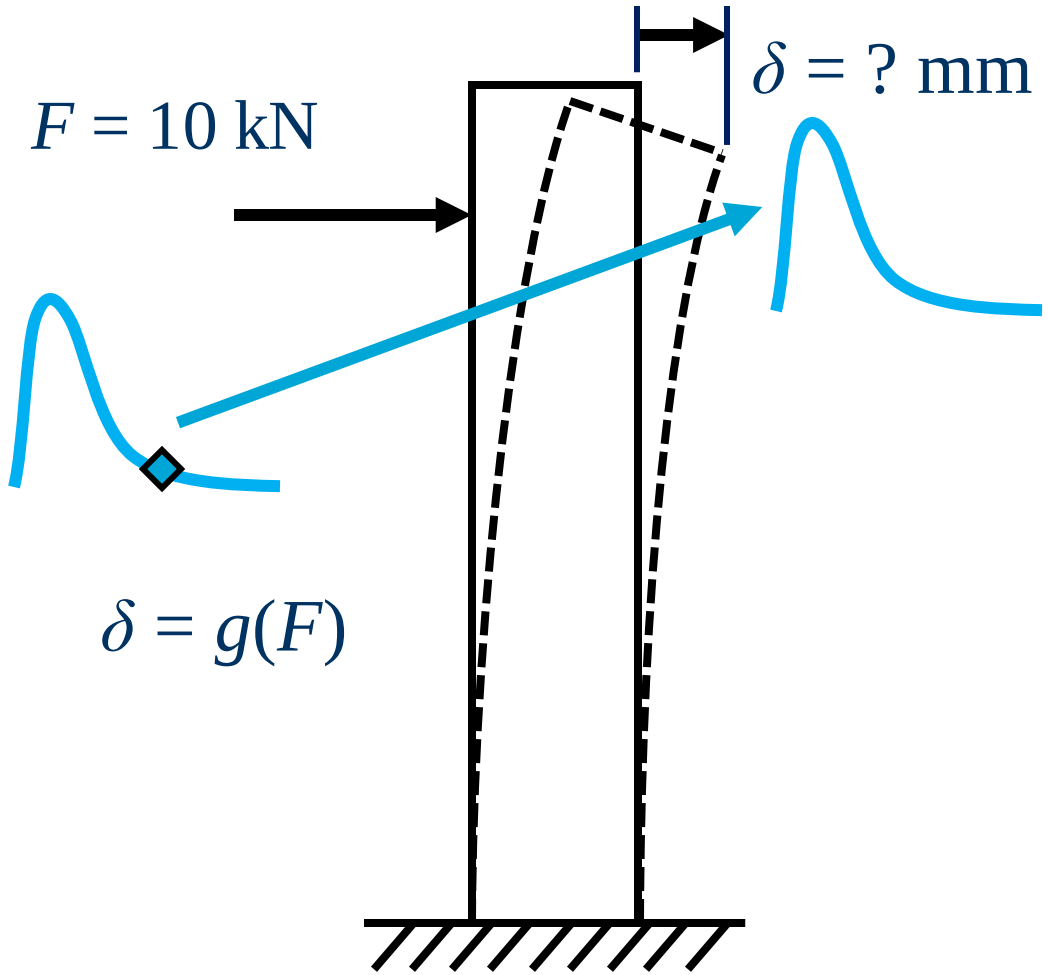
And how do I choose between a deterministic and stochastic model?

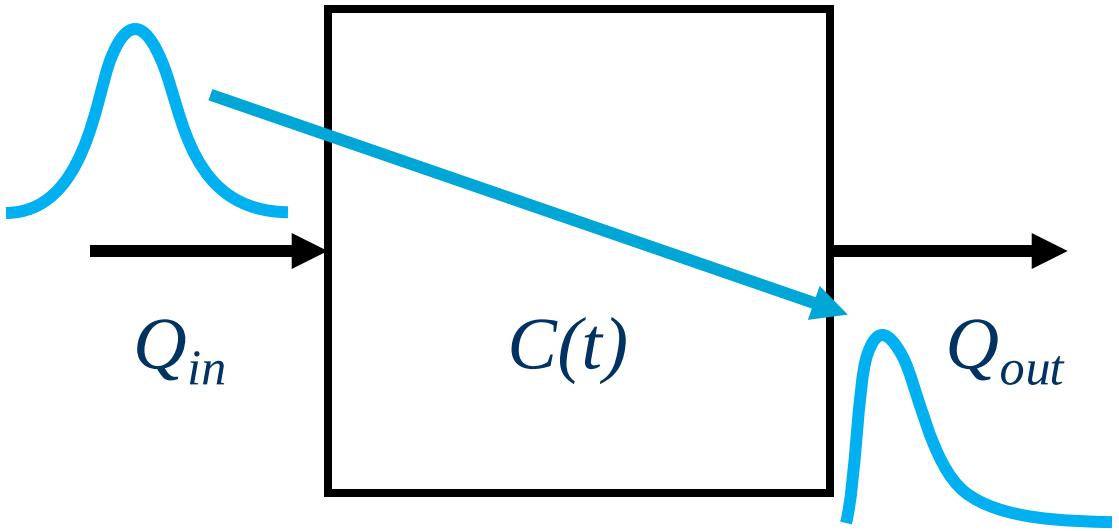
All systems, in reality, are stochastic to our eyes, since we never truly know the actual properties and inputs. However, under certain circumstances, this *stochasticity* can be neglected. Let us take a look to some examples of deterministic and stochastic systems:

Deterministic → If input is 'a', output will always be 'b'

Stochastic → If input is 'a', what is the probability of 'b'

Recap of Q1 – Uncertainty weeks

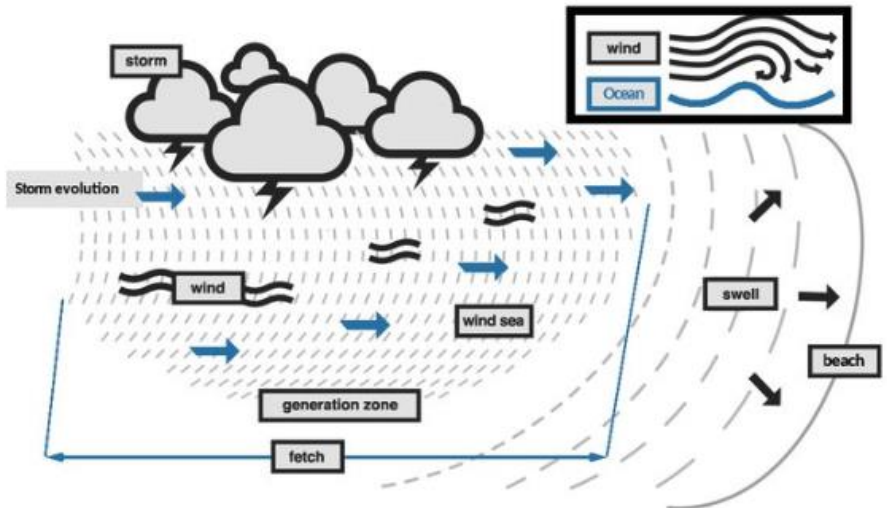




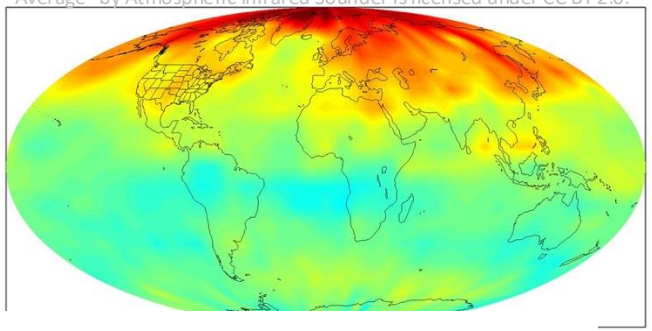
$$Q_{out} = g(Q_{in}, C(t))$$

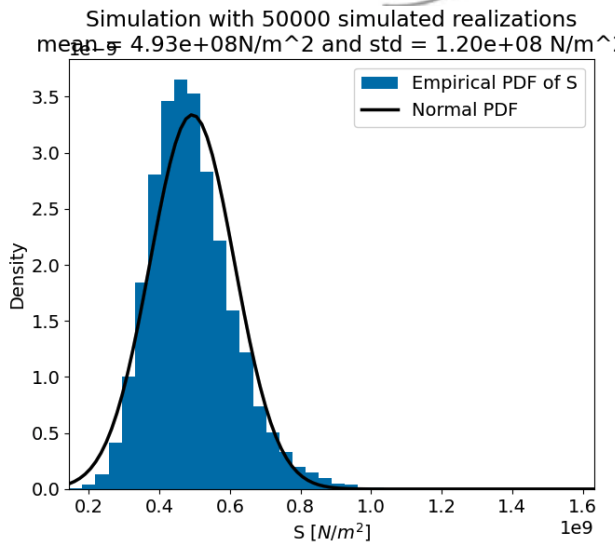
Recap of Q1 – Week 7

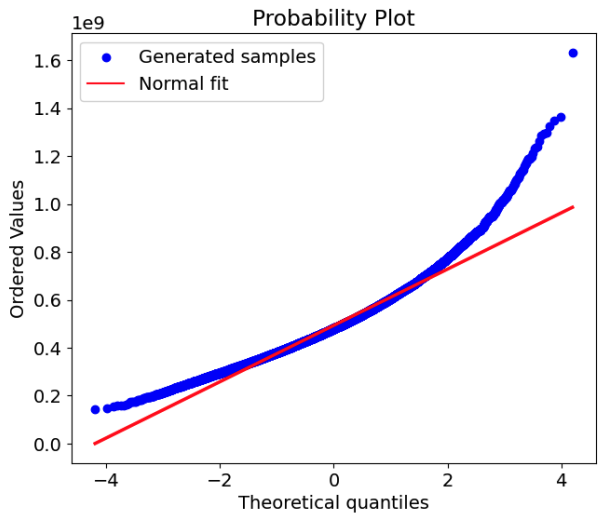
From C.E. Stringari (2020)



"Carbon Dioxide in Earth's Mid-Troposphere, April 2013 Monthly Average" by Atmospheric Infrared Sounder is licensed under CC BY 2.0.







Aleatoric

- intrinsic phenomenon; typically associated with variations that occur in nature

Epistemic

- lack of knowledge; often called model uncertainty

Error

- deficiency in any stage of modelling/simulation not due to lack of knowledge

Variables are NOT necessarily Gaussian-distributed!

Recap of Q1 – PDF

- Continuous random variables

- Distribution function: Mathematical model which relates the values of a random variable and their probability

Join the Vevox session

Go to **vevox.app**

Enter the session ID: **108-740-284**

Or scan the QR code

TU Delft

Which probability functions have you used previously in MUDE?

Which probability functions have you used previously in MUDE?

Recap of Q1 – PDF

- Continuous random variables

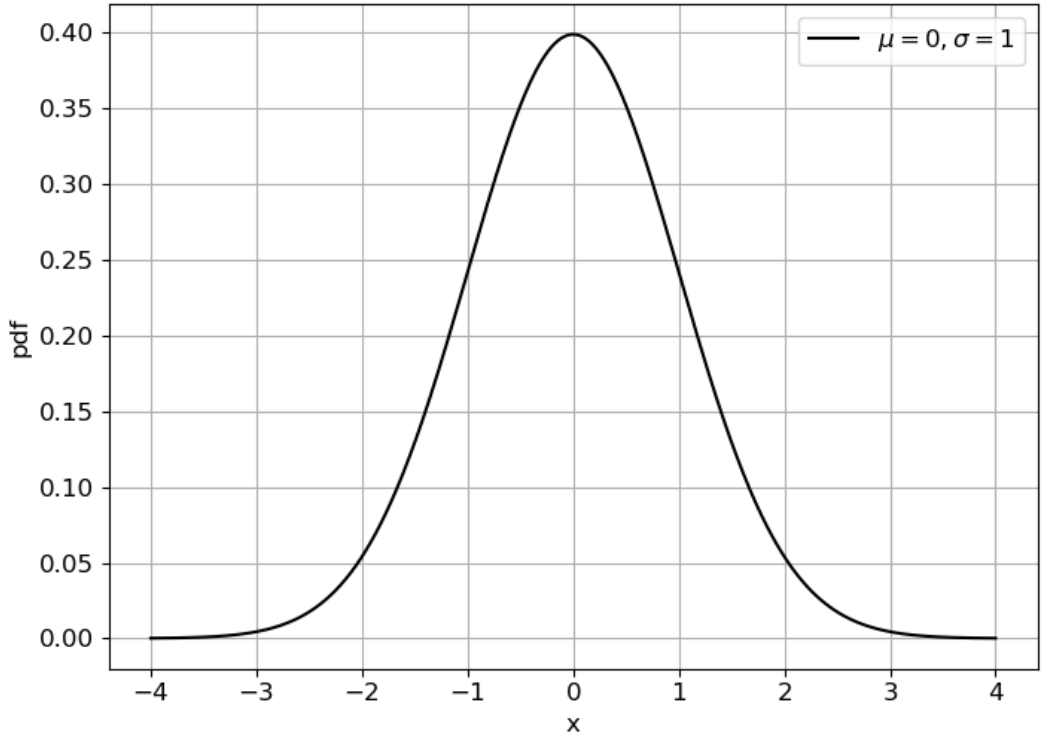
- Distribution function: mathematical model which relates the values of a random variable and their probability

- Probability density function (PDF) $f_X(x)$

$$f_X(x)dx = P(x < X \leq x + dx)$$

$$f_X(x) \geq 0$$

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1$$



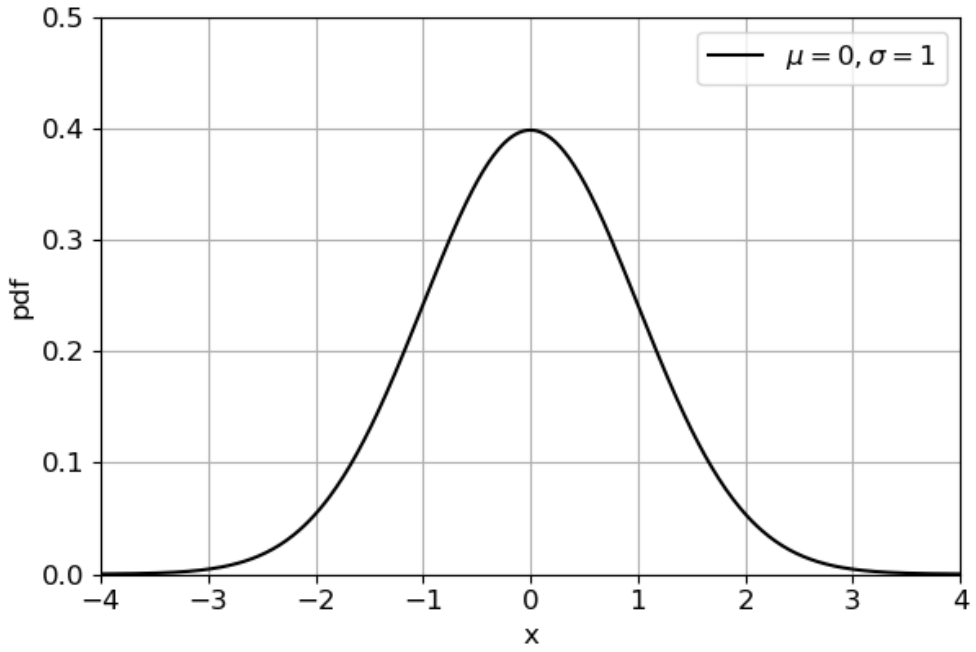
PDF of the Gaussian distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

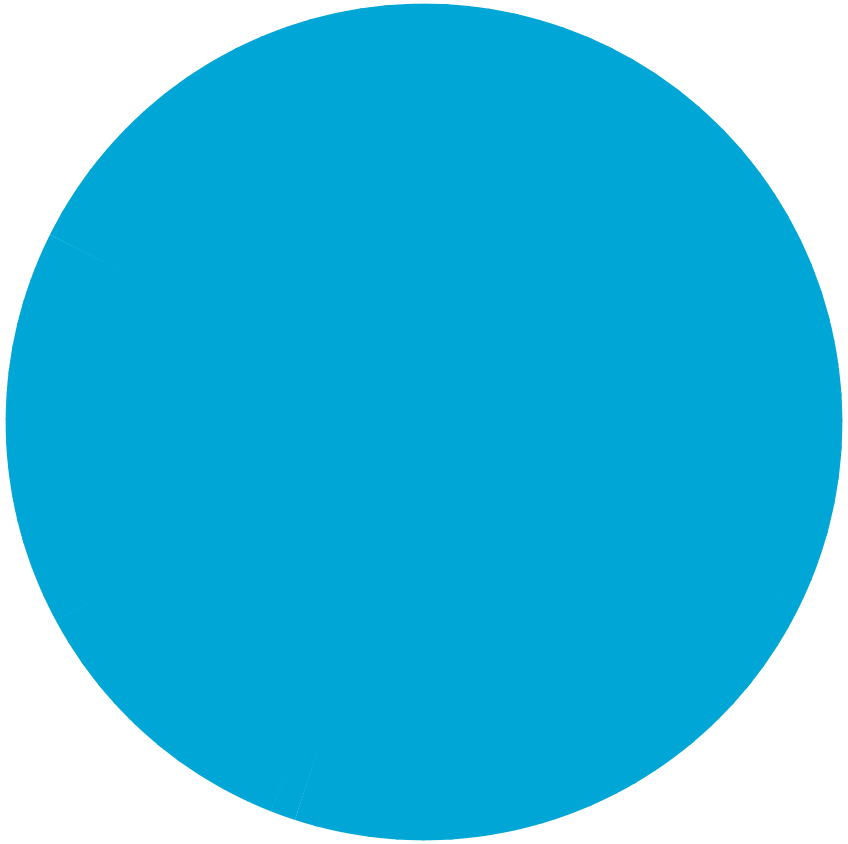
Recap of Q1 – From PDF to CDF

- Probability density function (PDF) $f_X(x)$

- Cumulative distribution function (CDF) $F(x) = \int_{-\infty}^{x} f(x)dx$



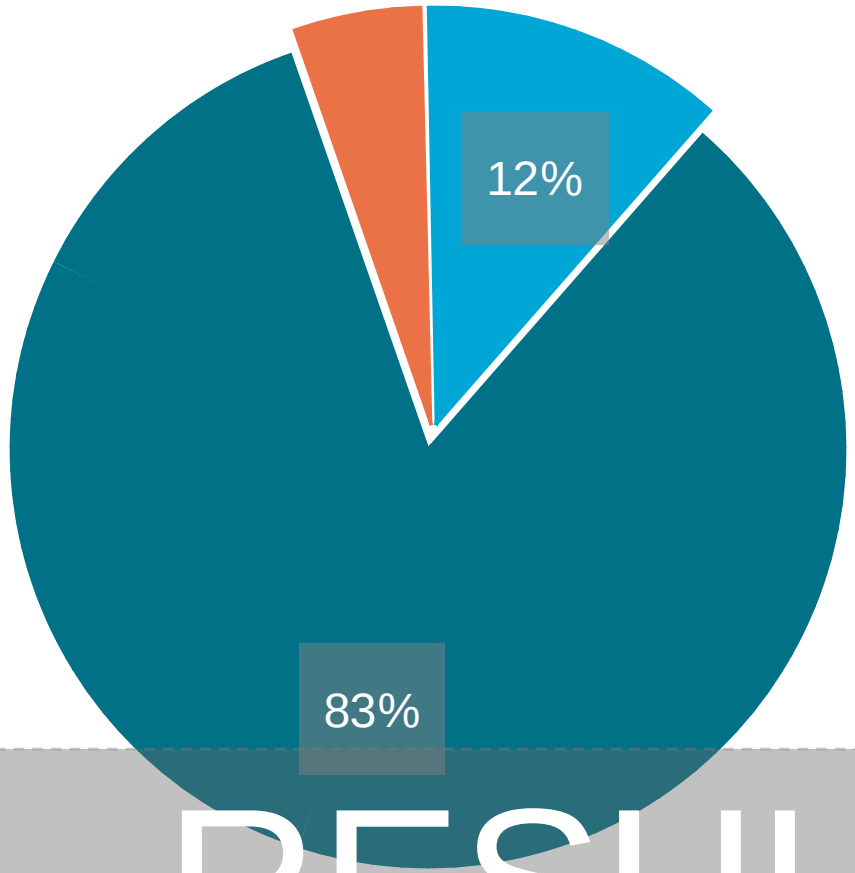
Which probability do I obtain when evaluating the CDF (Cumulative Distribution Function)?



Exceedance probability, P(X>x)	0%
Non-exceedance probability, P(X<x)	0%
Exact probability, P(X=x)	0%

TU Delft

Which probability do I obtain when evaluating the CDF (Cumulative Distribution Function)?



Exceedance probability, P(X>x)	12%
Non-exceedance probability, P(X<x)	83%
Exact probability, P(X=x)	5%

TU Delft RESULTS SLIDE

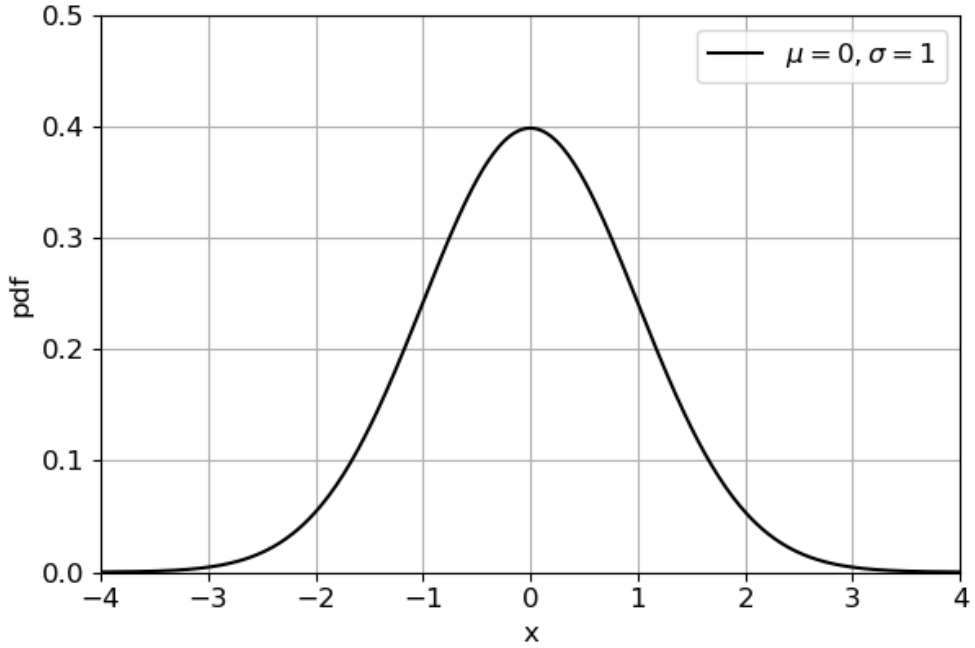
Recap of Q1 – From PDF to CDF

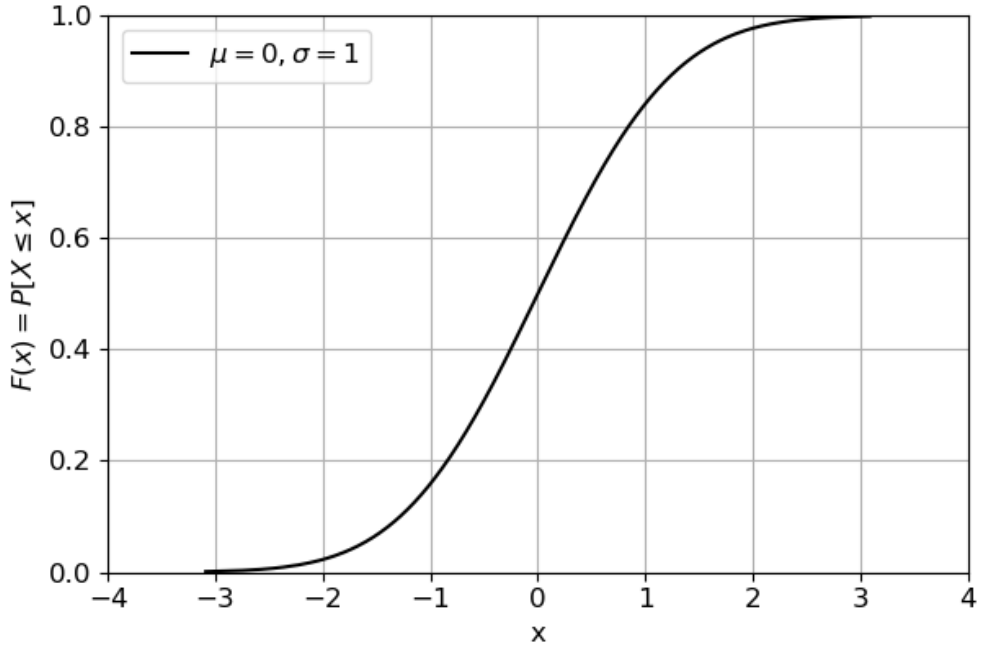
- Probability density function (PDF) $f_X(x)$

- Cumulative distribution function (CDF) $F(x) = \int_{-\infty}^{x} f(x)dx$

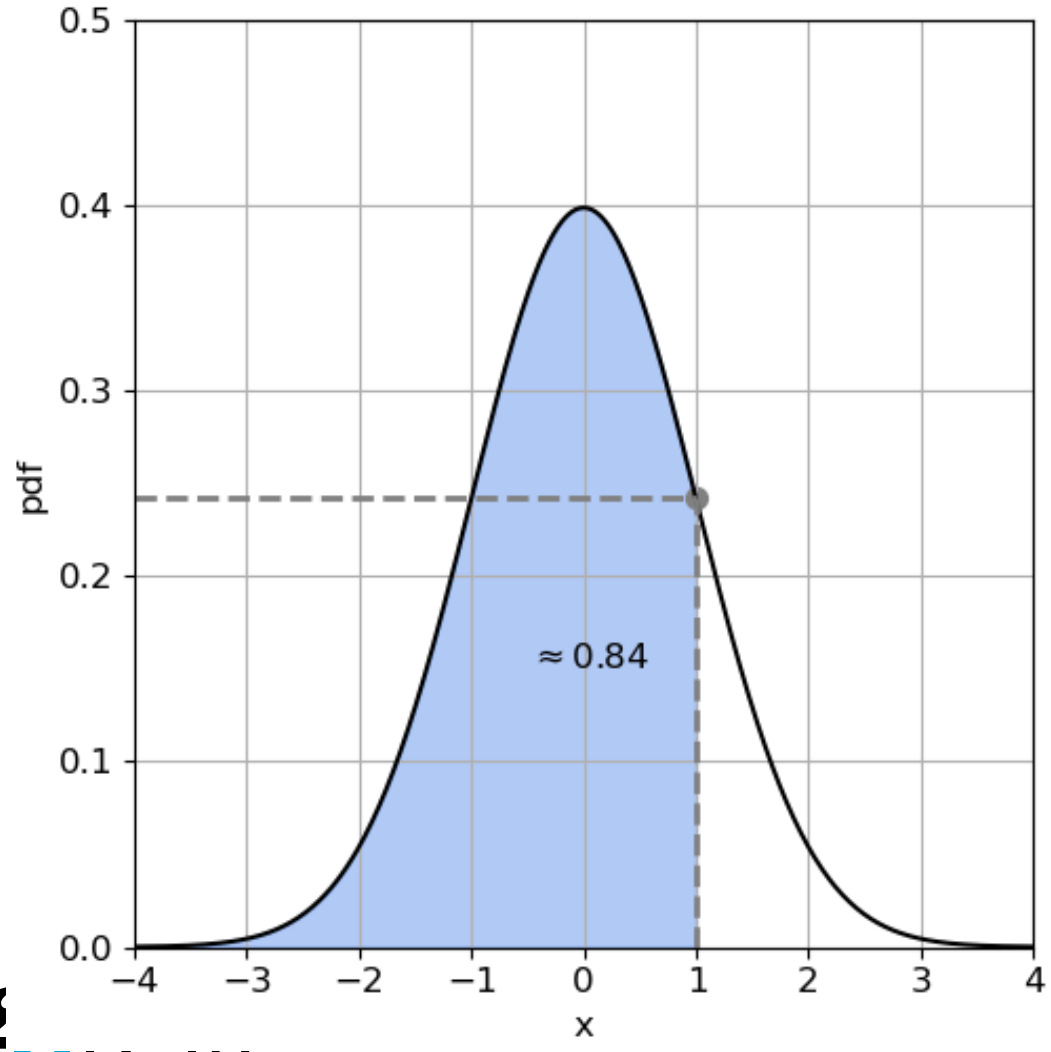
CDF of the Gaussian distribution

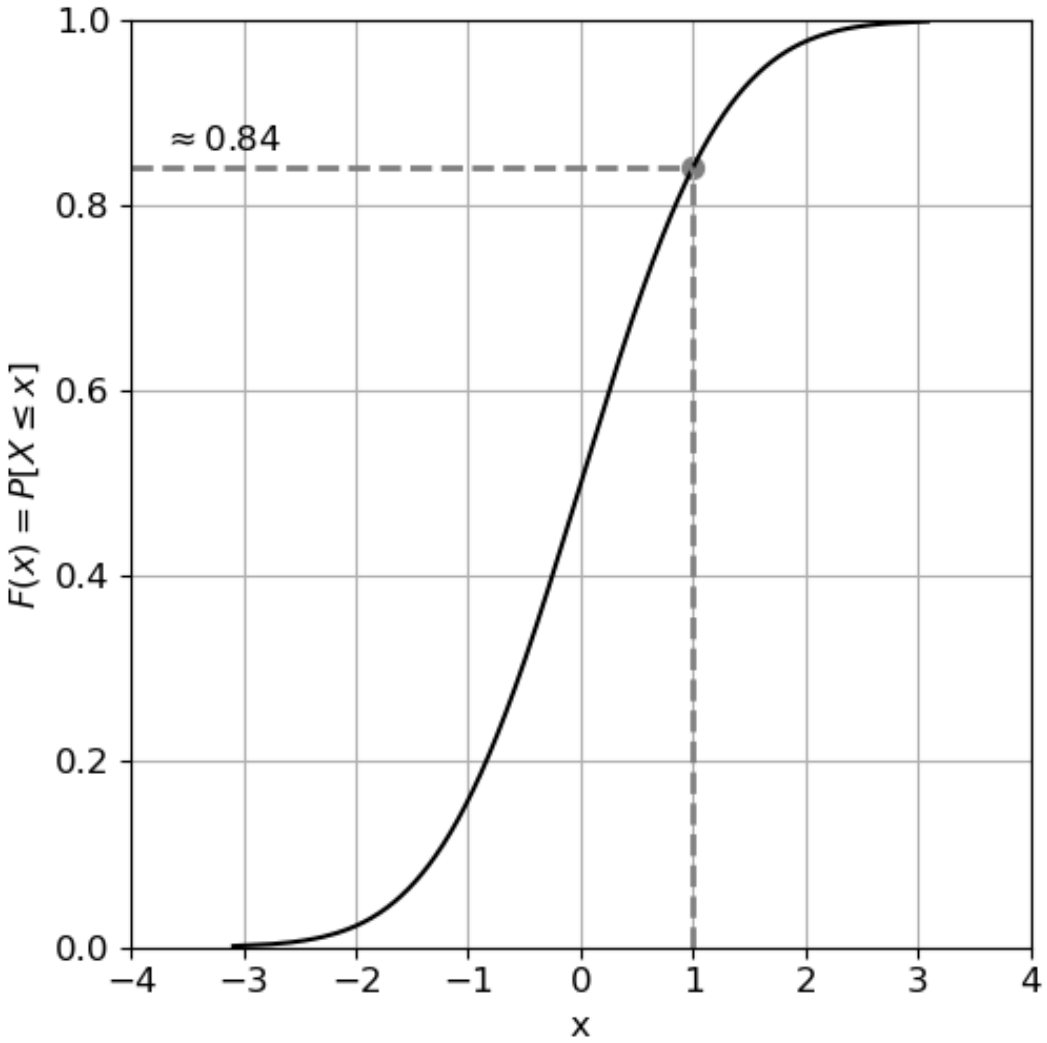
$$F(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$



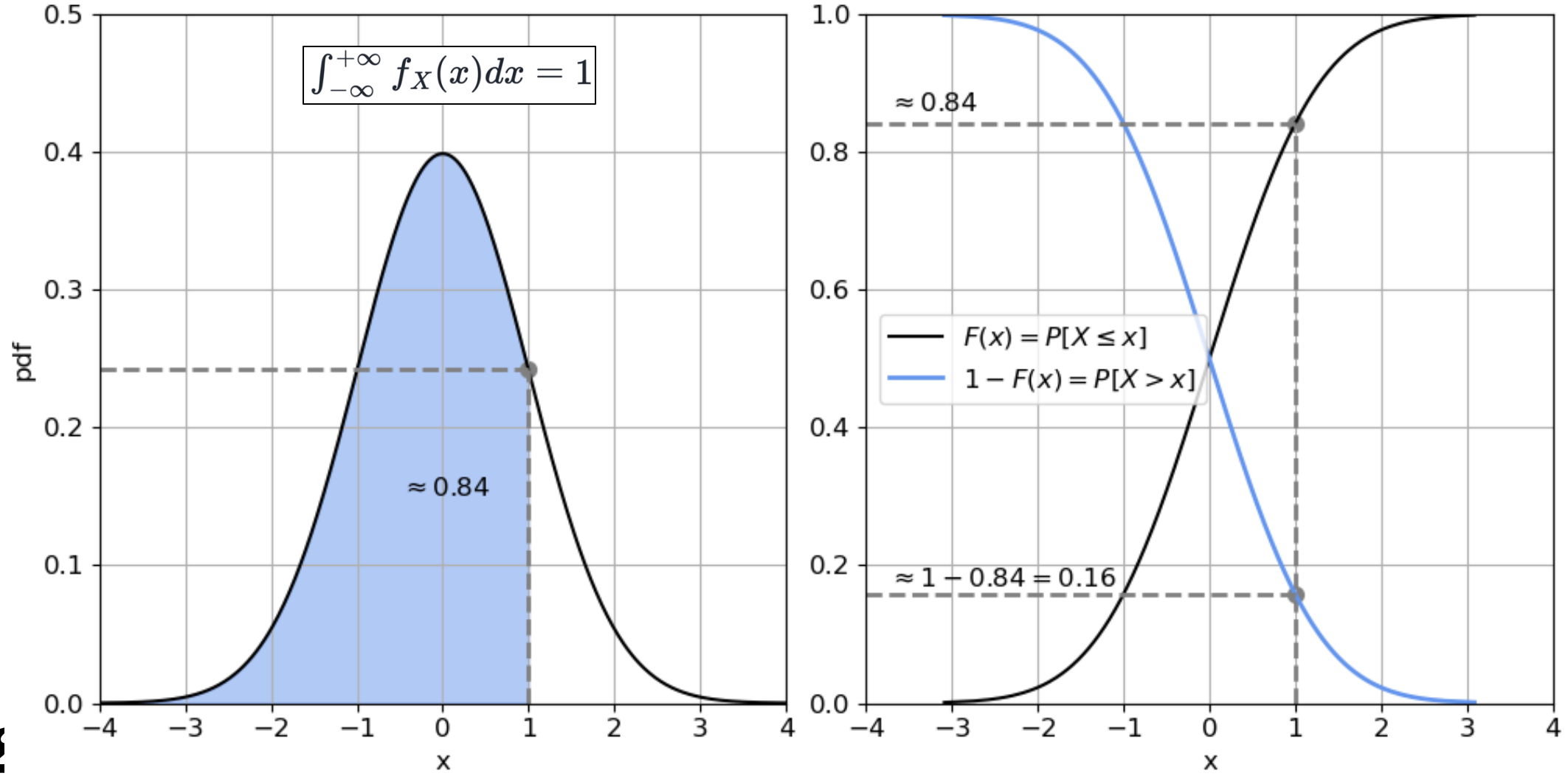


From PDF to CDF





Recap of Q1 – Exceedance



Continuous distribution functions

Mathematical model which relates the values of a random variable and their probability

But what do I want to model?

Observations  Empirical distribution function

Parametric distribution: model
Empirical distribution: observations

I want a model which is able to reproduce the probabilistic behavior in the observations

Empirical distribution function (ECDF)

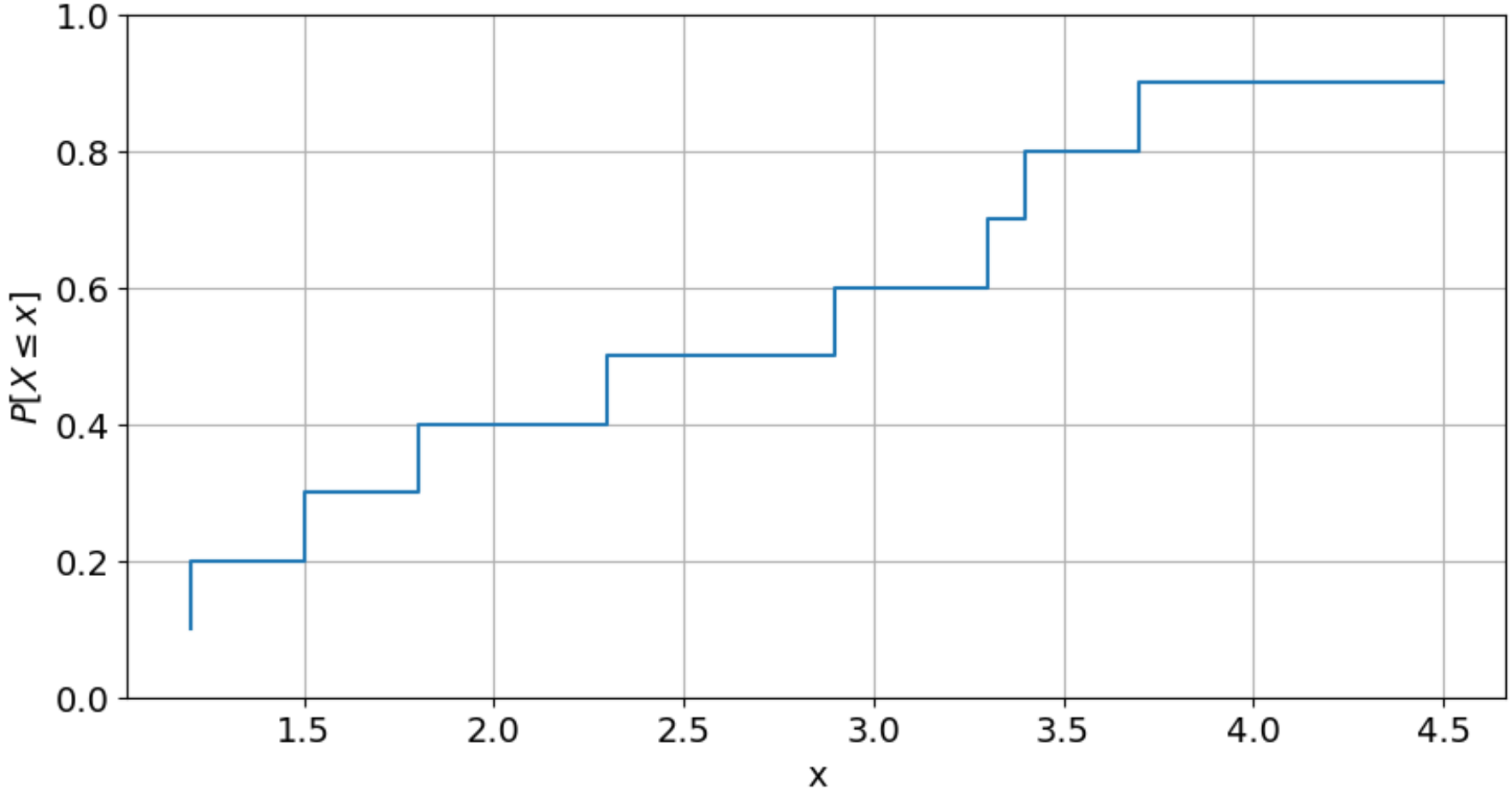
We need to assign a non-exceedance probability to each observation

Observations
2.3
1.2
4.5
1.8
3.4
3.7
3.3
2.9
1.5

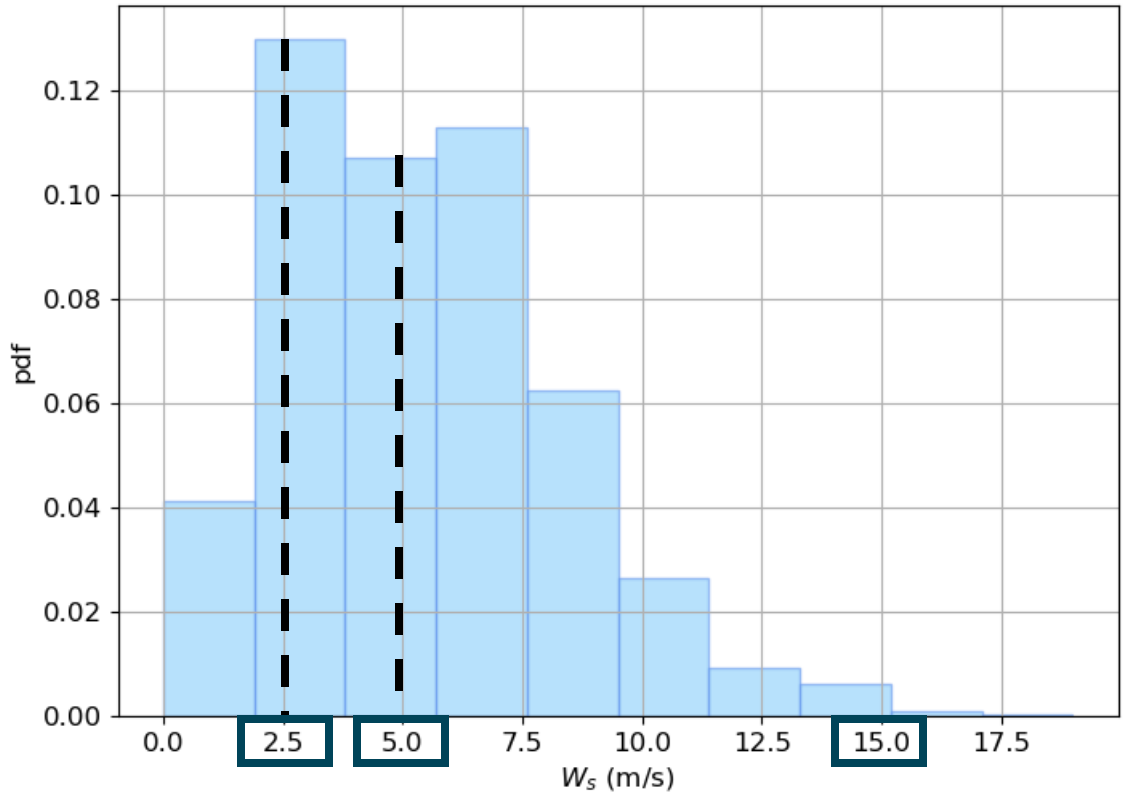
#observations = 9

Empirical distribution function (ECDF)

We need to assign a non-exceedance probability to each observation



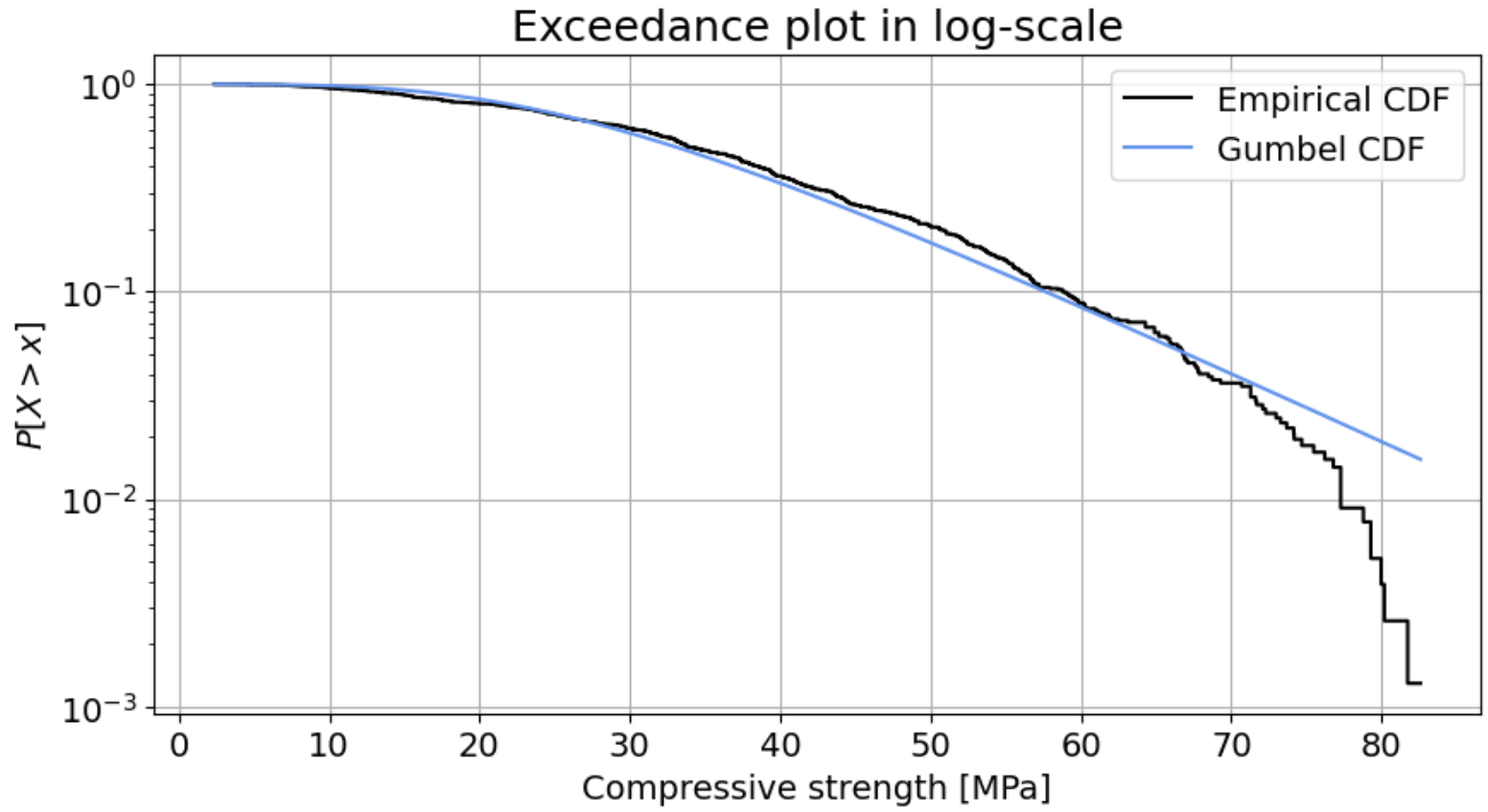
Recap of Q1 – Why is the tail important?



- You are designing a building against wind loading
- Which value would you use for design?
- We want the building to **perform in ordinary conditions** (around central moments)
- *We also want the building to **withstand the storms***

Recap of Q1 – Modelling the tail of the distribution

- Wednesday workshop in week 1.7 – modelling compressive strength of concrete



Focusing on the tail of the distribution: EVA

- Systems and infrastructures are typically designed and assessed for extreme conditions.

- Extreme events are located in the tails of the distribution

- Extreme events are typically scarce: short timeseries (e.g.: 20 years) in comparison with the design events that the system needs to withstand (e.g.: 1,000 years event).

- **Extreme Value Analysis (EVA) focuses on those events located at the tails of the distribution and provides a framework to identify and model the stochastic behavior of extreme events so events which have not been observed can be inferred.**

2. Extreme, design conditions and return period

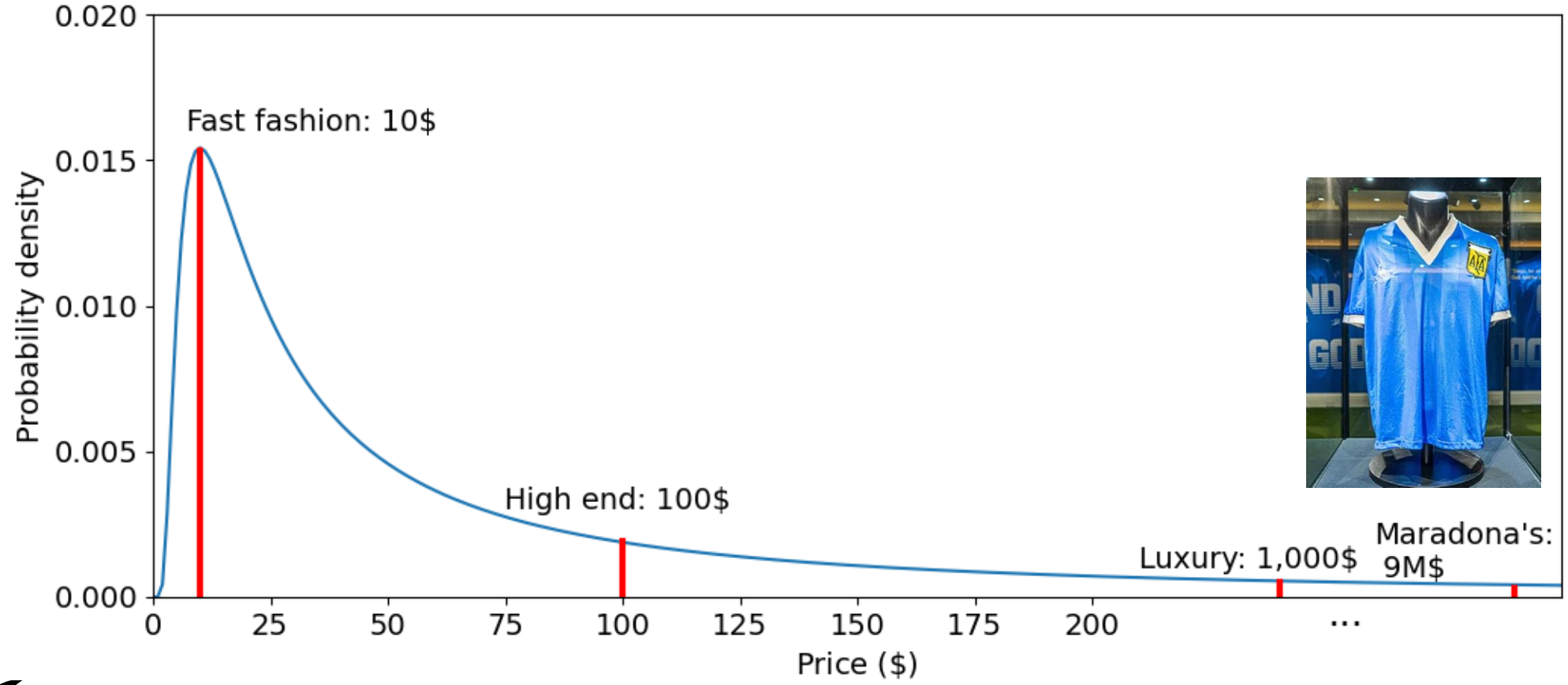
TU Delft

What's an extreme?

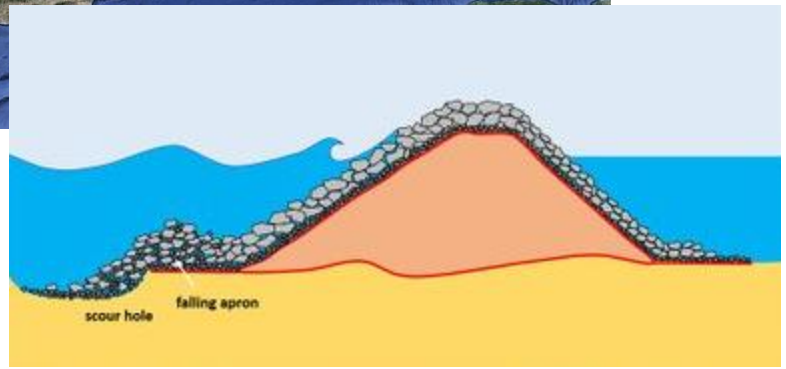
An **extreme observation** is an observation that **deviates from the average observations**

Example: t-shirt price

Extremes are located at the tail of the distribution!



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?***

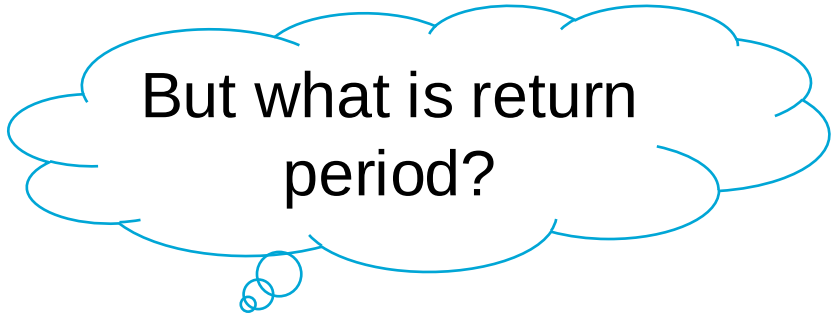
Design requirements

Which conditions does my intervention need to withstand?

Regulations and recommendations → Exceedance probability or **return period**

Country	Standard	**T_R (years)**	DL (years)	$p_{f,DL}$ (-)
England	BS 6349-1-1:2013	**50-100***	50-100	0.05*
Japan	TS Ports-2009	**50-100**	50	0.40-0.64
Spain	ROM 0.0-01/1.0-09	**113-4,975**	25-50	0.01-0.2

*Not well defined

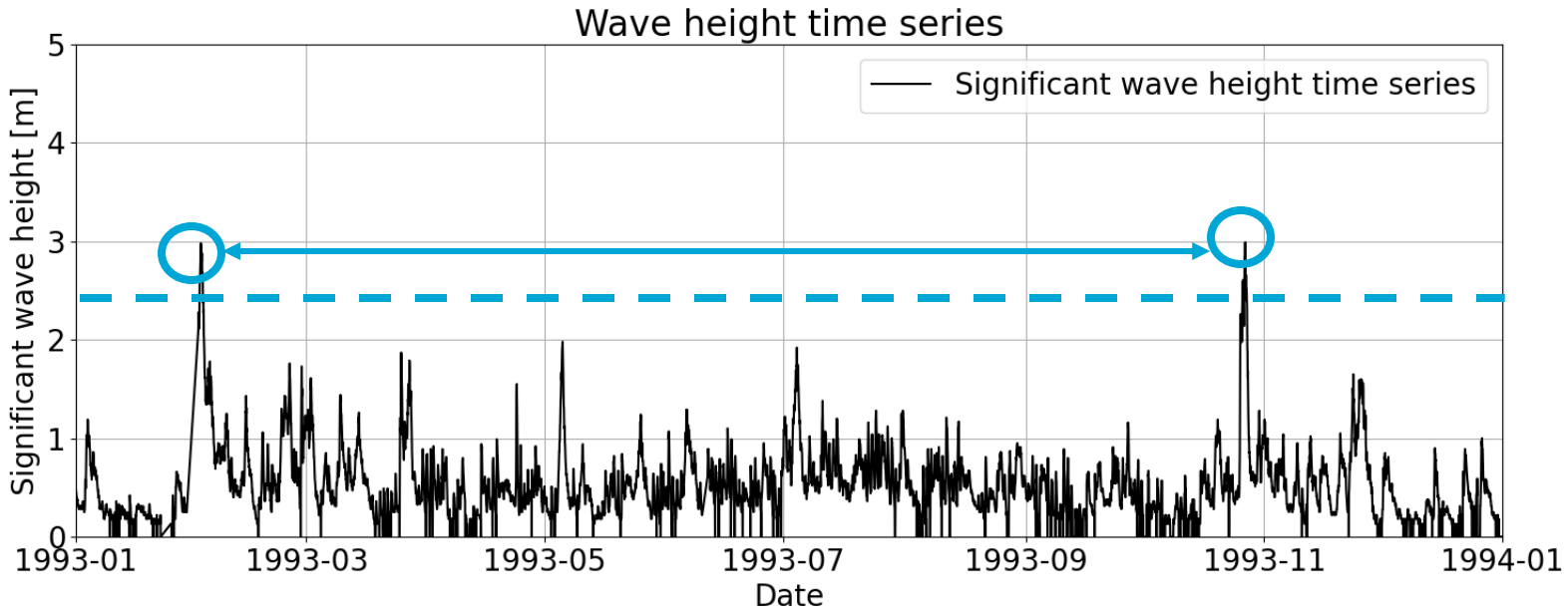


TU Delft

Design conditions: return period

- The return period is a concept in guidelines and recommendations in the Engineering and Geosciences field to parameterize the safety level. Then, the magnitude of the design event is given by a return period.

Return period is defined as is the expected time between two exceedances of extreme events.

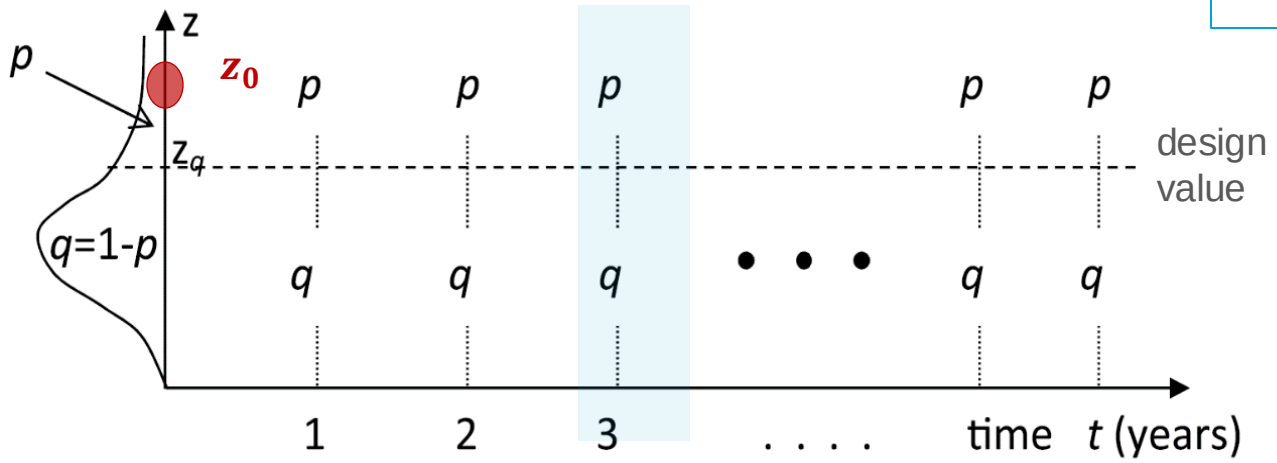


Return Period - Derivation

Mathematical definition

We are interested in estimating, on **average**, the **time** (e.g., year(*)) **at which an event** (here, the wave height) **higher than a given threshold**, (e.g. design value), **occurs**.
We know that $\Pr(Z > z_q) = 1 - q = p$

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



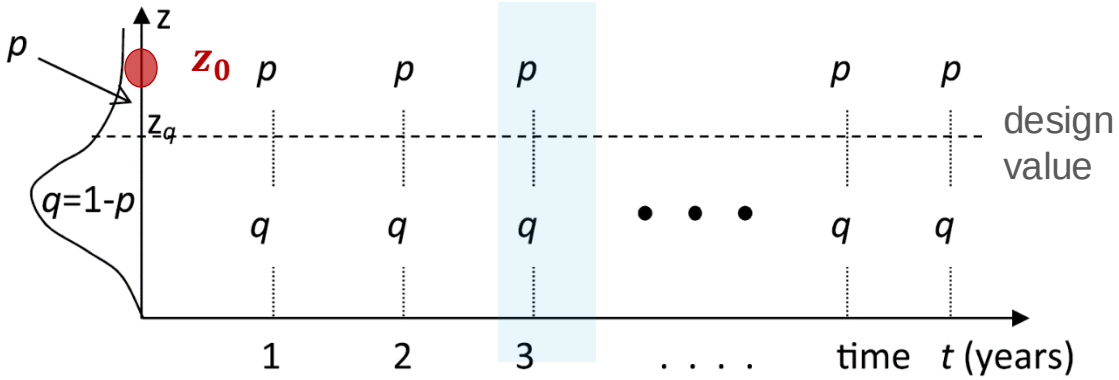
(*) the unit time reflects the interval time in which the observations are taken

Figure from Salas, et la (2013). *Journal of Hydrologic Engineering*, *19*(3), 554-568.

Return Period - Derivation

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

Let's calculate the probability that an event z_0 higher than the design value z_q occurs at time t



$$f(t) = Pr(z_0 \text{ at time } t) = (1-p)(1-p)\ldots(1-p)p$$

$$f(t) = Pr(z_0 \text{ at time } t) = q^{t-1}p$$

Geometric Distribution
it models the number of trials up to the first success (included)

$$T(t) = \frac{1}{p}$$

T(t) expectation
it will take on average 1/p trials to get a success

T is also defined as **Return Period** (in unit time).
"We have to make, on average, 1/p trials in order that the event happens once" (Gumbel)
or **wait 1/p years before the next occurrence**

Design requirements

Regulations and recommendations → Exceedance probability or **return period**

But also Design Life and the probability of failure during the design life ($p_{f,DL}$)

Country	Standard	T_R (years)	DL (years)	$p_{f,DL}$ (-)
England	BS 6349-1-1:2013	50-100*	50-100	0.05*
Japan	TS Ports-2009	50-100	50	0.40-0.64
Spain	ROM 0.0-01/1.0-09	113-4,975	25-50	0.01-0.2

*Not well defined

How do they relate to each other?

TU Delft

Back to basics – Bernoulli process

Extremes can be assimilated as a Bernoulli process

Bernoulli process	Extremes
Two possible outcomes: success or failure	✓ Each observation can be an over or below
Outcomes are mutually exclusive and collectively exhaustive	✓ over vs. below the design value
Constant probability of success	✓ stationarity
Independence between trials	✓ Hypothesis of EVA *iid* events

Back to basics – Binomial distribution

Extremes can be assimilated as a Bernoulli process

Number of exceedances (succeses) in a given number of trials follows a Binomial distribution

$$p_X(x) = P[X = x|n,p] = \binom{n}{x} p^x (1-p)^{n-x} \qquad for\ x = 0, 1, \dots, n; p \in [0,1]$$

$$p_X(x) = P[X = x|n,p] = 0 \qquad otherwise$$

where

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

Design requirements

Regulations and recommendations → Exceedance probability or **return period**

But also Design Life and the probability of failure during the design life (p_{DL})

Country	Standard	RT (years)	DL (years)	$p_{f,DL}$ (-)
England	BS 6349-1-1:2013	50-100*	50-100	0.05*
Japan	TS Ports-2009	50-100	50	0.40-0.64
Spain	ROM 0.0-01/1.0-09	113-4,975	25-50	0.01-0.2

*Not well defined

TU Delft

Design requirements – Binomial distribution

- $p_{f,DL}$ - $p_{f,y}$ - DL - T_R $\quad T_R = 1/p_{f,y}$
- Each year is a trial ⟹ Success (excess the design value) or failure (no excess)?
- The number of exceedances (successes) in a given number of years (trials) ~ Binomial

MODEL

- $p_{f,DL}$ is the probability of an excess at least once in the DL
- $p_{f,DL}$ = 1 – probability of no excess

$$p_X(0) = P[X = 0|DL, p_{f,y}] = \binom{DL}{0} {p_{f,y}}^0 (1 - p_{f,y})^{DL-0}$$

- $p_{f,DL} = 1 - (1 - p_{f,y})^{DL}$

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

TU Delft

Design requirements – Binomial distribution

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

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

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

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

Intermezzo – Poisson distribution

The Binomial distribution is defined as

$$p_X(x) = P[X = x|n, p] = \binom{n}{x} p^x (1-p)^{n-x}$$

If $n \to \infty$, x and p are finite and defined and p is very small, $\lambda = np$.

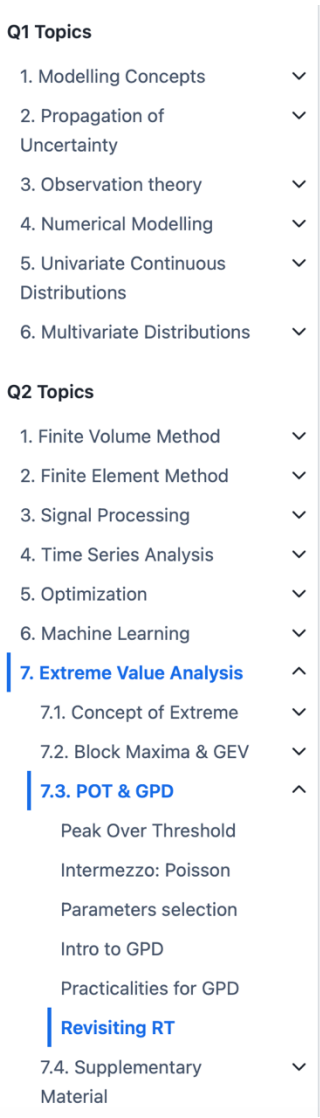
After some simplifications… **Poisson distribution**

$$p_X(x) = P[X = x|n, p] = \frac{\lambda^x e^{-\lambda}}{x!} \qquad for\ x = 0, 1, 2, \ldots\ and\ \lambda > 0$$

$$p_X(x) = P[X = x|p] = 0 \qquad otherwise$$

Binomial is based on **discrete events**, while the **Poisson** is based on **continuous events**. That is, in Poisson distribution $n \to \infty$ and p is very small, so you have an infinite number of trials with infinitesimal chance of success.

Also possible with the Poisson distribution!



Deriving the probability of failure along the design life

We already saw that extremes could be assimilated as a Bernoulli process: for each year (trial), we check if the observed value exceeds our design value (success) or not (failure). Thus, the number of exceedances over a design value (successes) in an infinite number of years (trials) will follow the Poisson distribution if each trial is independent and the probability of success (exceeding the threshold) is very small.

Let's calculate the probability of observing an event z_0 higher than the design value z_q at least once in DL years ($p_{f,DL}$).

First, we calculate the probability of not failing along DL applying the Poisson distribution. Remember that $\lambda = np$, being n the number of trials and p the probability of success.

$$p_X(0) = P[X = 0|DL, p_{f,y}] = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-DL \times p_{f,y}}$$

The probability of failing at least once in DL can be computed as $1 - p_X(0)$ (1 - no failure), so

$$p_{f,DL} = 1 - p_X(0) = 1 - e^{-DL \times p_{f,y}}$$

We defined $RT = 1/p_{f,y}$, so

$$p_{f,DL} = 1 - e^{-DL/RT} = 1 - e^{-DL/RT}$$

Rewriting it in terms of RT, we obtain

$$RT = \frac{-DL}{ln(1-p_{f,DL})}$$

And now you can compute the design RT based on the DL and $p_{f,DL}$ recommended in design guidelines!

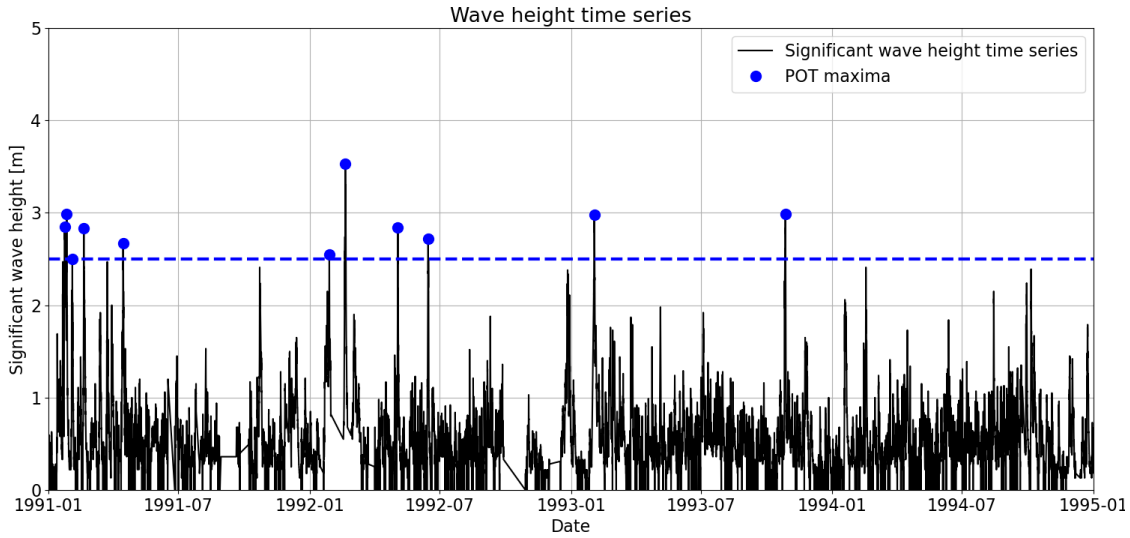
Break?

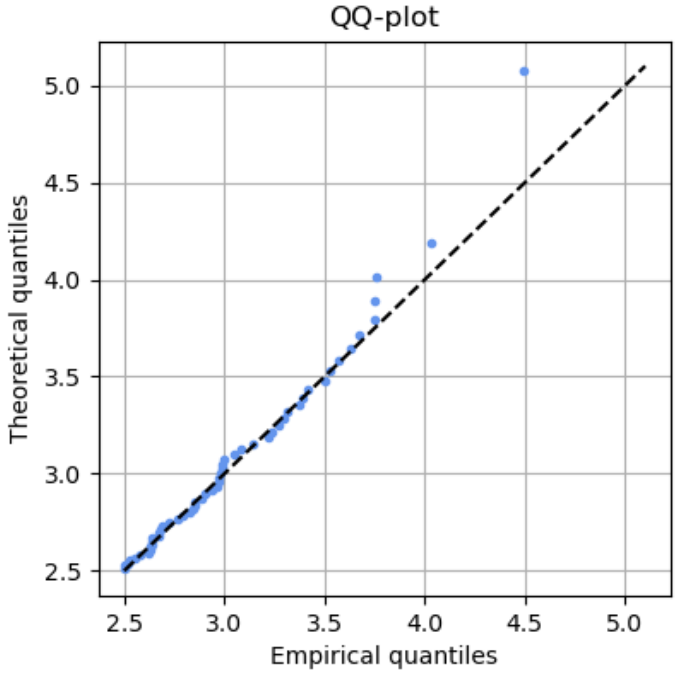
Please, leave the room through the door in the ground floor.

TU Delft

General outline of EVA

1. Select the extreme observations in the timeseries
2. Build a ECDF with the observations
3. Fit a parametric CDF to the ECDF (the distribution given by the method)
4. Check performance using GOF
5. Use the parametric CDF to infer extremes we have not observed yet



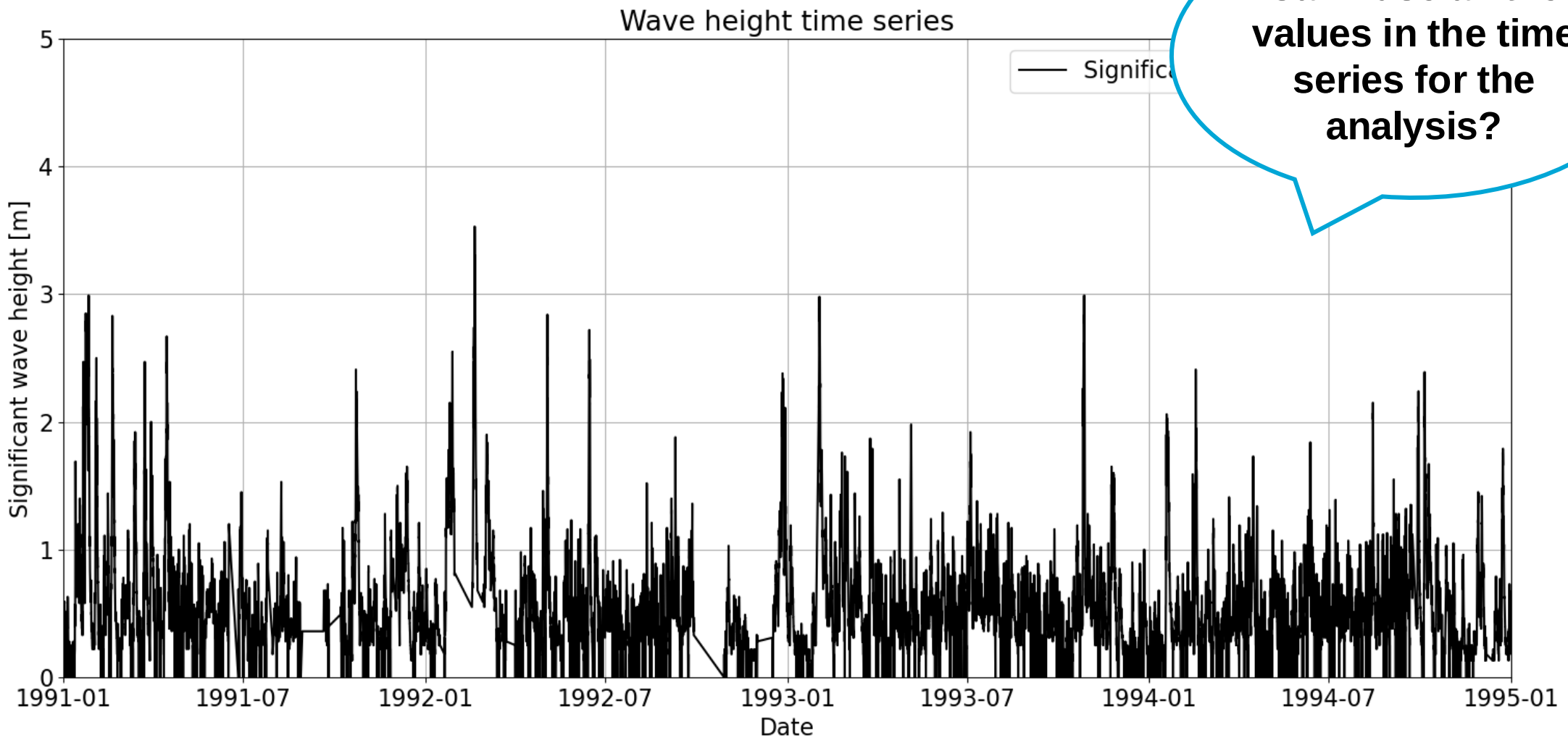


TU Delft

3. Block Maxima & Generalized Extreme Value (GEV) distribution

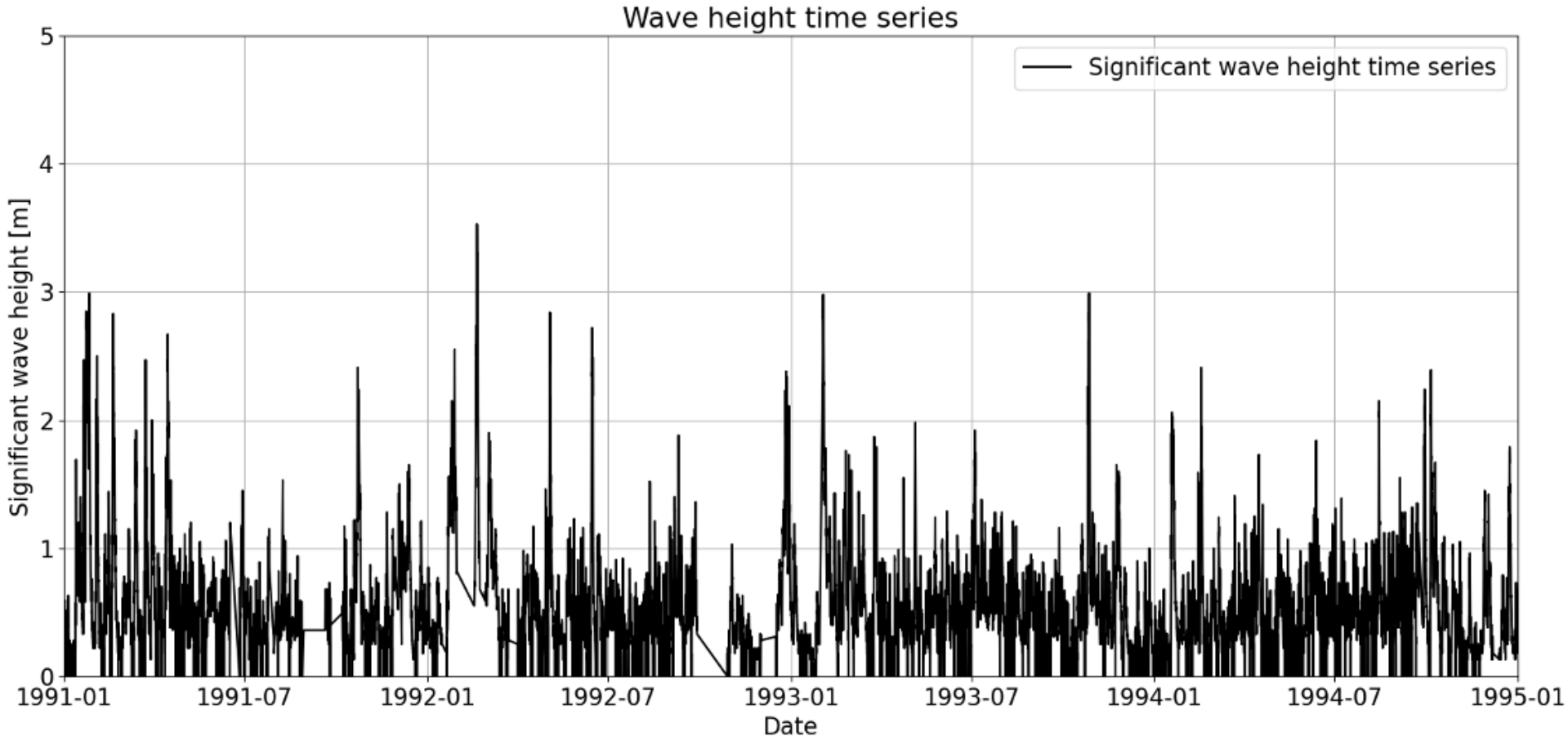
TU Delft

Time series



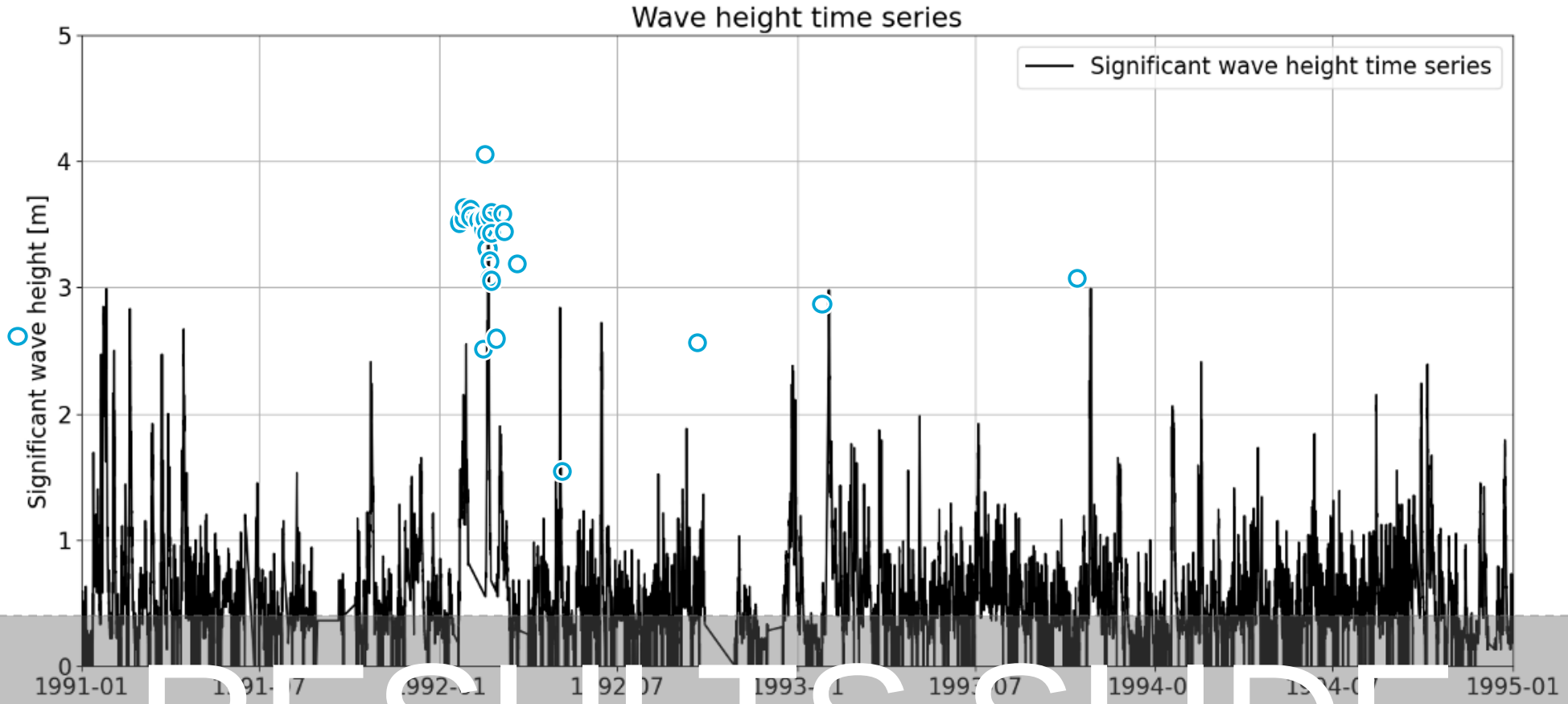
Can I use all the values in the time series for the analysis?

In the given time series, which observations are extreme and, thus, you would include in your analysis?



TU Delft

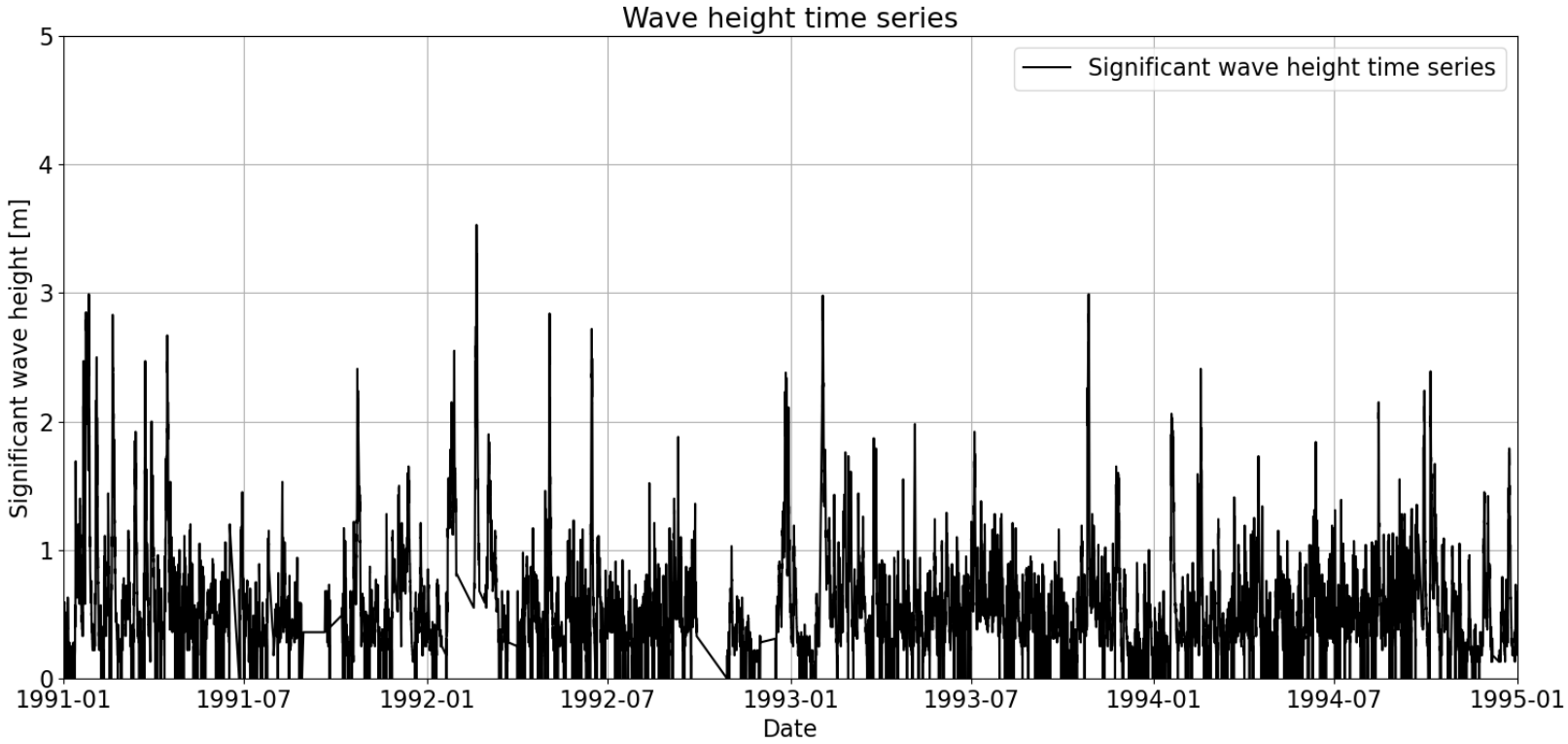
In the given time series, which observations are extreme and, thus, you would include in your analysis?



TU Delft RESULTS SLIDE

We need a systematic way to sample extreme values!

Each sampling technique defines a method of EVA

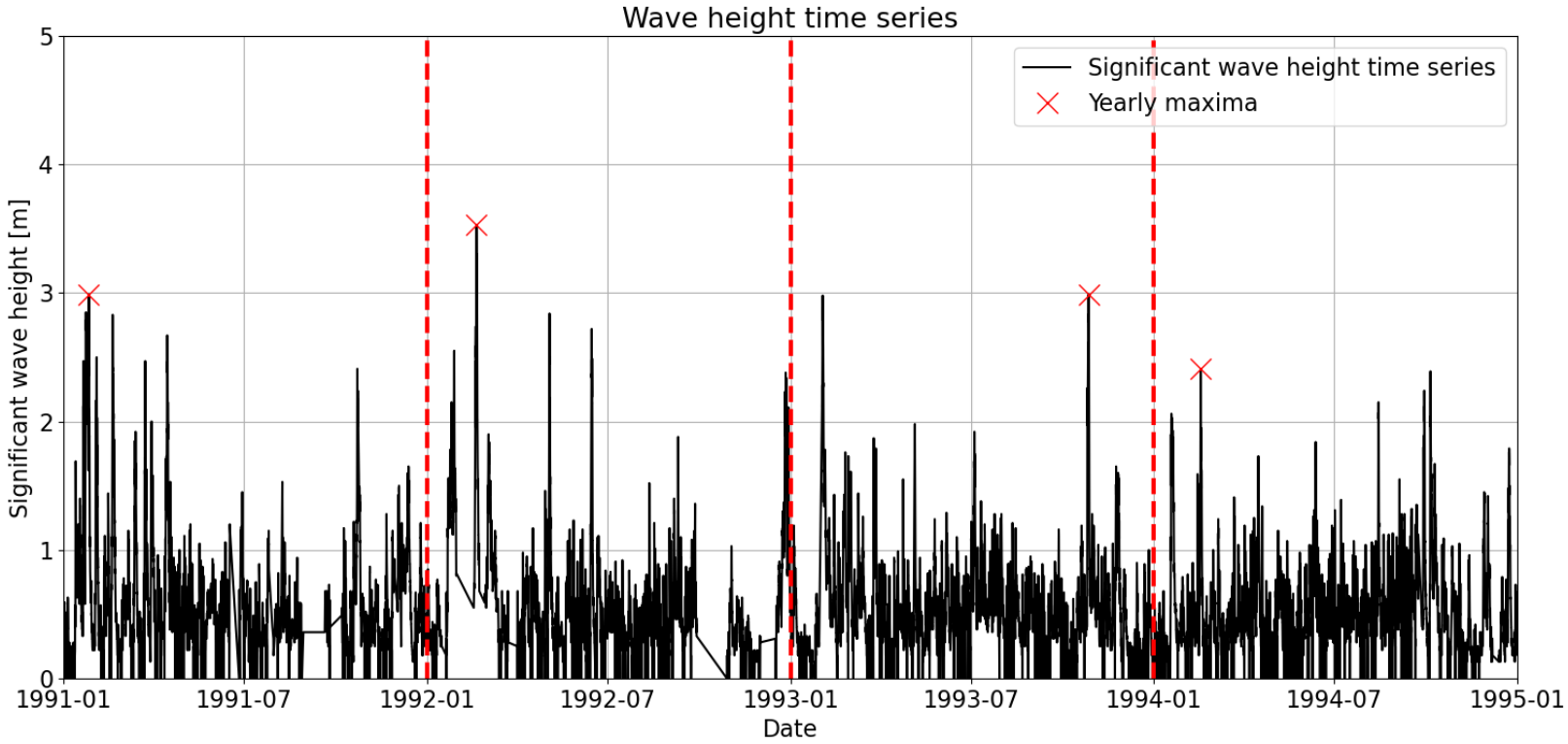


Two techniques:

1. **Block Maxima**
2. **Peak Over Threshold (POT)**

TU Delft

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

Generalized Extreme Value Distribution

- We are interested in modelling 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.

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

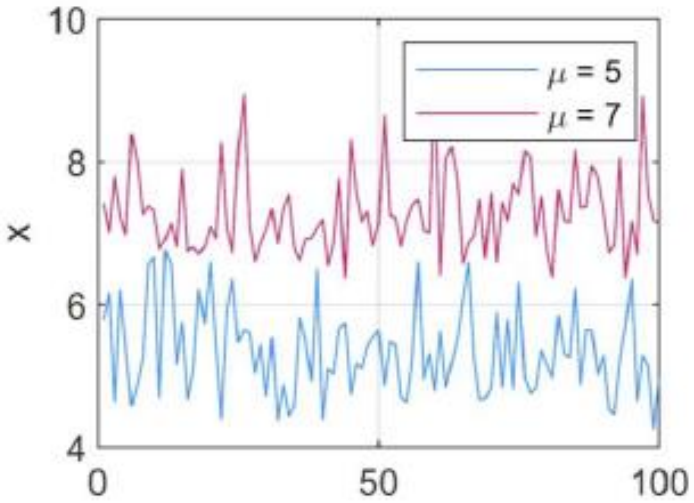
$$P[M_n \leq x] \rightarrow G(x)$$

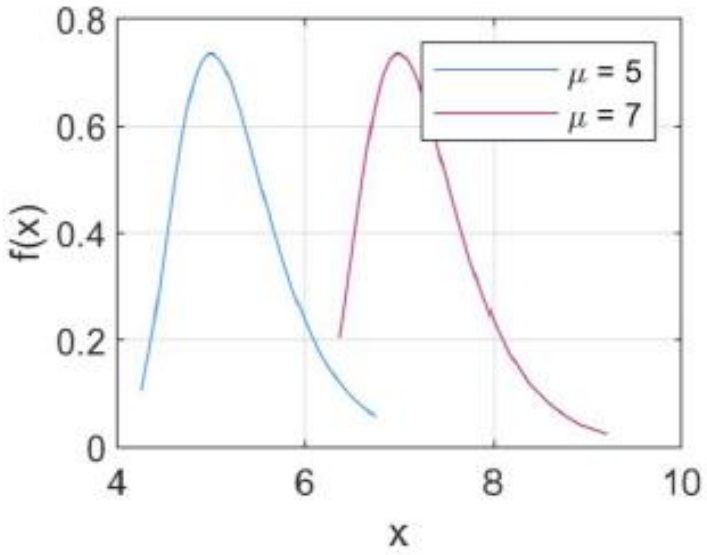
Generalized Extreme Value Distribution

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)

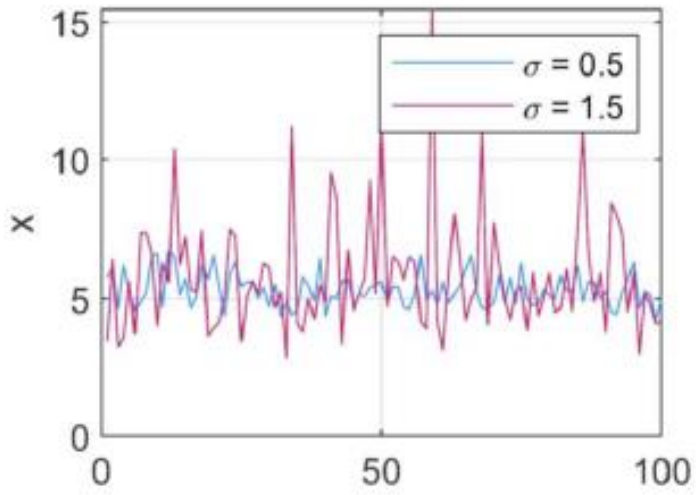
Higher μ, right displacement of the distribution, higher values.

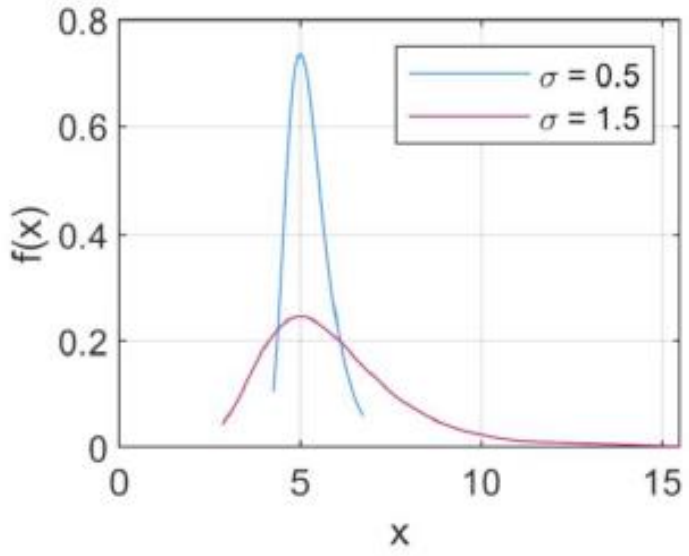
Generalized Extreme Value Distribution

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 ($-\infty < \xi < \infty$).





Scale parameter (σ)

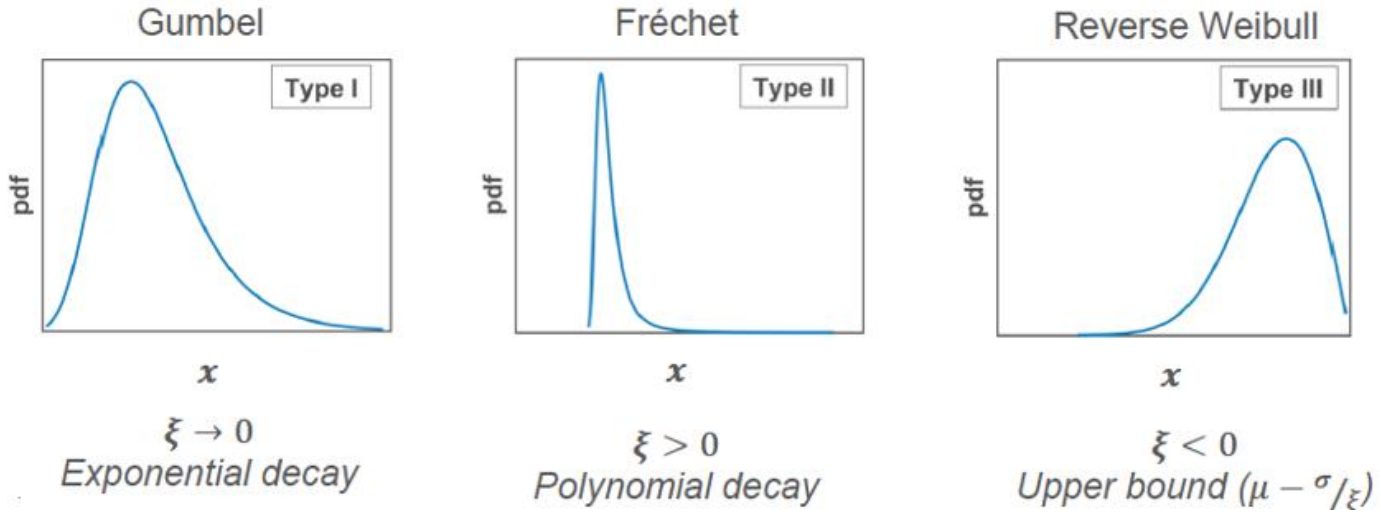
Higher σ, wider distribution.

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 ($-\infty < \xi < \infty$).



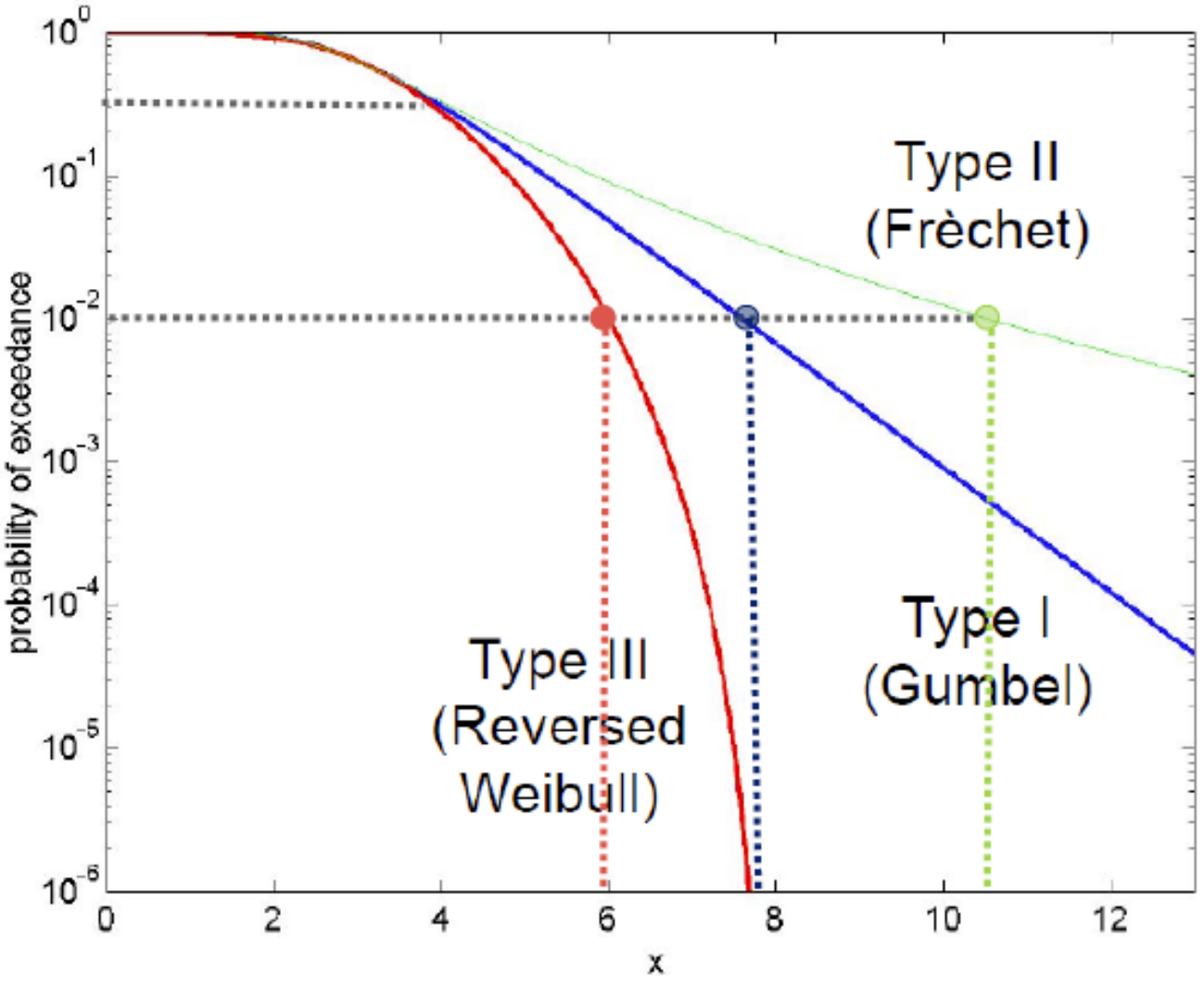
Shape parameter (ξ)

Determines the tail of the distribution.

Generalized Extreme Value Distribution

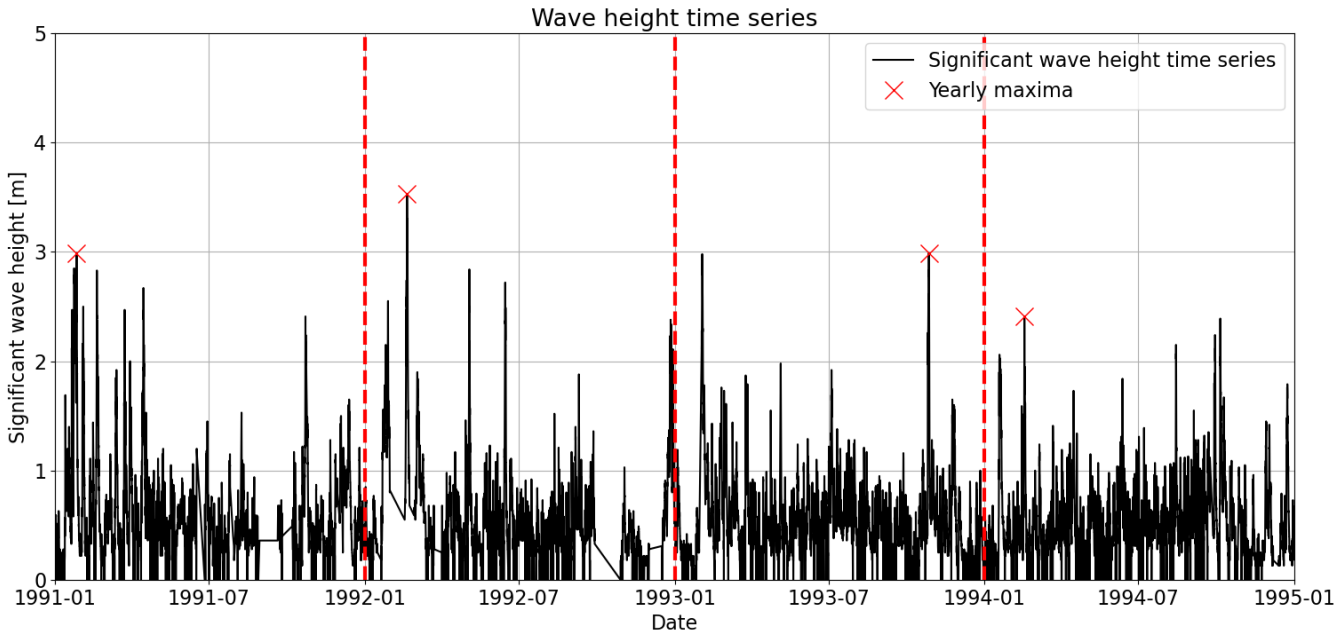
Plotting the tails…

- **Gumbel:** light tail
- **Fréchet:** heavy tail
- **Reversed Weibull:** bounded at $x = \mu - \frac{\sigma}{\xi}$



Let's apply it

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



$$T_R = \frac{1}{p_{f,y}} \rightarrow p_{f,y} = \frac{1}{90} = 0.011$$

- **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)

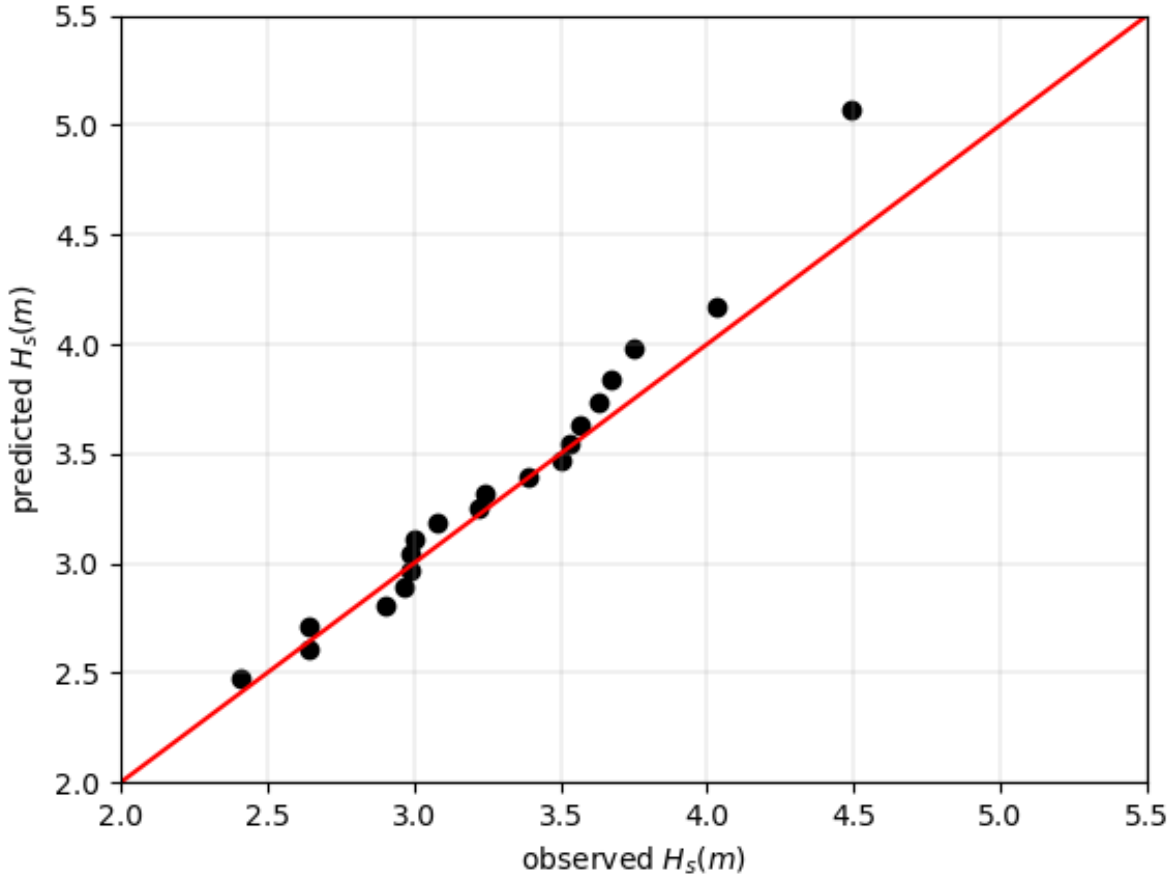
check fit (e.g., QQ-plot or Kolmogorov-Smirnov test)

inverse GEV to determine the design event

TU Delft

Let's apply it

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



$$T_R = \frac{1}{p_{f,y}} \rightarrow p_{f,y} = \frac{1}{90} = 0.011$$

- **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

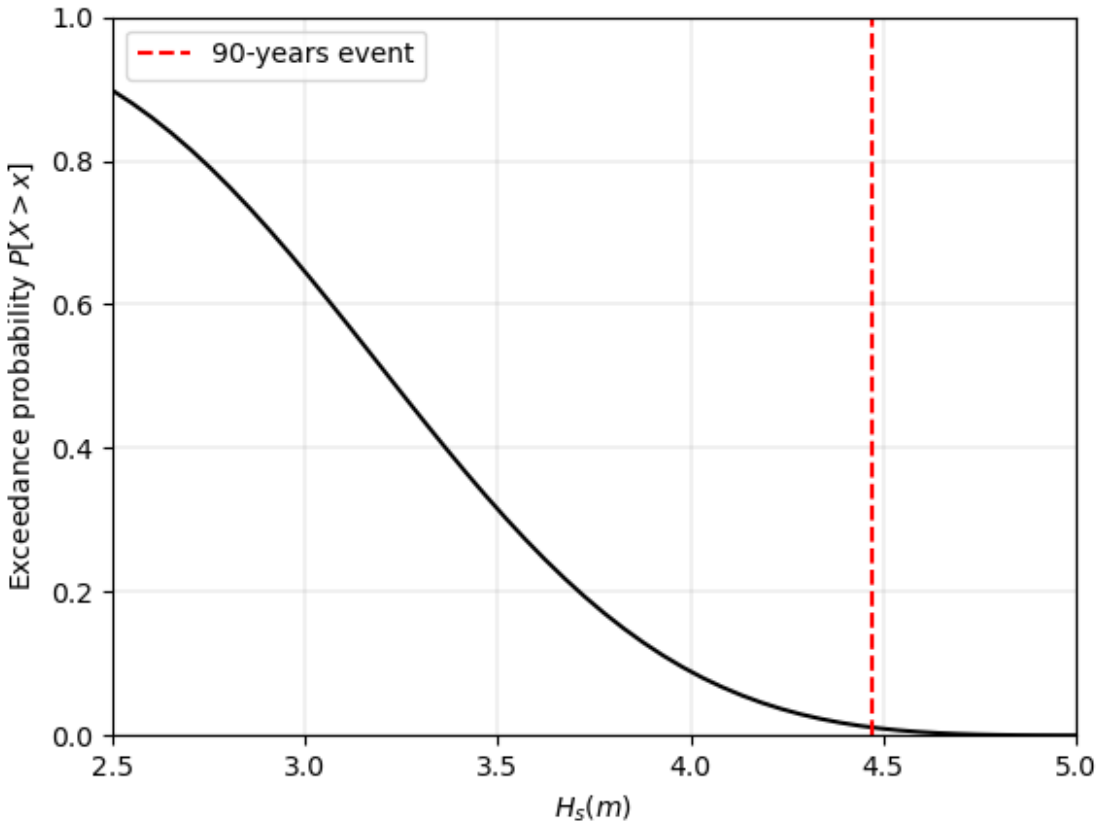
fit GEV(obs_max)

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

inverse GEV to determine the design event

Let's apply it

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



$$T_R = \frac{1}{p_{f,y}} \rightarrow p_{f,y} = \frac{1}{90} = 0.011$$

- **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**.

??

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

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

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

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

1-$p_{f,d}$

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

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

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

- Can you compute the return level of Q for $\mu = 50, \sigma = 25$ and $\xi = -1$?

Computing the return level

$\mu = 50, \sigma = 25$ and $\xi = -1$
$p_{f,d} = 2.7 \cdot 10^{-5}$ 1/days

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

$$-\ln(G(x)) = \left[1 + \xi\frac{x-\mu}{\sigma}\right]^{-\frac{1}{\xi}}$$

$$-\ln(G(x))^{-\xi} - 1 = \xi\frac{x-\mu}{\sigma}$$

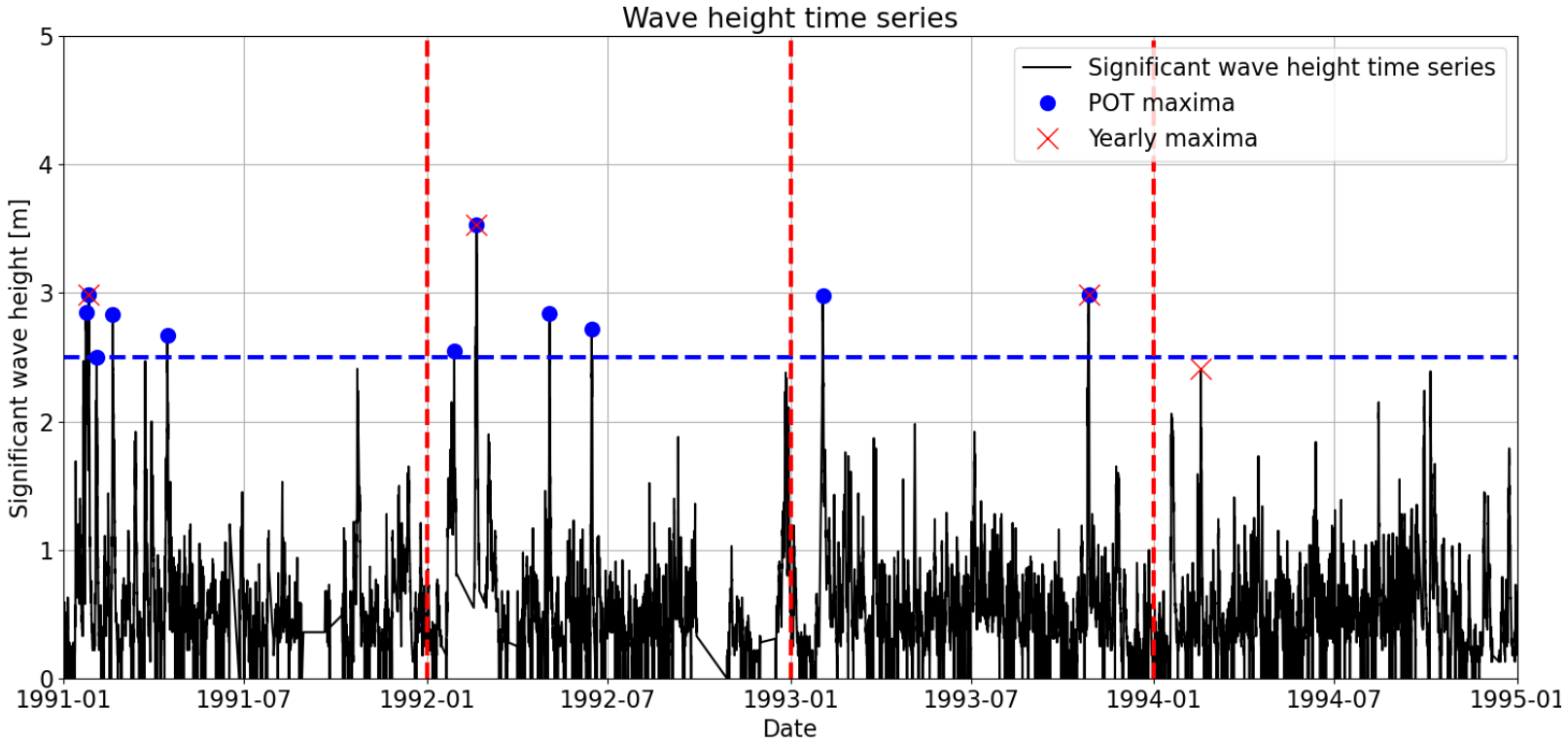
$$x = \frac{\sigma}{\xi}\left[-\ln(G(x))^{-\xi} - 1\right] + \mu$$

$$x = \frac{25}{-1}[-\ln(1 - 2.7 \cdot 10^{-5})^{1} - 1] + 50 \approx 75 m^3/s$$

4. Peak Over Threshold (POT) & Generalized Pareto distribution (GPD)

TU Delft

Sampling extremes: Peak Over Threshold (POT)

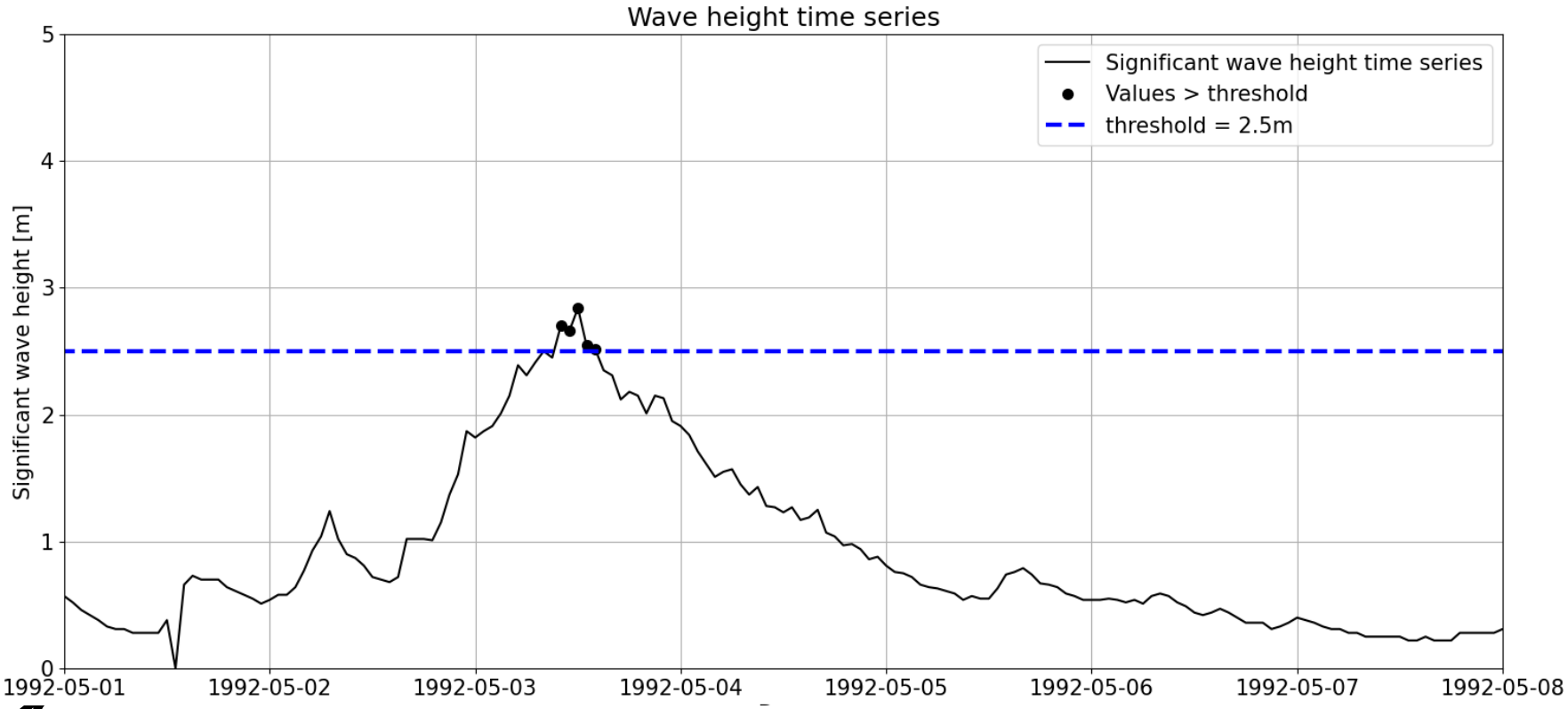


2. Peak Over Threshold (POT)

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

Choosing POT parameters

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

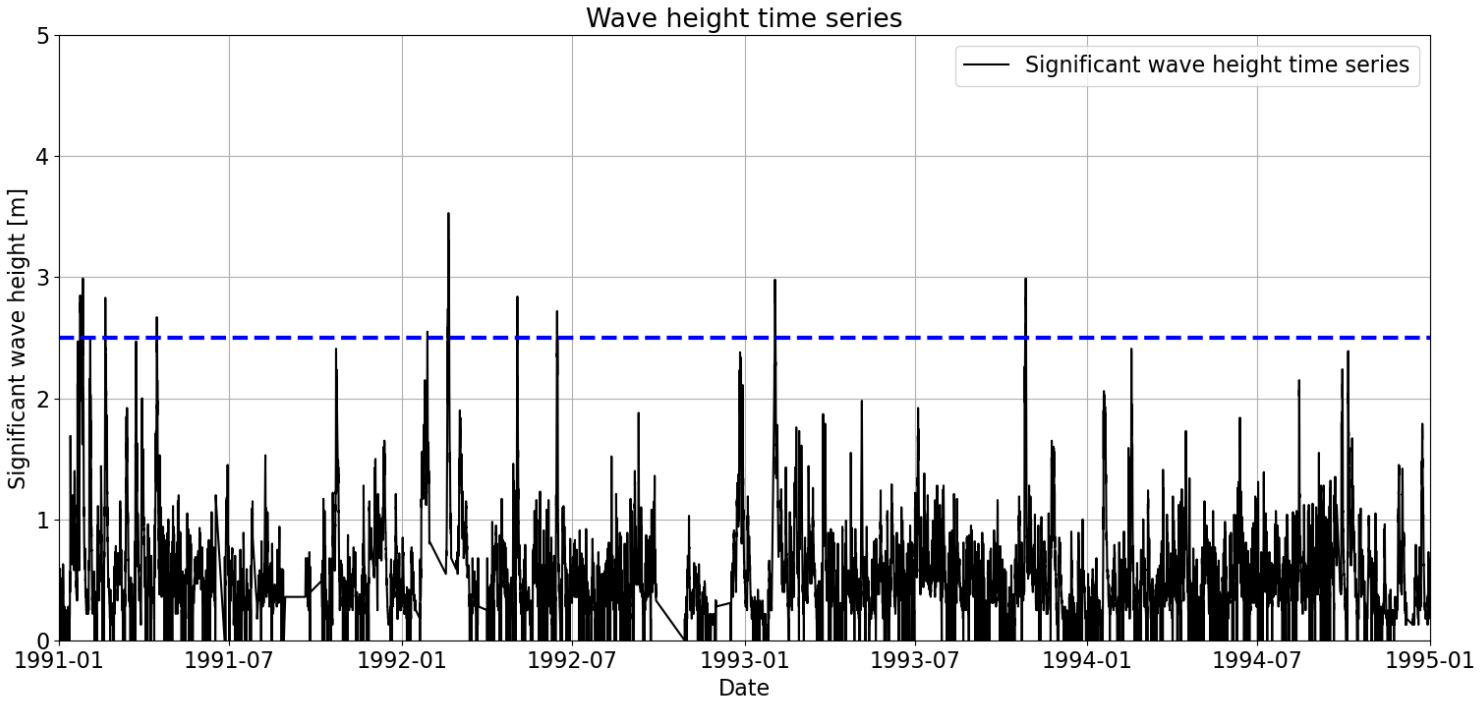


Extremes cluster in time!

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



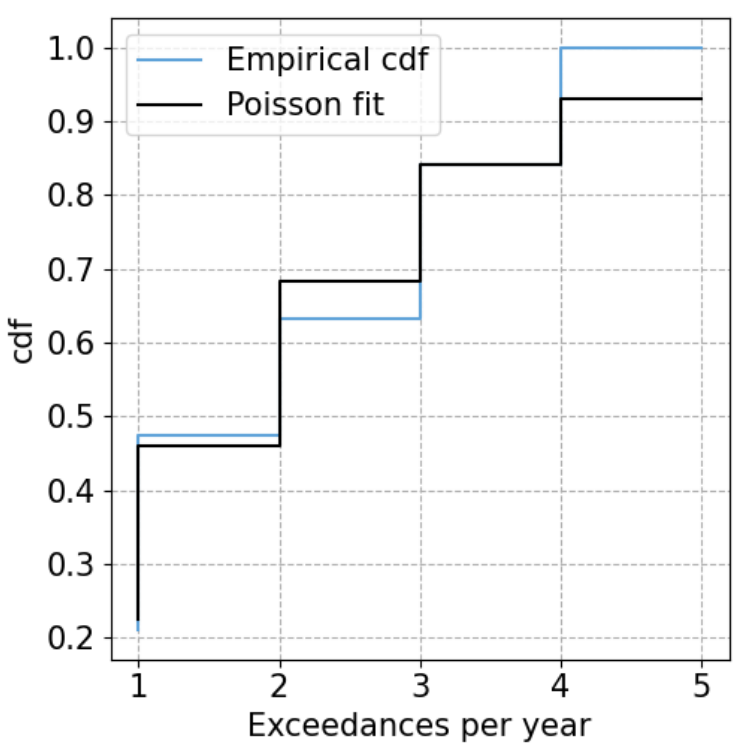
- Each hour is a trial ($n \rightarrow \infty$)
- Over or below the threshold?
- p_{above} is very small (tail of the distribution)
- Block = 1 year
- Number of excesses in each block over the threshold ~ Poisson

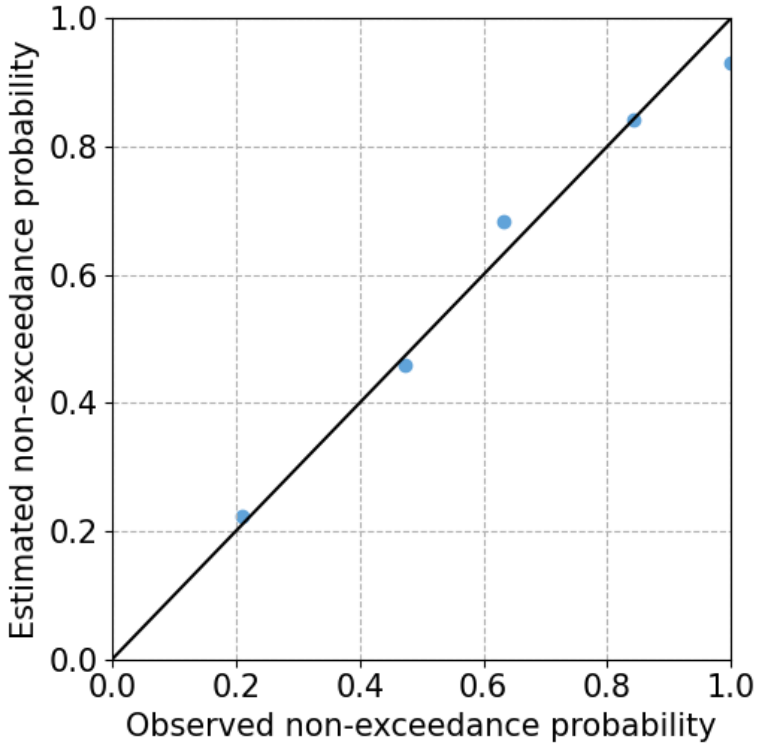
Almost all the techniques to formally select the threshold and declustering time for POT are based on the assumption that the sampled extremes should follow a Poisson distribution.

TU Delft

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

TU Delft

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 $X > th$

Generalized Pareto Distribution

Generalized Pareto distribution of the excesses is defined as

$$H(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma_{th}}\right)^{-1/\xi} & for\ \xi \neq 0 \\ 1 - exp\left(-\frac{y}{\sigma_{th}}\right) & for\ \xi = 0 \end{cases}$$

where $y \geq 0$ if $\xi \geq 0$, and $0 \leq y \leq -\frac{\sigma_{th}}{\xi}$ if $\xi < 0$.

These are conditional probabilities to *X>th*. As function of the random variable *X* and the threshold *th*

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

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 ($-\infty < \xi < \infty$).

Relationship with GEV's parameters

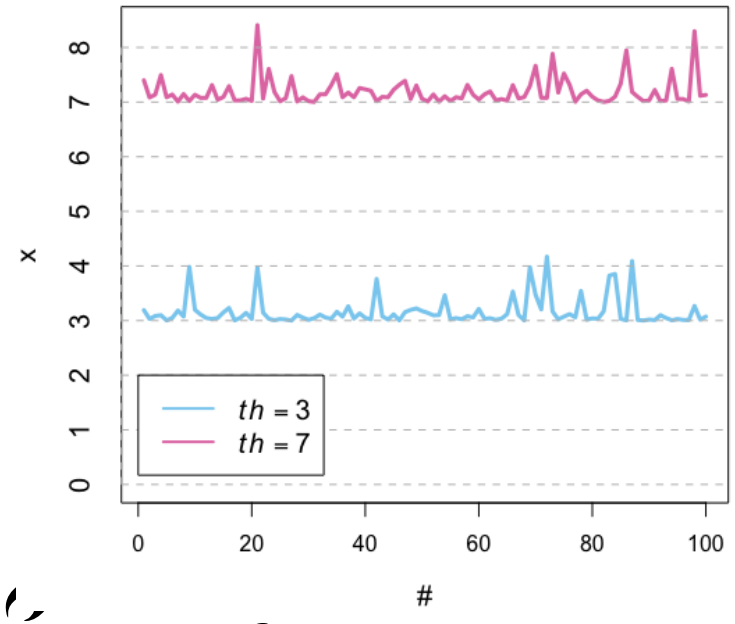
- Shape parameter is the same
- σ_{th} defined based on GEV's parameters as

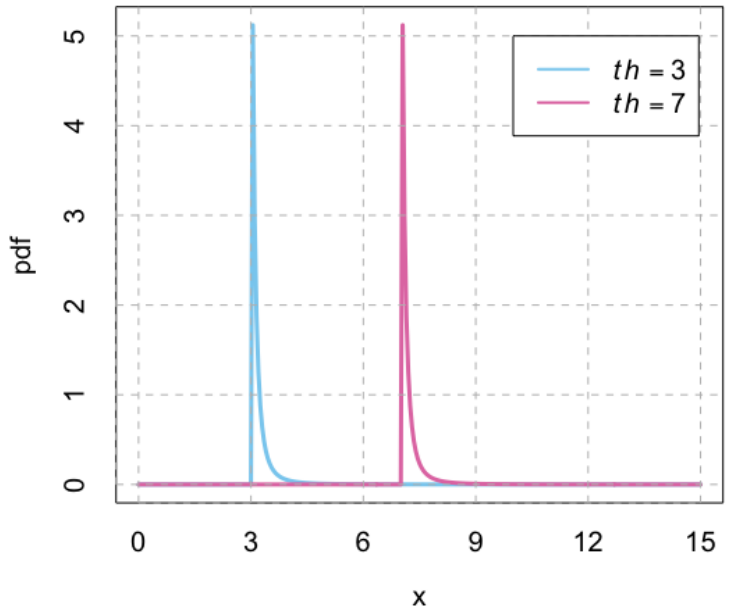
$$\sigma_{th} = \sigma + \xi(th - \mu)$$

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 ($-\infty < \xi < \infty$).





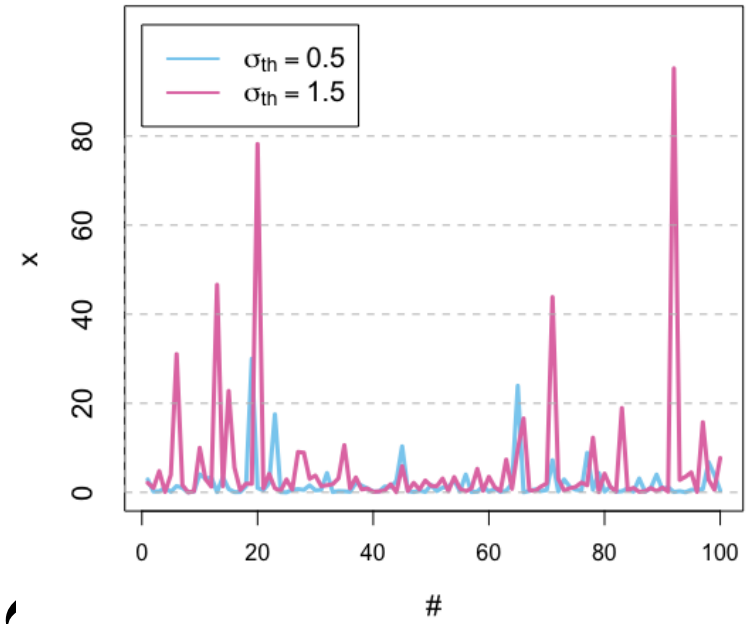
Threshold (*th*)

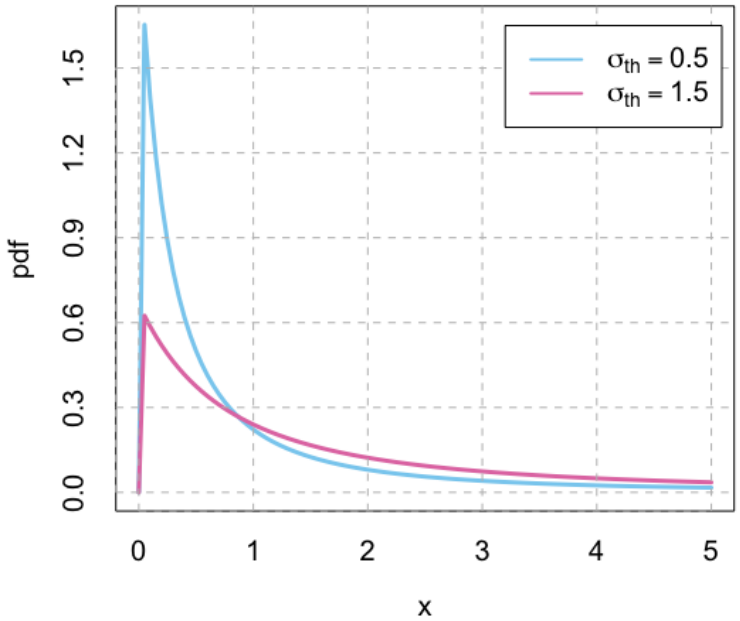
Acts like a location parameter.

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 ($-\infty < \xi < \infty$).





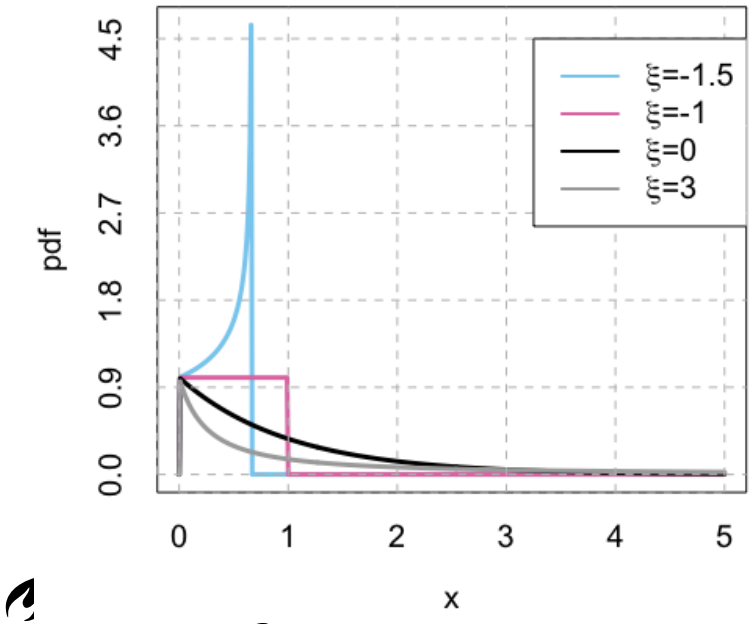
Scale parameter (σ_{th})

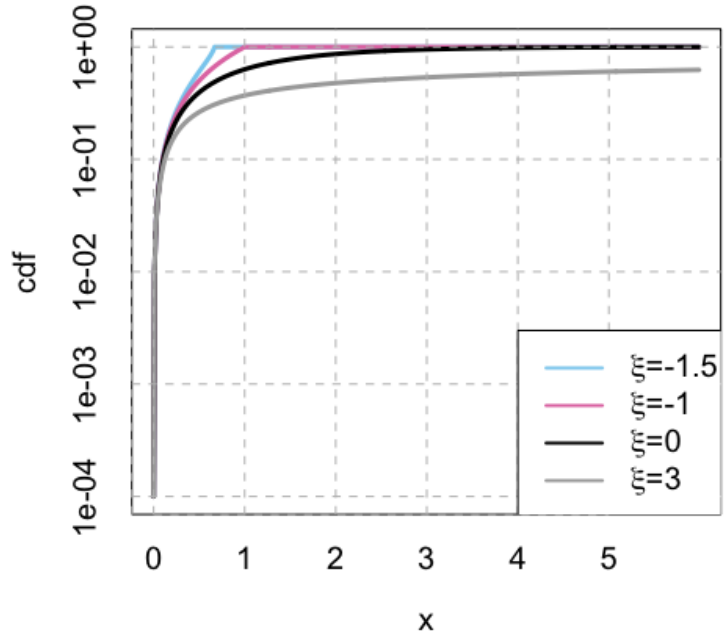
Higher σ_{th}, wider distribution.

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 see how to put it into practice

- Extremes of discharges Q are sampled with POT using *th* =50 m^3/s and *dl* = 24h on the observations which last for 5 years. We have sampled 2,000 extremes. A GPD is fitted to the excesses with $\sigma_{th} = 25$ and $\xi = 0.1$.
- We want to compute the discharge associated with a **return period of 100 years**.

$$H(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma_{th}}\right)^{-1/\xi} & for\ \xi \neq 0 \\ 1 - exp\left(-\frac{y}{\sigma_{th}}\right) & for\ \xi = 0 \end{cases}$$

where the excesses are defined as $Y=X-th$ for $X>th$.

Let's talk again about 'units'

- POT using *th* = 50 m^3/s and *dl* = 24h
- N = 5 years
- n_{th} = 1,000 extremes
- GPD is fitted with $\sigma_{th} = 25$ and $\xi = 0.1$
- T_R = 100 years

$$T_R = \frac{1}{p_{f,y}} = \frac{1}{100} = 0.1 \text{ 1/days}$$

→'Units' of the GPD?
'Event-wise', irregular number of events per year

We assume that the number of extremes per year is Poisson-distributed.

We use the average number of excesses each year: $\lambda = \frac{1{,}000}{5} = 200$

$$T_R = \frac{1}{p_{f,y}} \rightarrow p_{f,d} = \frac{1}{T_R} = \frac{1}{100\ years} \frac{1\ year}{200\ events} = 5\ 10^{-5} \text{ 1/events}$$

TU Delft

Remember the threshold!

- POT using *th* = 50 m^3/s and *dl* = 24h
- N = 5 years
- n_{th} = 1,000 extremes
- GPD is fitted with $\sigma_{th} = 25$ and $\xi = 0.1$
- T_R = 100 years

$$H(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma_{th}}\right)^{-1/\xi} & for\ \xi \neq 0 \\ 1 - exp\left(-\frac{y}{\sigma_{th}}\right) & for\ \xi = 0 \end{cases}$$

$$p_{f,d} = \frac{1}{T_R} = \frac{1}{100\ years}\frac{1\ year}{200\ events} = 5\ 10^{-5}\ 1/events$$

Since $\xi = 2 (\neq 0)$, $H(y) = 1 - \left(1 + \frac{\xi y}{\sigma_{th}}\right)^{-1/\xi}$

$$\left(1 + \frac{\xi y}{\sigma_{th}}\right)^{-1/\xi} = 1 - H(y)$$

$$1 + \frac{\xi y}{\sigma_{th}} = [1 - H(y)]^{-\xi}$$

$$y = \frac{\sigma_{th}([1 - H(y)]^{-\xi} - 1)}{\xi} =$$

$$= \frac{25([1 - (1 - 0.00005)]^{-0.1} - 1)}{0.1} = 249\frac{m^3}{s}$$

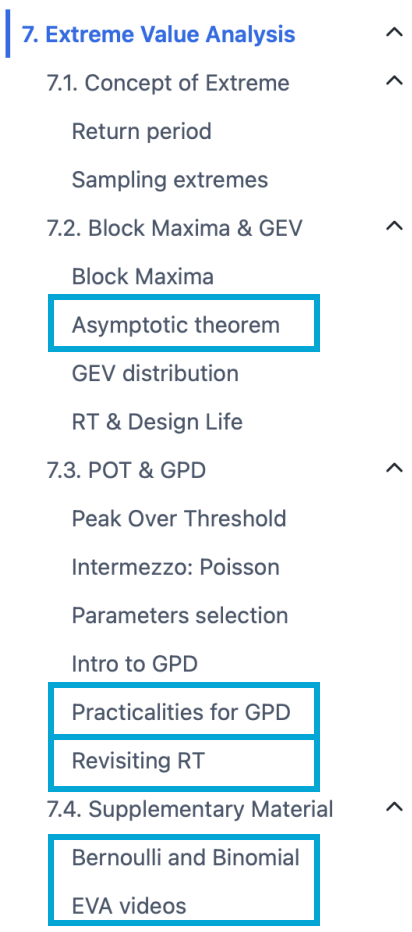
Design discharge: y + th = $299\frac{m^3}{s}$

Practicalities: what's next?

TU Delft

What's next?

- This lecture: basic theory and how to apply it.
- Much more theory and demonstrations in the textbook!
  - In 7.2, asymptotic model and domains of attraction
  - In 7.3, RT&DL based on Poisson distribution, theory behind application of the GPD
  - In 7.4, extra material: Bernoulli&Binomial and videos



7. Extreme Value Analysis

The preceding chapters have focused on a variety of data-driven and physics-based modelling techniques which we primarily used to interpolate between known data (e.g., machine learning), or make predictions about phenomena where randomness did not play a significant role (e.g., finite volume or finite element methods applied to simple physics problems). Most of the methods and problems considered ignored (or greatly simplified) the stochastic nature of the underlying processes. When uncertainty was considered explicitly, it focused primarily on error and epistemic types, for example, the inclusion of various types of noise in Time Series Analysis, or measurement precision in Observation Theory. In these cases, the focus was on applications that were governed by variations around a central value (as opposed to the tails of the distribution), which are often modelled sufficiently using a Gaussian distribution.

The chapter on Continuous Distributions introduced additional asymmetric parametric distributions, such as the Gumbel or Exponential, which are better able to represent the observations that have a small frequency in a data set (i.e., rare events). Regardless of their flexibility, these distributions are not possible to validate for cases where data simply does not exist, which is where concepts introduced in this chapter become useful.

Continuous parametric distributions are relatively simple models to apply, and allow one to make inferences of values of the modelled random variable that occur infrequently, or not at all, within the available observations (i.e., the data set) due to a key concept: the *tail of the distribution*. When extrapolating to values outside a set of observations, fitting of the parametric distribution to the tail is crucial to provide a reasonable extrapolation. Consider the following figure, which was covered earlier in this book:

TU Delft

What's next?

- Wednesday:
  - Extreme temperature
- Friday
  - Extreme precipitation
  - Case of the flood caused by a DANA in Valencia in October 2024

Figure source: Revista Ejercitos

And enjoy the journey!

Questions?

Please, leave the room through the door in the ground floor.

TU Delft