# Extreme Value Analysis

The statistical analysis of low frequency, high severity events

Alfred J. Reich, PhD

20 October 2022

# Extreme Value Analysis (EVA)

- EVA is a statistical methodology for making inferences about rare events (weather, finance, public health, materials, etc.)
  - It is also very often referred to as Extreme Value Theory (EVT)
- Disambiguation:
  - Extreme Value Theory (Analysis) has <u>nothing</u> to do with the Extreme Value Theorem, from elementary calculus.
- This talk will be limited to:
  - ”Classical” EVA (mostly)
  - Univariate, continuous probability distributions
  - Maxima, since $\min(X_1, X_2, \dots, X_n) = -\max(-X_1, -X_2, \dots, -X_n)$

# North Sea Flood of 1953

Losses:

- 1836 people killed
- 72,000 people evacuated
- 49,000 houses and farms flooded
- 201,000 cattle drowned
- 500 km coastal defenses destroyed
- More than 200,000 ha flooded

Effect on Study of Extreme Events:

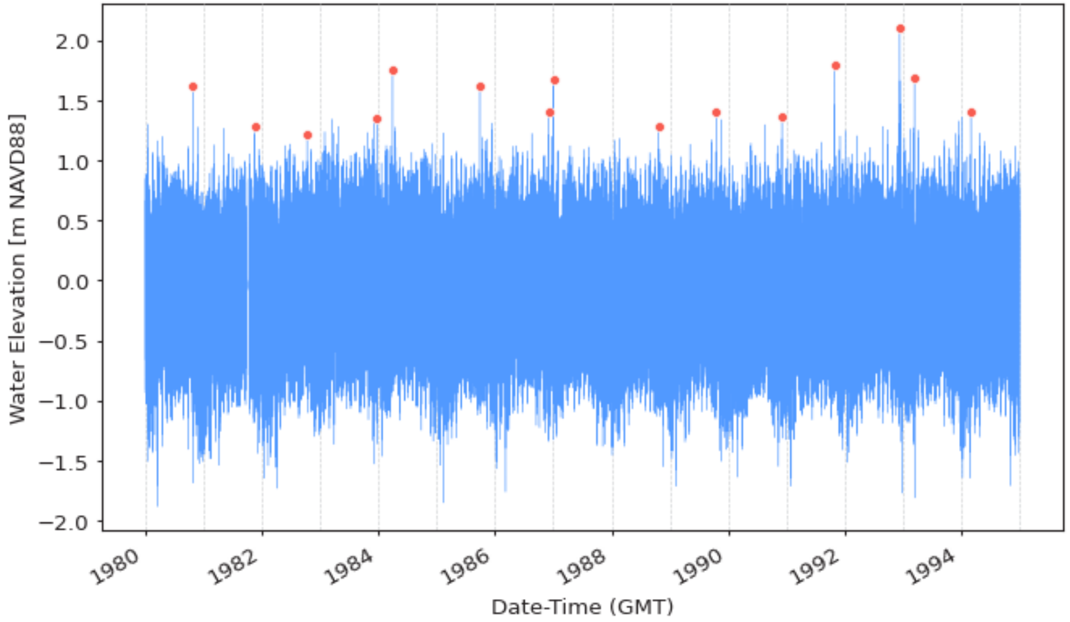
- Very little systematic statistical research w.r.t. height of the dikes was done before 1953
  - Flood of 1570 was mean-sea-level + 4m
- Gave EVA research a decisive push
- Needed height estimate **well outside range of existing data**
  - Van Dantzig report estimated $p=1-10^{-4}$ quantile (one-in-ten-thousand-year surge height) of mean-sea-level + 5.14m

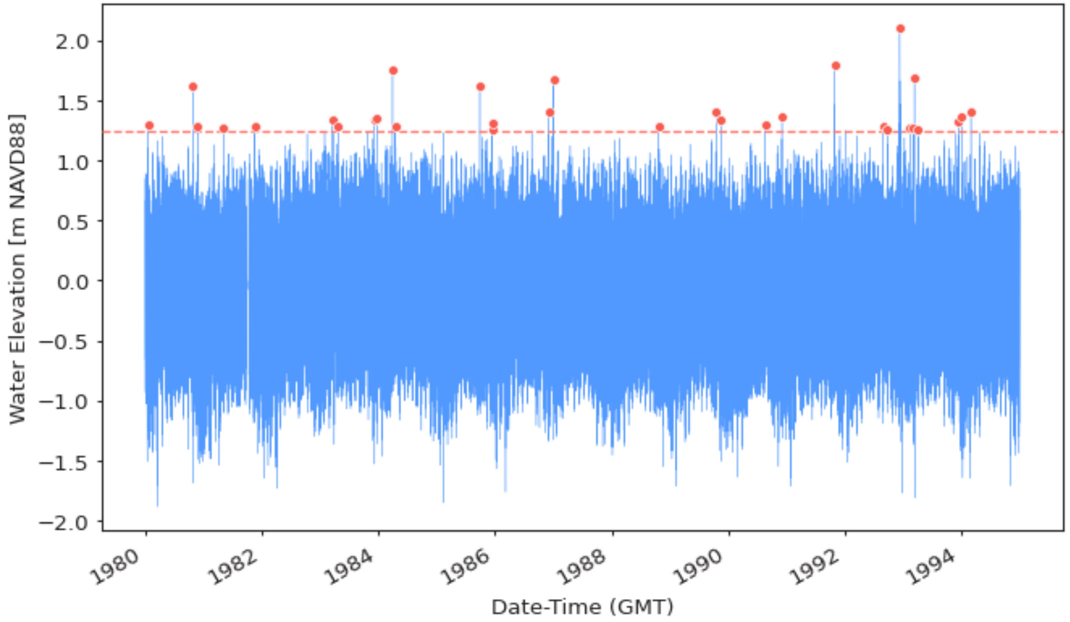
Source: [Embrechts 1997]

Netherlands, during 1953 North Sea Flood.
Viewed from a U.S. Army helicopter.

Source: https://en.wikipedia.org/wiki/North_Sea_flood_of_1953

# Two Primary Approaches to EVA




**Block Maxima (BM)**

Divide the data into large/long blocks and use the maximum/minimum value in each block

**Points Over Threshold (POT)**

Use all data that exceeds a specific threshold

Source: PyExtremes User Guide

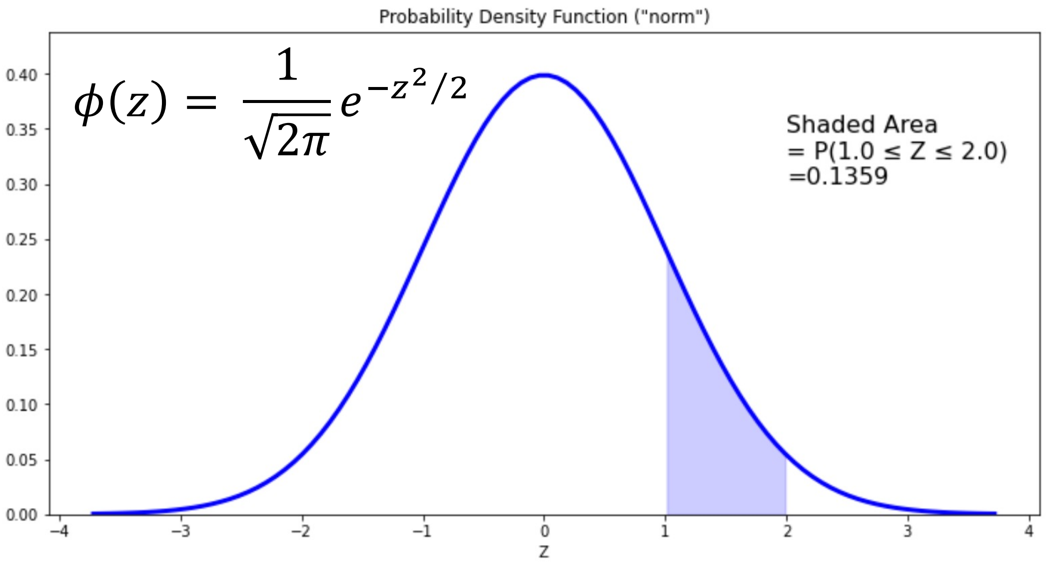
# A Very Brief Refresher

Probability Theory: PDFs, CDFs, CLT, etc.

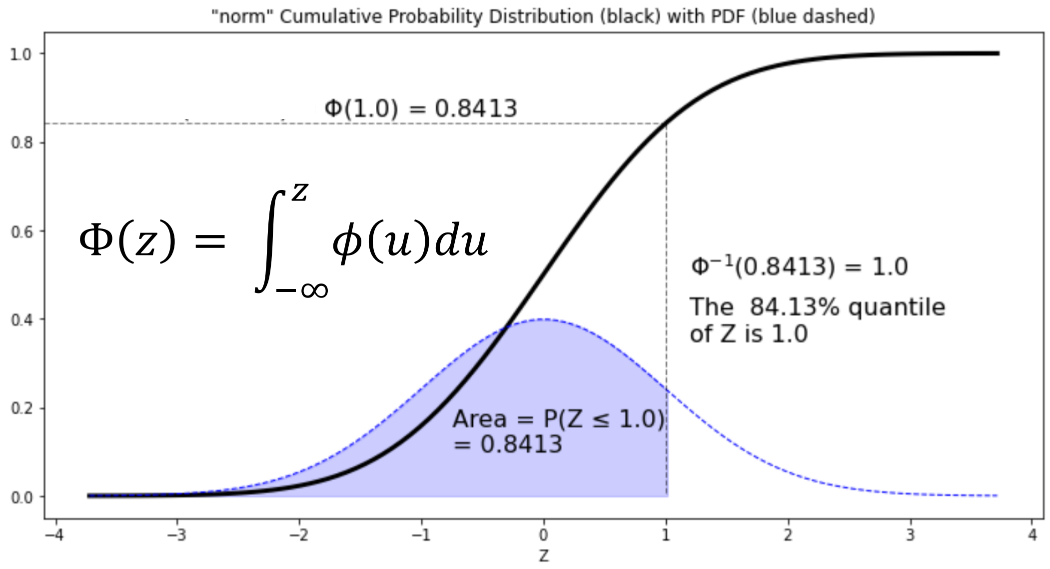
# Continuous PDFs and CDFs

Standard Normal Distribution

Probability Density Function (PDF)




Cumulative Distribution Function (CDF)



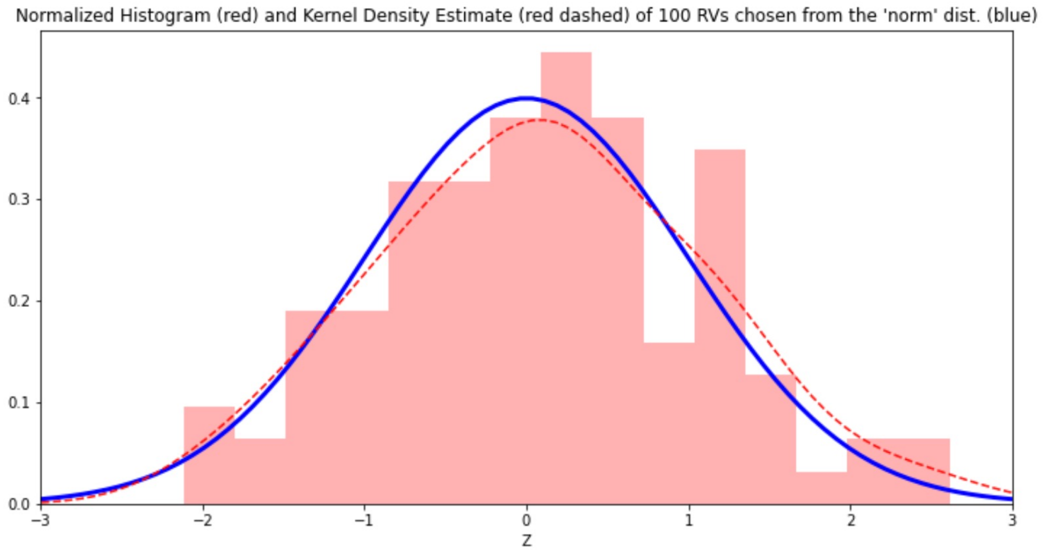

If $X \sim N(\mu, \sigma)$ and $Z = \frac{x-\mu}{\sigma}$,

then $F_X(x) = P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$, i.e., a location-scale family†

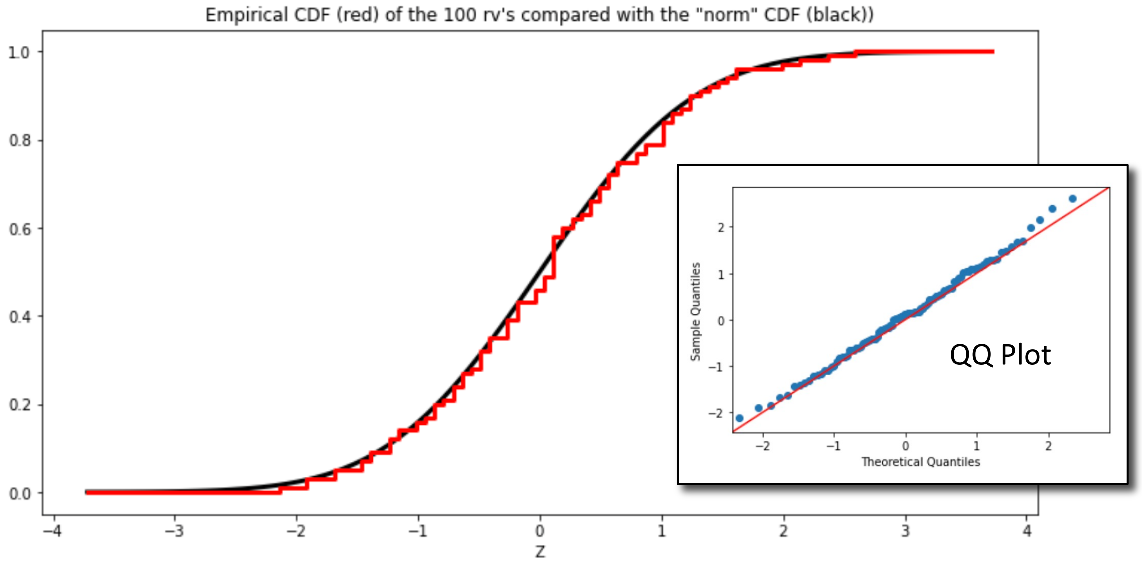
†A location-scale family is a family of distributions formed by translation and scaling of a *standard* family member.

# Histograms and Empirical CDFs

## Normalized Histogram & Kernel Density Estimate of a Random Sample




## Empirical CDF of the Same Random Sample




- Histograms (normalized) are empirical estimates of PDFs, but their shape is sensitive to bin size
- Kernel density estimates are another form of PDF estimate, but their shape is sensitive to the type of kernel used

- The Empirical CDF is less susceptible to subjective choices,
- so it is often used for model checking, for example, using Quantile-Quantile Plots.

# Parameter Estimation

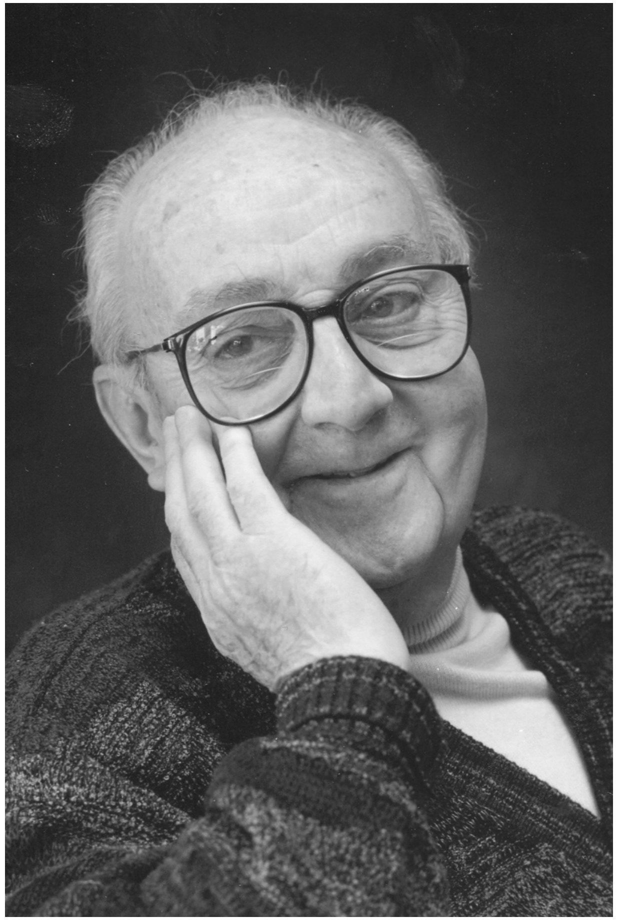
$X_1, \dots, X_n \sim F(x; \underline{\theta})$ iid and $f = F'$

- Maximum Likelihood Est. (MLE)
  - $\underline{x} = (x_1, \dots, x_n)^T$
  - $\mathcal{L}_n(\underline{\theta}; \underline{x}) = \prod_{i=1}^{n} f(x_i; \underline{\theta})$
  - $\underline{\hat{\theta}} = \underset{\underline{\theta}}{\operatorname{argmax}} \ln[\mathcal{L}_n(\underline{\theta}; \underline{x})]$
- Other Methods
  - MOM / PWM / L-Moments
  - Bayesian Parameter Estimation
  - …

- Confidence Intervals (CI)
  - Wald CI ("classical" method)
    - MLE: $\underline{\hat{\theta}} \dot{\sim} MVN_d(\theta, I^{-1}(\theta))$
  - Profile Likelihood CI
    - Likelihood ratio is asymptotically $\chi^2_{df}$
  - Bootstrapping (resampling w/ replacement)
  - Credible Interval / Highest Posterior Density (HPD) Interval/Region (Bayesian)
  - …

“All models are wrong, but some are useful”

-- George E. P. Box



# Maximum Values

An experiment using pseudo-random numbers, along with some theory

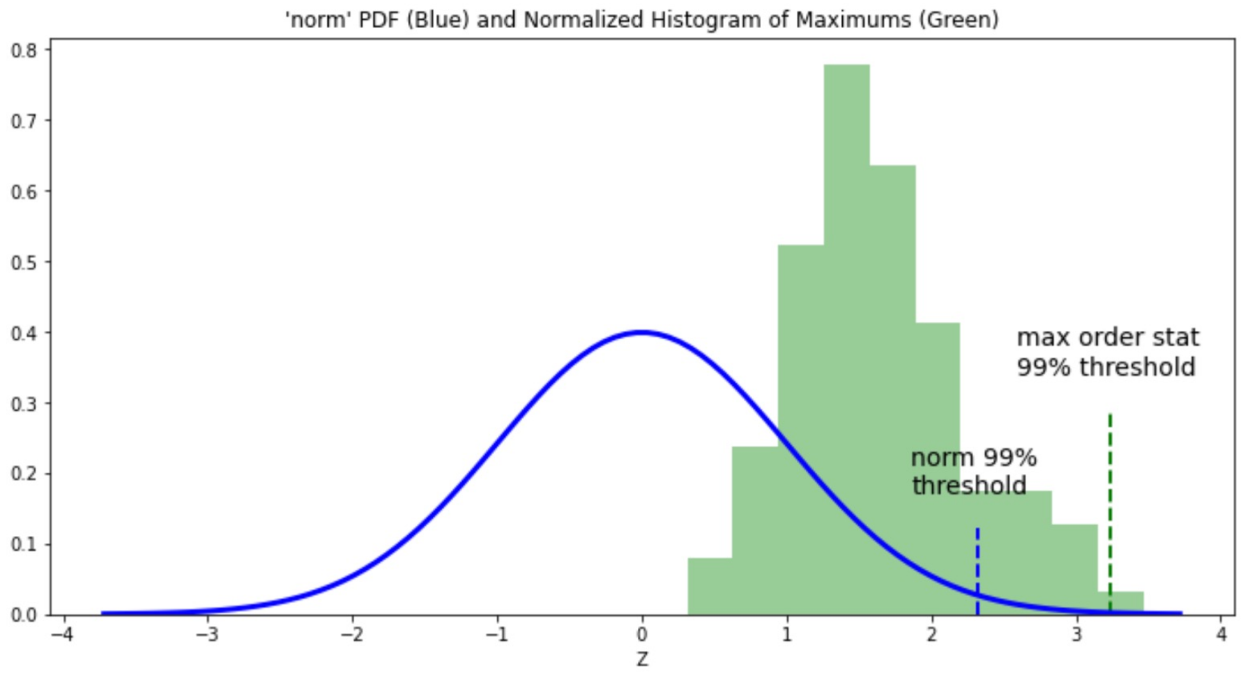
# Order Statistics

- Let $X_1, X_2, X_3, \ldots, X_n$ be iid RVs
  - with CDF: $F(x)$
  - and PDF: $f(x) = \frac{d}{dx} F(x)$
- Also, let $Y_1 \leq Y_2 \leq Y_3 \leq \ \ldots \leq Y_n$ be the $X$'s in ascending order
- The $Y$'s are *Order Statistics* based on the $X$'s
- We'll focus on the Maximum Order Statistic, $Y_n$

# A Random Sample of Maximum Order Statistics (based on Std. Normal Dist.)

The figure at right depicts:

- Standard Normal PDF (blue)
- A normalized histogram of 200 maximum order statistics (green)
  - Where each maximum came from a random sample of 12 standard normal RVs
- Two quantiles (“thresholds”):
  - Standard normal 99% quantile (~2.33)
  - Empirical 99% quantile of the 200 maxs (~3.24)
- Note that there is almost a full N(0,1) standard deviation between the quantiles.



# Distribution of the Maximum Order Statistic

The maximum order statistic …

- has CDF, $Y_n \sim G$, where $G(y) = [F(y)]^n$

$$
\begin{aligned}
G(y) &= P(Y_n \leq y) \\
&= P(X_1 \leq y, \ldots, X_n \leq y) \\
&= P(X_1 \leq y) \ldots P(X_n \leq y) \\
&= [F(y)]^n
\end{aligned}
$$

- The PDF is $g(y) = n[F(y)]^{n-1} f(y)$
- Note: If $F(x) < 1$, then $F^n(x) \to 0$, as $n \to \infty$

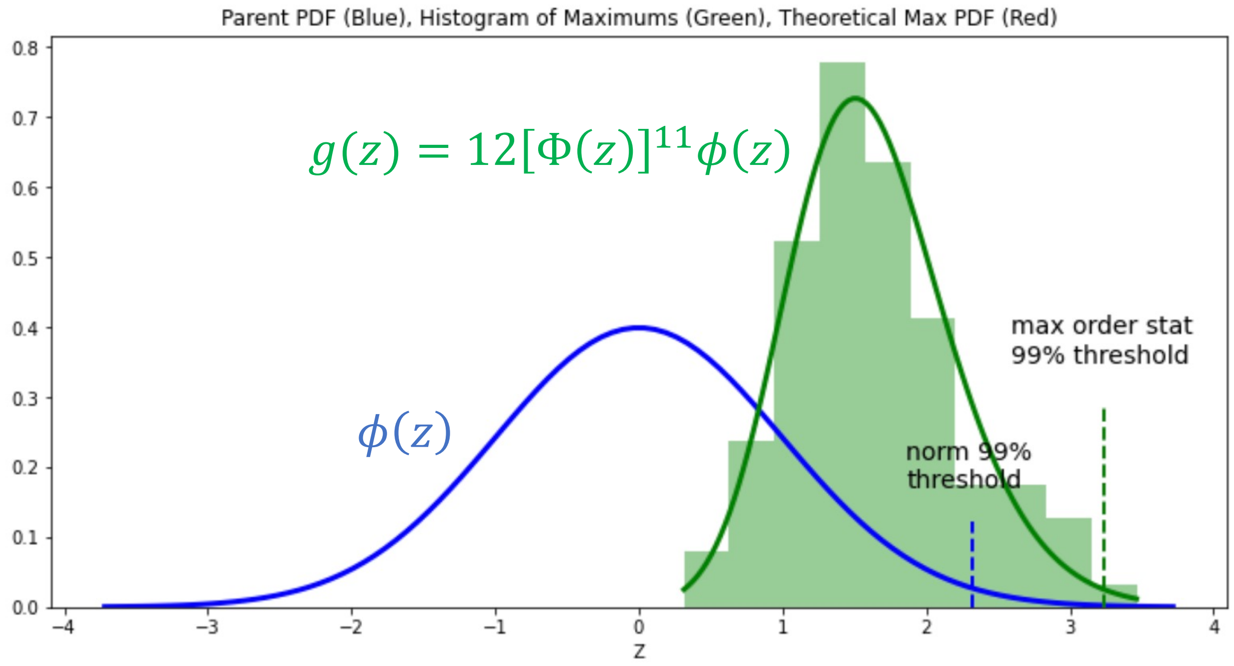
# Max. Order Stat. Distribution (based on Std. Normal Dist.)

If $X_1, \ldots, X_n \sim N(0,1)$ iid and $Y_n$ is the Maximum Order Statistic, then its CDF and PDF are as follows, resp.:

$$G(y) = [\Phi(y)]^n$$

$$g(y) = n[\Phi(y)]^{n-1}\phi(y)$$

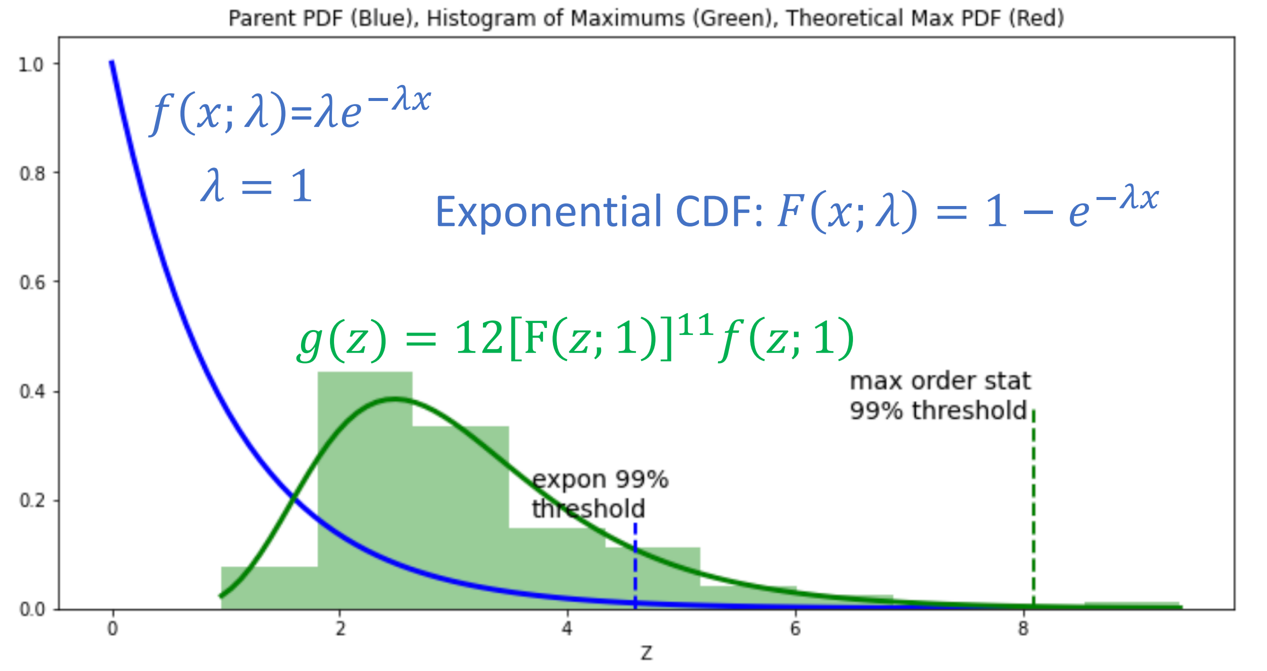
At right, the PDF, $g$, is plotted (in green) along with the histogram of maximums from 200 samples, each of size $n = 12$.



# Max. Order Stat. Distribution (based on Exponential Dist.)

The figure at right depicts:

- Exponential PDF, $\lambda = 1$ (blue)
- A histogram of 200 maximum order statistics (green)
    - Where each maximum came from a random sample (iid) of 12 exponential RVs ($\lambda = 1$)
- Two quantiles ("thresholds"):
    - Exponential 99% quantile (~4.6)
    - Empirical 99% quantile of the 200 maxs (~8.1)



# Extreme Value Theory (EVT)

The Generalized Extreme Value (GEV) Distribution and Theorem

# Generalized Extreme Value (GEV) Distribution

CDF

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

defined on the set $\left\{z: 1+\frac{\xi(z-\mu)}{\sigma}>0\right\}$,

where $-\infty<\mu<\infty$, $\sigma>0$ and $-\infty<\xi<\infty$.

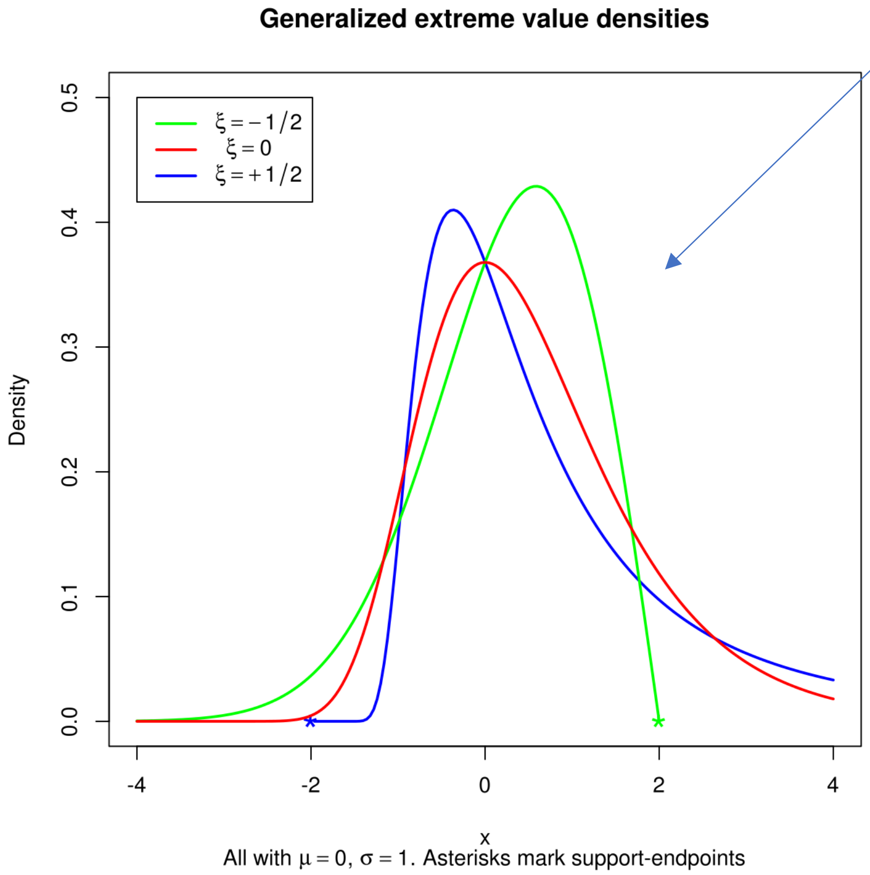
- If $\xi>0$, then $G$ is the **Fréchet** dist. (heavy tailed)
- If $\xi<0$, then $G$ is the **Weibull** dist. (upper-bounded)
- Taking the limit as $\xi\to 0$, obtains the **Gumbel** dist. (light-tailed)

$$G(z) = exp\left[-exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right], -\infty<z<\infty$$

Source: [Coles 2001]

*NOTE: scipy.stats reverses the sign of $\xi$. ($c=-\xi$)*

PDFs




Source: [Wikipedia: GEV]

# Extreme Value Theorem

Let $Y_n$ be the maximum order statistic of $X_1, \dots, X_n$, a sequence of iid random variables with common CDF, $F$, then if there exists sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\left(\frac{Y_n - b_n}{a_n} \leq z\right) \to G(z) \text{ as } n \to \infty$$

for a non-degenerate distribution function $G$,

then $G$ is a member of the GEV family (described on the previous slide).

# GEV is a Location-Scale Family of Distributions

**$P(Y_n \leq w)$ is approximated by another member, $G^*$, of the GEV location-scale family.**

To see this, observe that the GEV distribution $G$ is a family of distributions formed by translation and scaling of a standard family member.

The standard member is $H(x;\xi) = \exp\left[-(1+\xi x)^{-1/\xi}\right]$, for $\xi \neq 0$,

and so, $G(z) = H\left(\frac{z-\mu}{\sigma};\xi\right)$, for $\xi \neq 0$ and $\left\{z\colon 1 + \frac{\xi(z-\mu)}{\sigma} > 0\right\}$

For $\xi = 0$, define $H(x;0) = \exp[-\exp(-x)]$

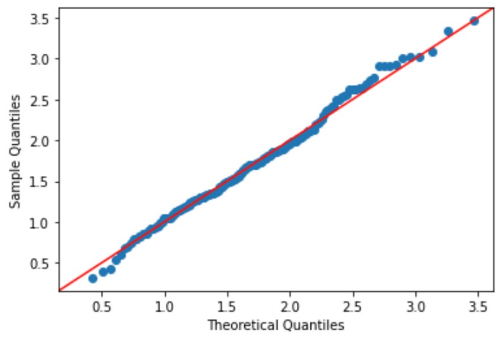
Furthermore, $P\left(\frac{Y_n-b_n}{a_n} \leq z\right) \approx H\left(\frac{z-\mu}{\sigma};\xi\right)$ for large $n$, which, after some algebraic manipulation, can be written as

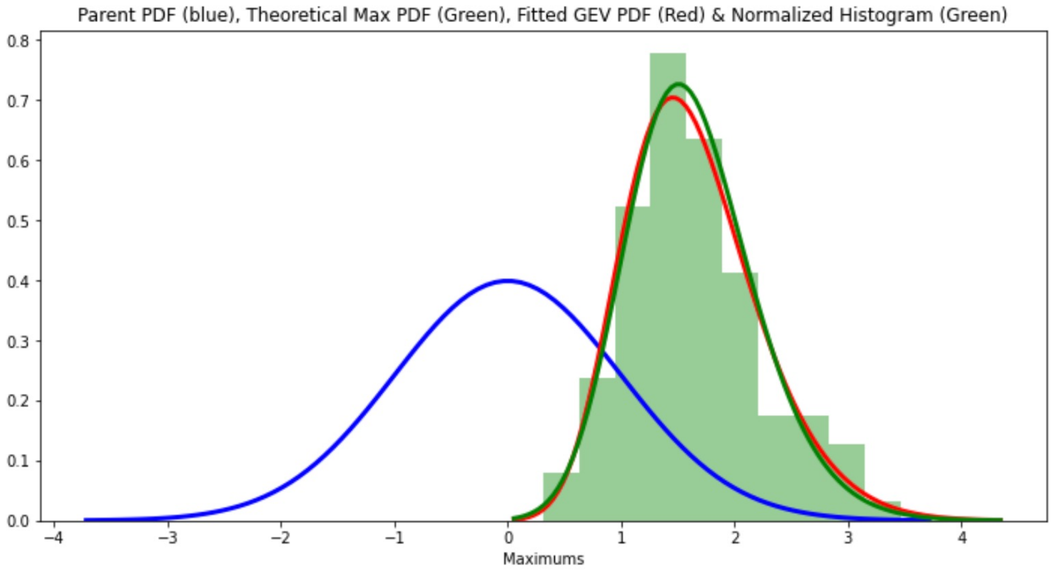
$$P(Y_n \leq w) \approx H\left(\frac{w-b_n^*}{a_n^*};\xi\right) = G^*(w)$$

using different location and scale values, $a_n^*$ and $b_n^*$.

# GEV MLE Fit to Maximums based on Std. Normal RVs

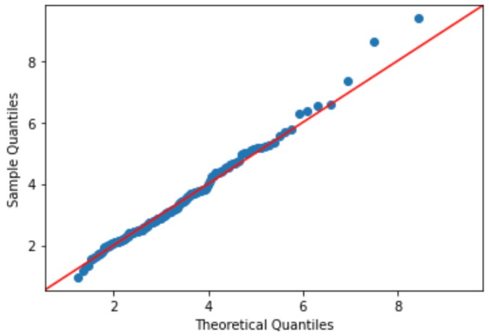
- Same as earlier normal plot, except that now the GEV fit (via MLE) is also shown (red)
- The QQ-plot shows the normal-based maximums (data) vs. the GEV fit

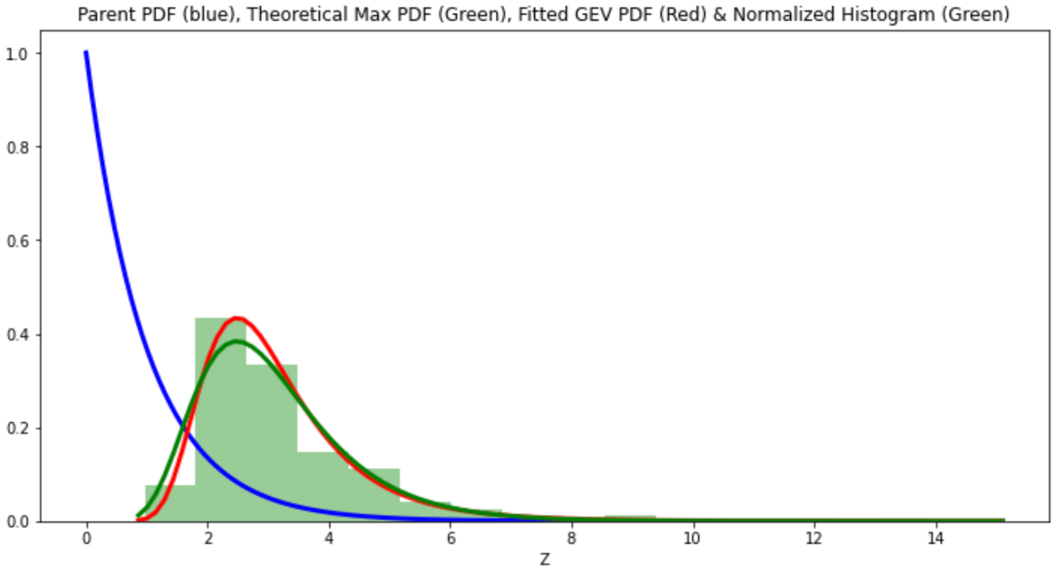


# GEV MLE Fit to Maximums based on Exponential RVs

- Same as earlier exponential plot, except that now the GEV fit (via MLE) is also shown (red)
- The QQ-plot shows the exponential-based maximums (data) vs. the GEV fit



# Return Levels & Return Periods

- The quantiles of the GEV can be interpreted as **return levels**
- A **return level** is the value expected to be exceeded on average once every 1/p periods, where 1 – p is the specific probability associated with the quantile $G\left(z_p\right) = 1 - p$

$$\Longrightarrow z_p = \mu - \frac{\sigma}{\xi}\left\{1 - [-\log(1-p)]^{-\xi}\right\}, \xi \neq 0$$
$$z_p = \mu - \sigma \log\{-\log(1-p)\}, \xi = 0$$

- $z_p$ is the **return level** associated with the **return period** of ${}^{1}/_{p}$

# EVA Software

- R:
  - **ismev**: https://cran.r-project.org/web/packages/ismev/index.html
  - **extRemes**: https://cran.r-project.org/web/packages/extRemes/index.html
  - and many, many more… see the following…
    - https://cran.r-project.org/web/views/ExtremeValue.html (Many links to other EVA packages in R)
    - *"A modeler's guide to extreme value software"*, Belzile, et al., arXiv:2205.07714v1, 16 May 2022
- Python:
  - **Pyextremes**: https://georgebv.github.io/pyextremes/
  - Scikit-extremes: https://kikocorreoso.github.io/scikit-extremes/
  - Wafo: https://pypi.org/project/wafo/
- Documentation
  - Both extRemes (R) and Pyextremes (Python) have excellent documentation
    - Not that other packages don't, but the docs for these two packages make good starting points for learning more about EVA.
  - W.r.t. a Good Book, I recommend Coles' book:
    - "An Introduction to Statistical Modeling of Extreme Values"
    - …that is, *ISMEV*

# Example 1

GEV Fit using ISMEV (R)

# Example 1: ismev (R)

```
> library(ismev)
Loading required package: mgcv
Loading required package: nlme
This is mgcv 1.8-40. For overview type 'help("mgcv-package")'.
> data(portpirie)
> plot(portpirie)
> ppfit <- gev.fit(portpirie[,2])
$conv
[1] 0

$nllh
[1] -4.339058

$mle
[1]  3.87474692  0.19804120 -0.05008773

$se
[1] 0.02793211 0.02024610 0.09825633

>
```

Information, derived during the MLE fit, is stored here.

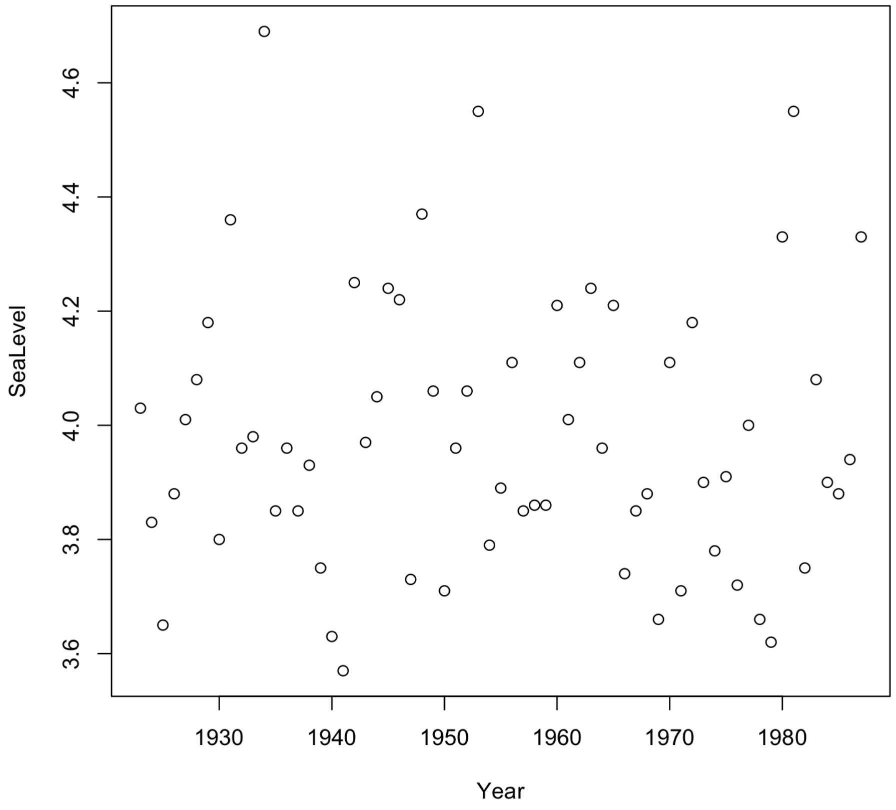
$\mu$ $\sigma$ $\xi$

**95% CI**

**-0.0500 ± 1.96 * 0.0983**

$-0.243 < \xi < 0.142$

Annual Maximum Sea-Levels at Port Pirie, South Australia



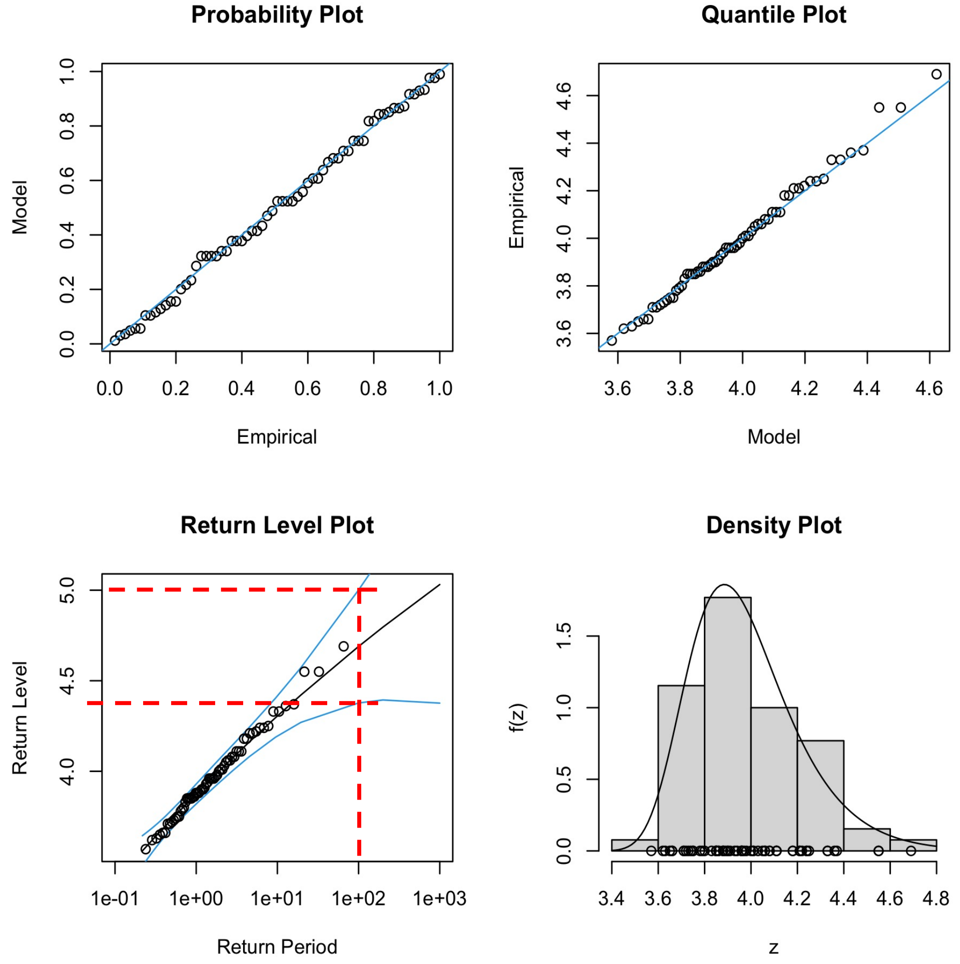
# Example 1: Diagnostics

R function call:

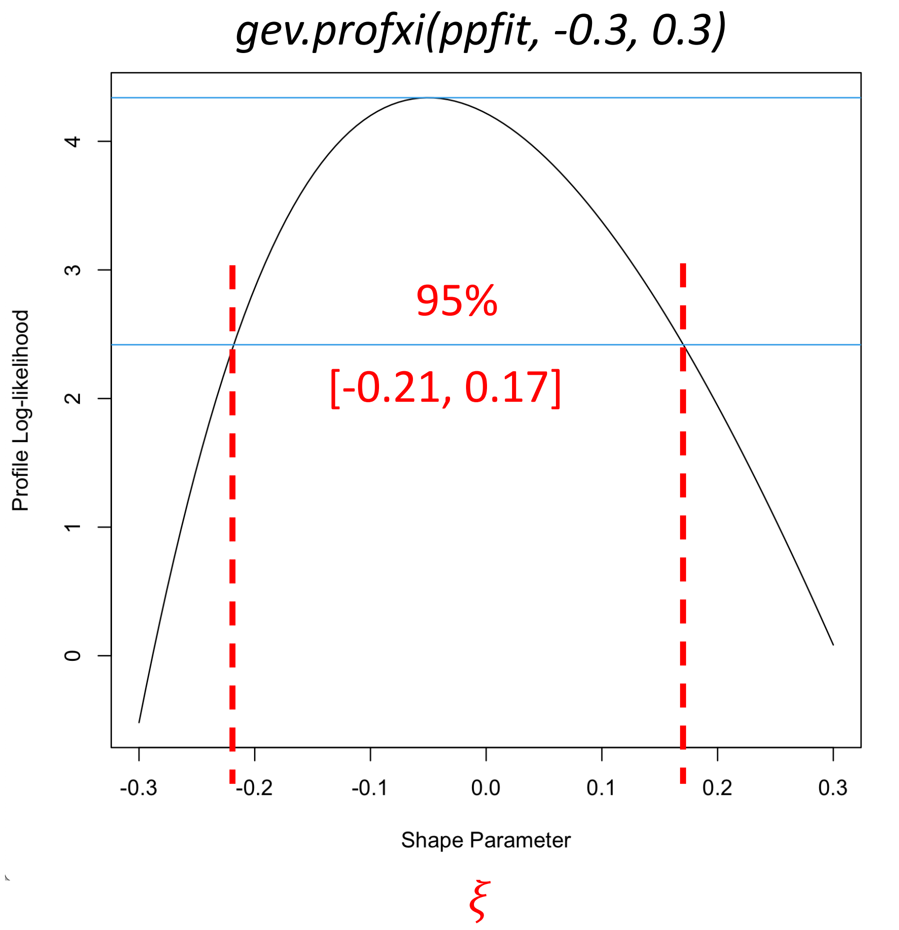
*gev.diag(ppfit)*

Return Levels (& Wald 95% CI)

- 10 year: $\hat{z}_{0.1} = 4.30$, [4.19, 4.41]
- 100 year: $\hat{z}_{0.01} = 4.69$, [4.38, 5.00]



# Example 1: Profile Likelihood




- Confidence intervals produced using the profile likelihood method are derived from the asymptotic Chi-Square distribution of the likelihood ratio.
- They are “better” for asymmetric, sparse datasets, like those encountered in EVA.

Recall from the first slide of this example, we had:

$$-0.243 < \xi < 0.142$$

# Peaks Over Threshold (POT)

and the Generalized Pareto Distribution (GPD)

# Peaks Over Threshold (POT) & the Generalized Pareto Distribution (GPD)

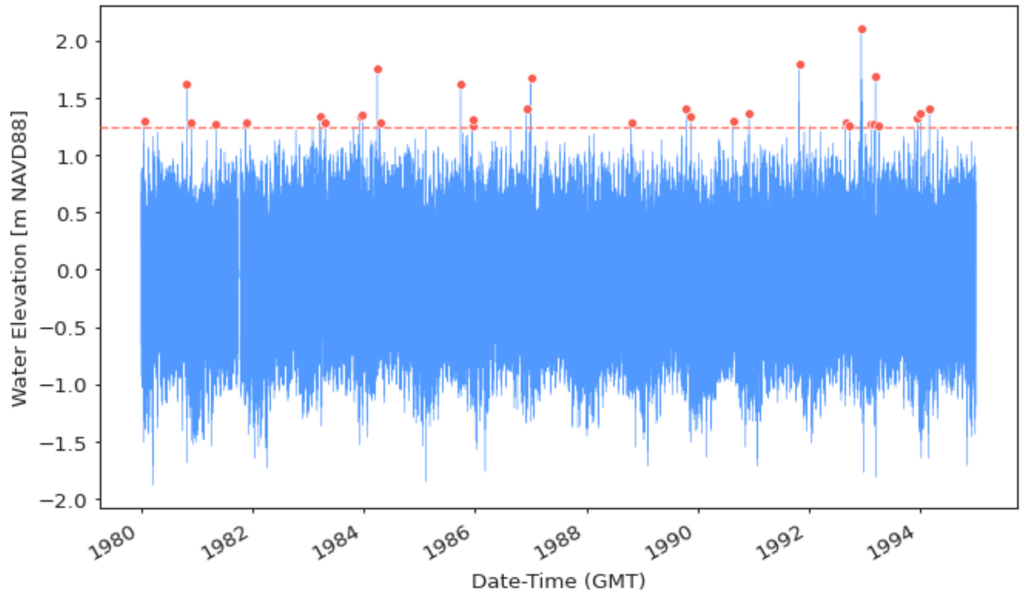
- Let $X_1, \dots, X_n$ iid $\sim F$
- Define *extreme events* as those $X_i$'s that exceed some high threshold, $u$.
- If $F$ is known, then the distribution of threshold **exceedances** is:

$$P(X > u + y \mid X > u) = \frac{1 - F(u + y)}{1 - F(u)}$$

- Otherwise, if $Y_n = \max(X_1, \dots, X_n)$ and $Y_n \dot{\sim} GEV(x; \mu, \sigma, \xi)$ then, for large enough $u$, the distribution of exceedances is approximately the **Generalized Pareto Distribution:**

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

where $\{y: y > 0 \; and \; (1 + \xi y/\tilde{\sigma}) > 0\}$
and $\tilde{\sigma} = \sigma + \xi(u - \mu)$

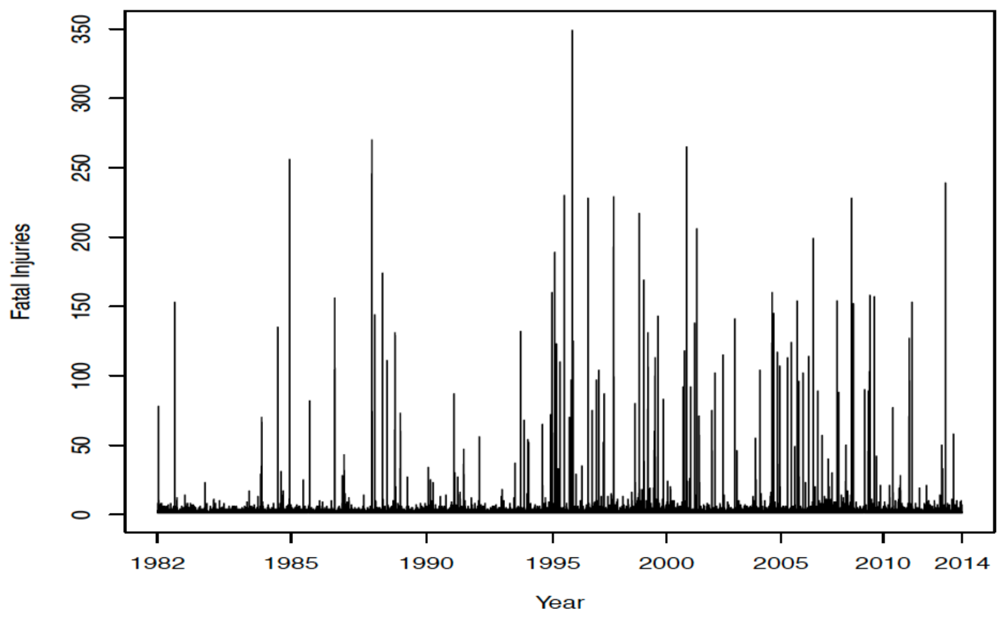



But now there's a problem…
How do we choose the threshold?

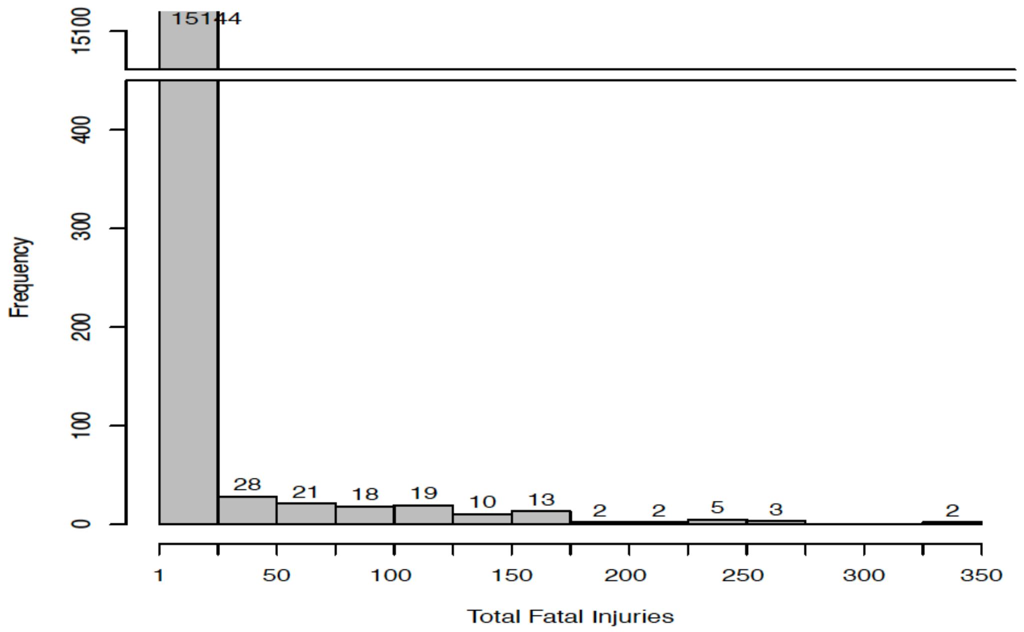
# Choosing a Threshold for POT: Example

## Fatal Injuries from Aviation Accidents

- "Quantification of the large accidents which have far reaching effect (fatality) would provide objective guidance in long-term planning and response for manufacturers, insurers and re-insurers." [Das 2016]




Numbers of fatal injuries from aviation accidents, 1982 - 2014




Histogram of fatal injuries from aviation accidents

Source: [Das 2016]

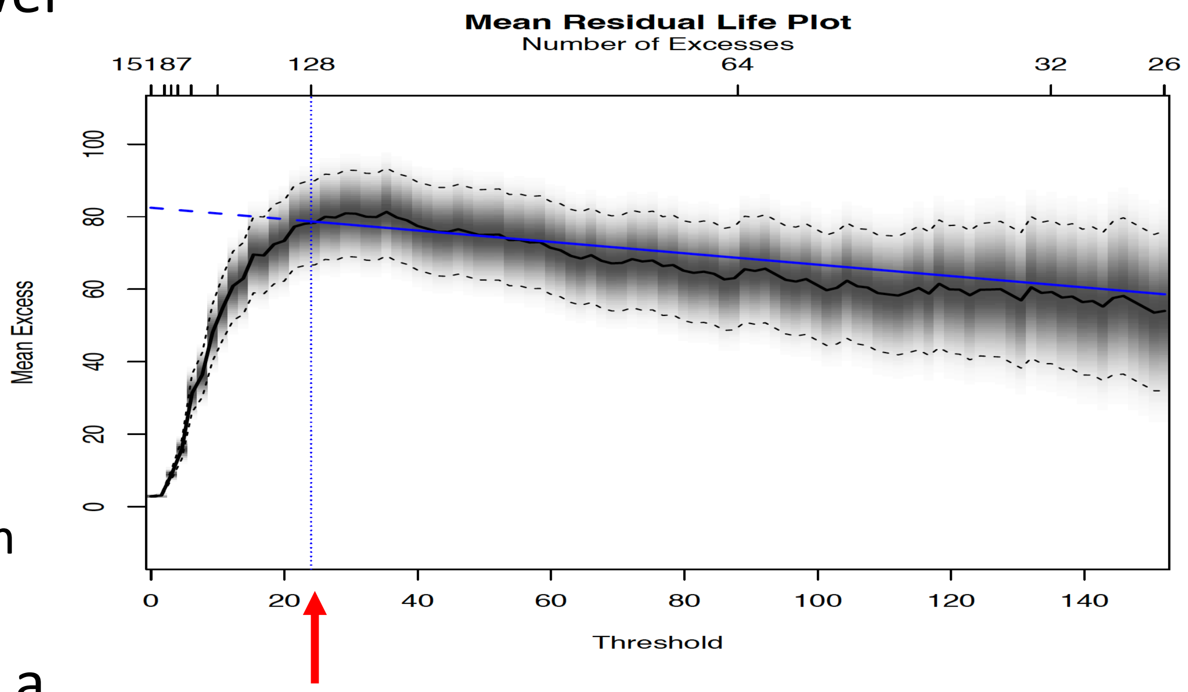
# Choosing a Threshold for POT: MRL

- If the tail data follow a GPD with lower bound of u, then the Mean Residual Life (MRL) plot should be approx. linear for values above u.

$$mrl(x) = E(X - x \mid X > x)$$

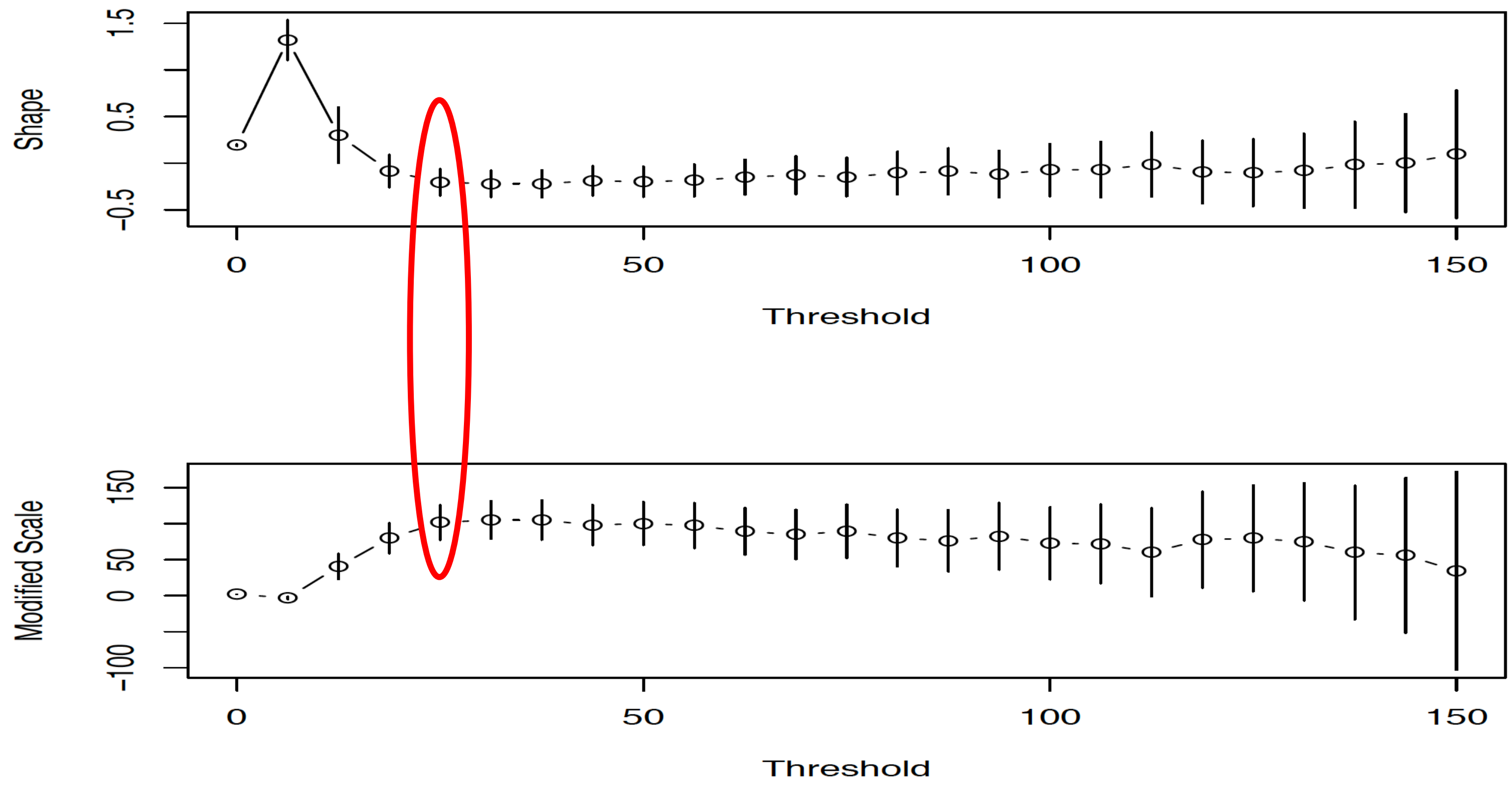
  - MRL is also sometimes called the Mean Excess Function

- So, select the smallest u which gives a linear MRL plot.




Source: [Das 2016]

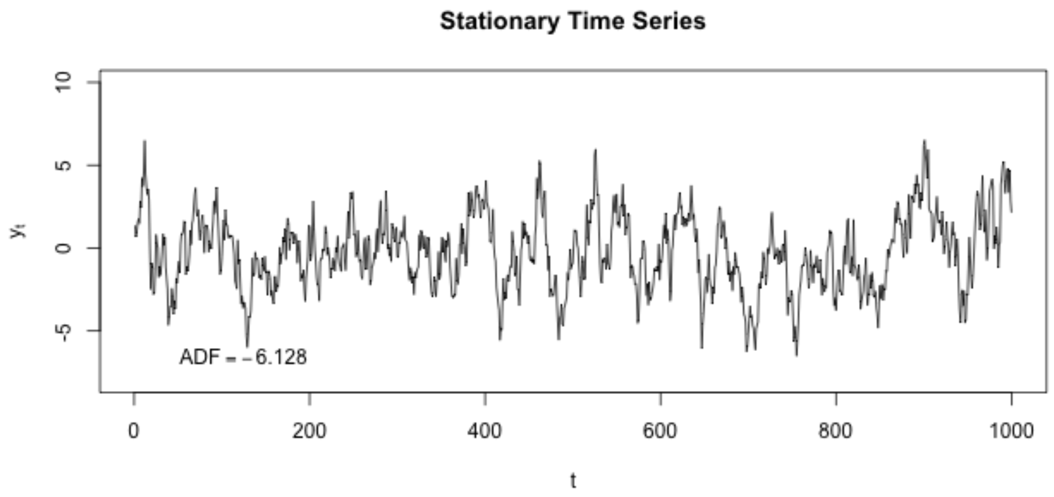
# Choosing a Threshold for POT: Parameter Stability

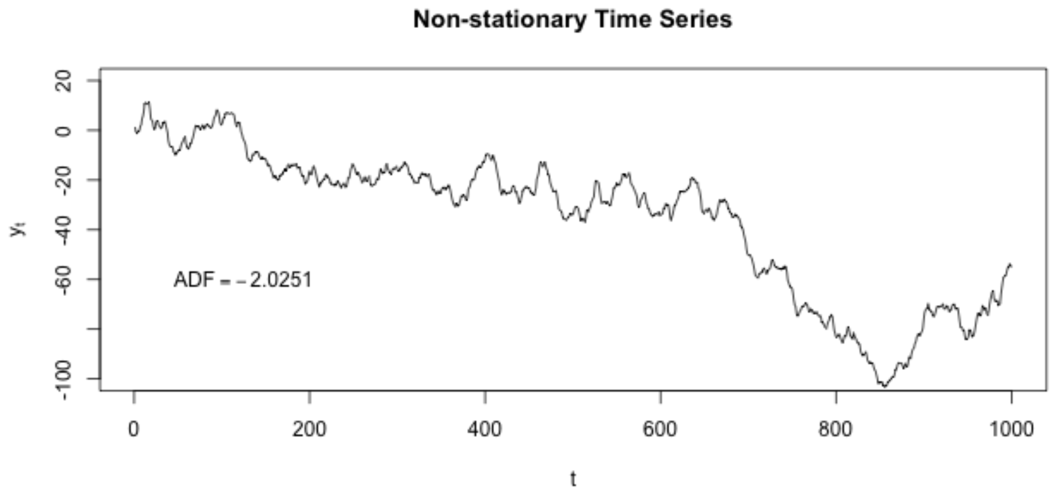


# Stationarity & Non-Stationarity

# Stationarity

- $X_1, X_2, \ldots$ is a **stationary** random process if for any set of integers $\{i_1, \ldots, i_k\}$ and any integer $m$, the joint distributions of $(X_{i_1}, \ldots, X_{i_k})$ and $\left(X_{i_1+m}, \ldots, X_{i_k+m}\right)$ are identical.




https://en.wikipedia.org/wiki/Stationary_process

# Dealing with Non-Stationarity

- Data often contains trends and seasonal cycles (financial, weather)
- Using BM with annual maximums can avoid seasonal cycles (weather)
- Trends and cycles can be removed via regression or time-series modeling.

$$\mu(t) = \mu_0 + \mu_1 t + \mu_2 t^2,$$

$$\sigma(t) = \sigma_0 + \sigma_1 t,$$

$$\xi(t) = \begin{cases} \xi_0, & t \leq t_0, \\ \xi_1, & t > t_0. \end{cases}$$

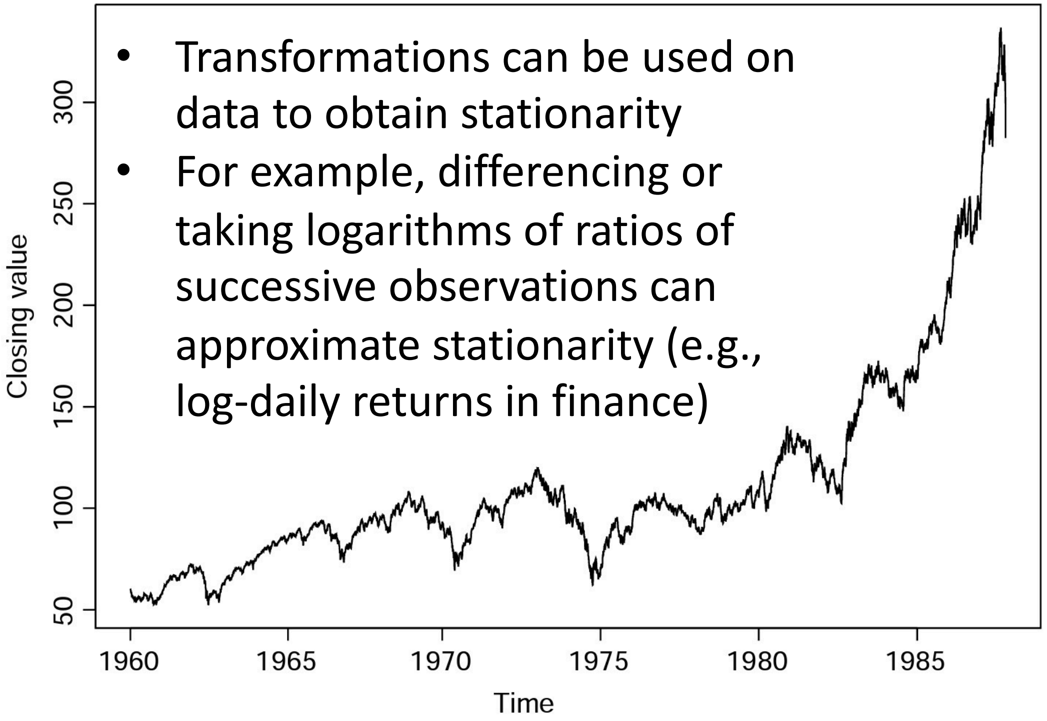
Example from *extRemes* [Gilliland 2016]

# Dealing with Non-Stationarity (cont.)

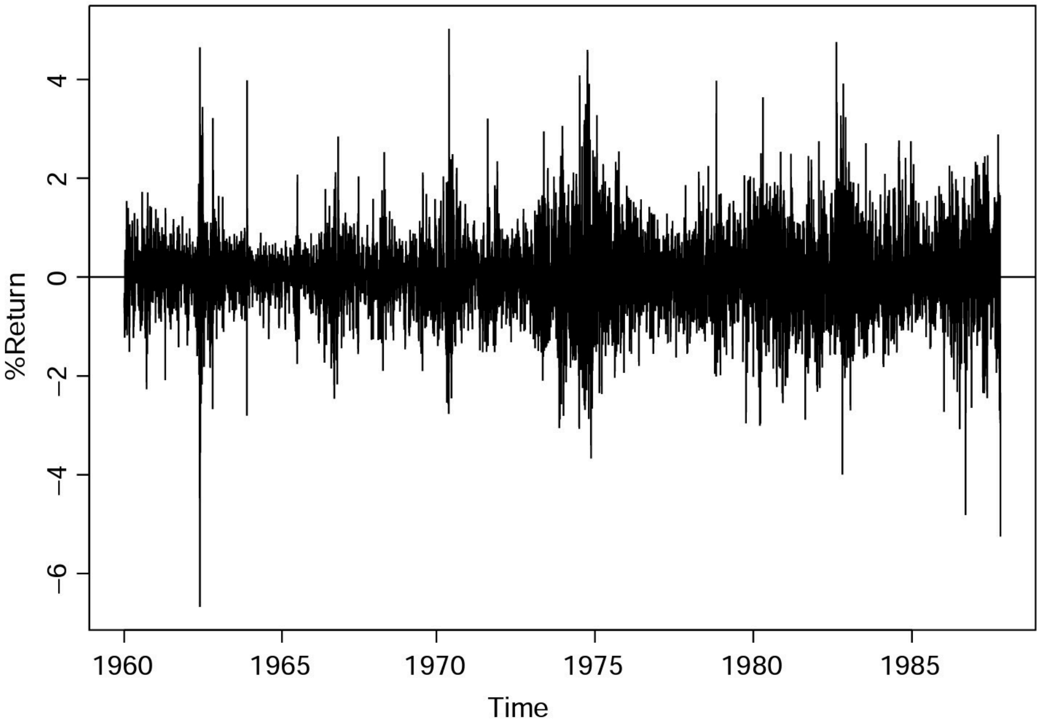
Financial Application: Estimate VaR (Value-at-Risk) for a given portfolio

S&P 500 Closing Values

- Transformations can be used on data to obtain stationarity
- For example, differencing or taking logarithms of ratios of successive observations can approximate stationarity (e.g., log-daily returns in finance)




Daily % Returns




Source: [Beirlant 2004]

# References


[Beirlant 2004] J. Beirlant, et al., “Statistics of Extremes: Theory and Applications” (2004)

[Belzile 2022] Belzile, et al., “A modeler’s guide to extreme value software”, arXiv:2205.07714v1 (2022)

[Coles 2001] S. Coles, “An Introduction to Statistical Modeling of Extreme Values”, Springer (2001)

[Das 2016] K. Das & A. Dey, “Analyzing fatal accidents in aviation using extreme value theory” (2016)

[Einmahl 2019] J. J. Einmahl, et al., “Limits to Human Life Span Through Extreme Value Theory”, JASA, 114:527, 1075-1080 (2019)

[Embrechts 1997] Embrechts, Paul, et al. “Modelling Extremal Events for Insurance and Finance”, Springer (1997)

[Gilleland 2016] E. Gilleland, R.W. Katz “extRemes 2.0: An Extreme Value Analysis Package in R”, J. of Statistical Software, 72(8), 1-39 (2016)

[Gumbel 1958] E.J. Gumbel, “Statistics of Extremes”, Columbia University Press, New York (1958)

[Robeson 2015] Robeson, S. M., “Revisiting the recent California drought as an extreme value”, Geophys. Res. Lett., 42, 6771–6779 (2015)

[Tsiftsi 2018] Tsiftsi, T., & De la Luz, V., “Extreme value analysis of solar flare events”, Space Weather, 16, 1984–1996 (2018)

[Wikipedia GEV] “Generalized extreme value distribution”

# Backup Slides

# Central Limit Theorem[†]

- Let $X_1, X_2, X_3, \ldots, X_n$ be independent & identically distributed (iid) random variables (RVs)
- from a distribution that has mean $\mu$ and positive variance $\sigma^2$,
- and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

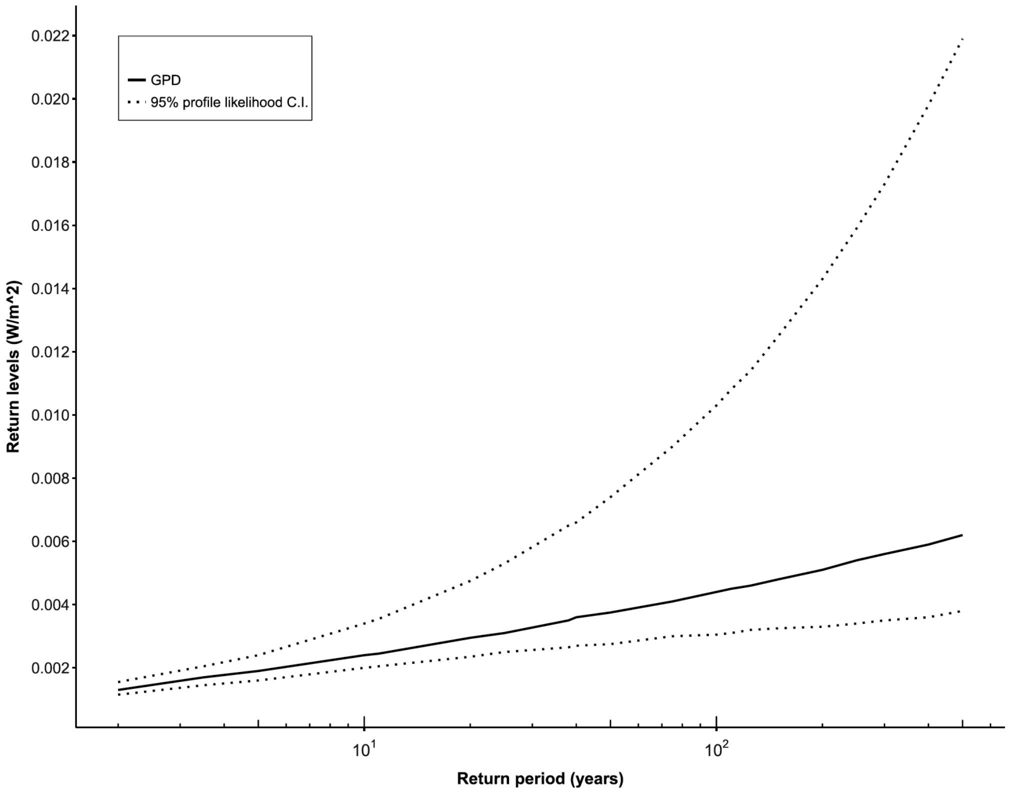
[†] This is a limited form of the CLT; other variants impose fewer conditions.

# Case Studies (brief)

Solar Flares, California Droughts, and Human Life Span

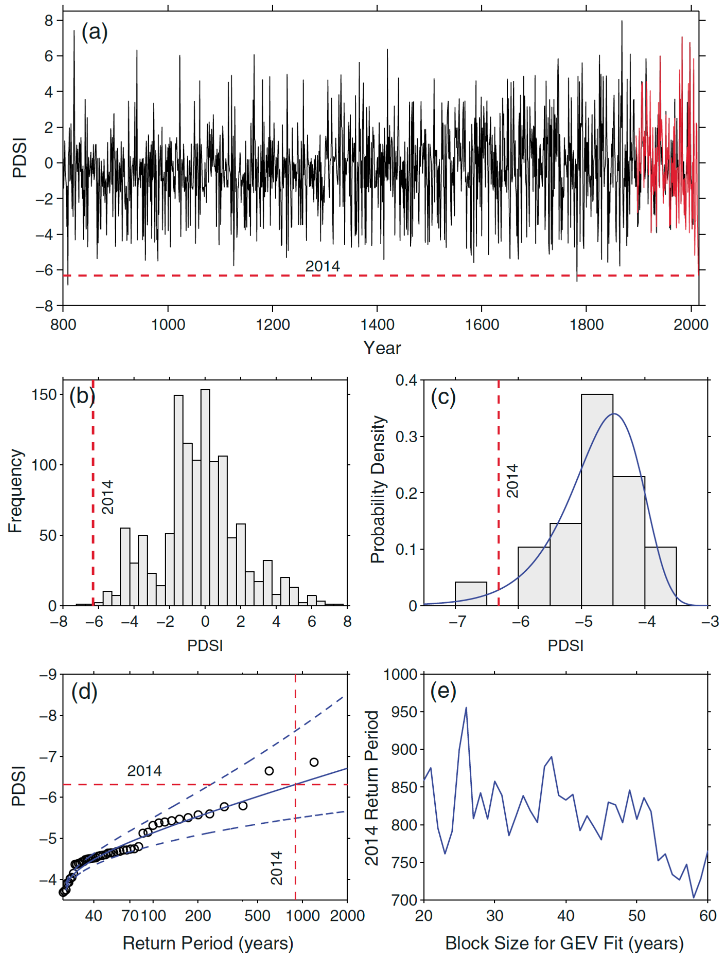
# “EVA of Solar Flare Events” [Tsiftsi 2018]

- 1859 – the “Carrington Event” (X45), most intense geomagnetic storm in recorded history
  - Est. cost to U.S. of similar event: **$670 billion to $2.9 trillion (~ 3.6% – 15.5% annual GDP)**
    - https://en.wikipedia.org/wiki/Carrington_Event#Similar_events
  - **Return Period: 110 years**,
    - with profile likelihood CI ~ (20, 6500) years.
  - Probability of a Carrington-like event happening in the next decade is 9%
- 2003 – “Halloween solar storms” (X35) generated largest solar flare ever recorded by GOES
  - **Return Period: 38 years,**
    - with profile likelihood CI ~ (10, 300) years.
  - A Halloween-like event is expected in the next decade with probability 23.8%



# California Droughts [Robeson 2015]

- The 1-year 2014 drought was most severe in the 1895–2014 record
  - Has a return period of 140–180 years,
  - however, *quantile mapping* produces return periods of 700–900 years
- Cumulative 3- and 4-year droughts are estimated to be much more severe
  - 2012–2014 drought is nearly a 10,000-year event
  - 2012–2015 drought has an almost incalculable return period and is completely without precedent




PDSI – Palmer Drought Severity Index

# Limits to Human Life Span [Einmahl 2019]

- Used EVA to consider whether the human life span is bounded:
- 30 years of data from Dutch residents
- The estimated extreme value indices ($\xi$), exhibited in Figure 2, at right, are all negative, hinting at a finite upper endpoint, that is, a **finite maximum life span**.

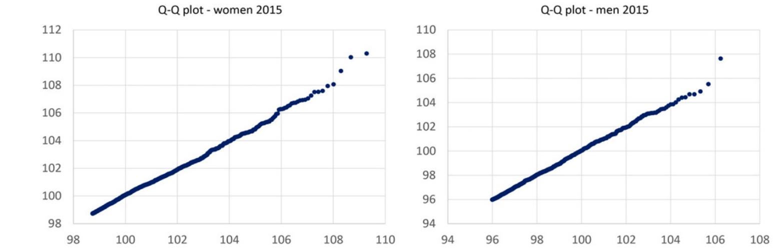



Figure 1. Generalized Pareto Q–Q plots for women and men for the year 2015.

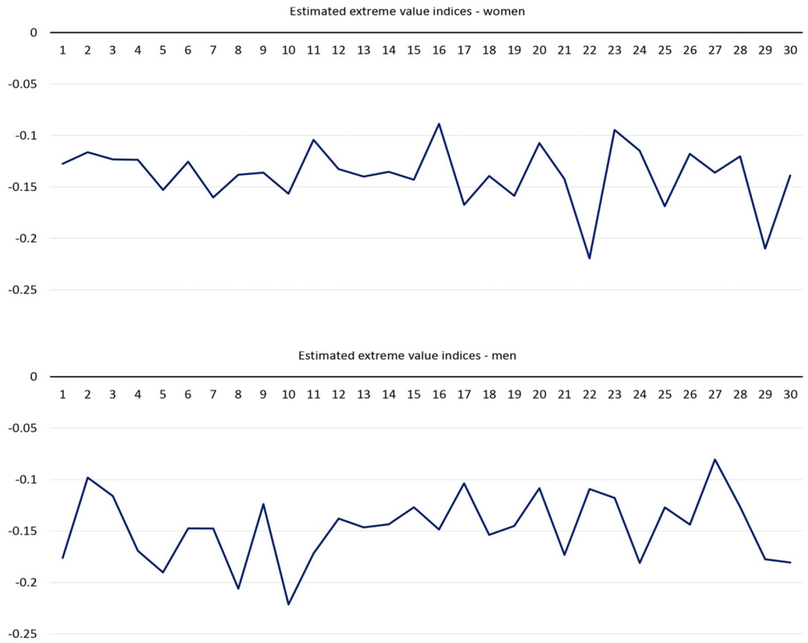



Figure 2. Estimated extreme values indices for the years of death $1985 + j, j = 1, \ldots, 30$.

# Example 2

GEV Fit using PyExtremes (Python)

# Example 2: pyextremes (Python)

- Raw Data (.csv) read & ”cleaned”
  - Sorted in ascending order
  - NaN entries removed
  - Converted to Pandas.Series
    - https://pandas.pydata.org/
  - Trend removed (+2.87 mm/yr)

> from pyextremes import EVA

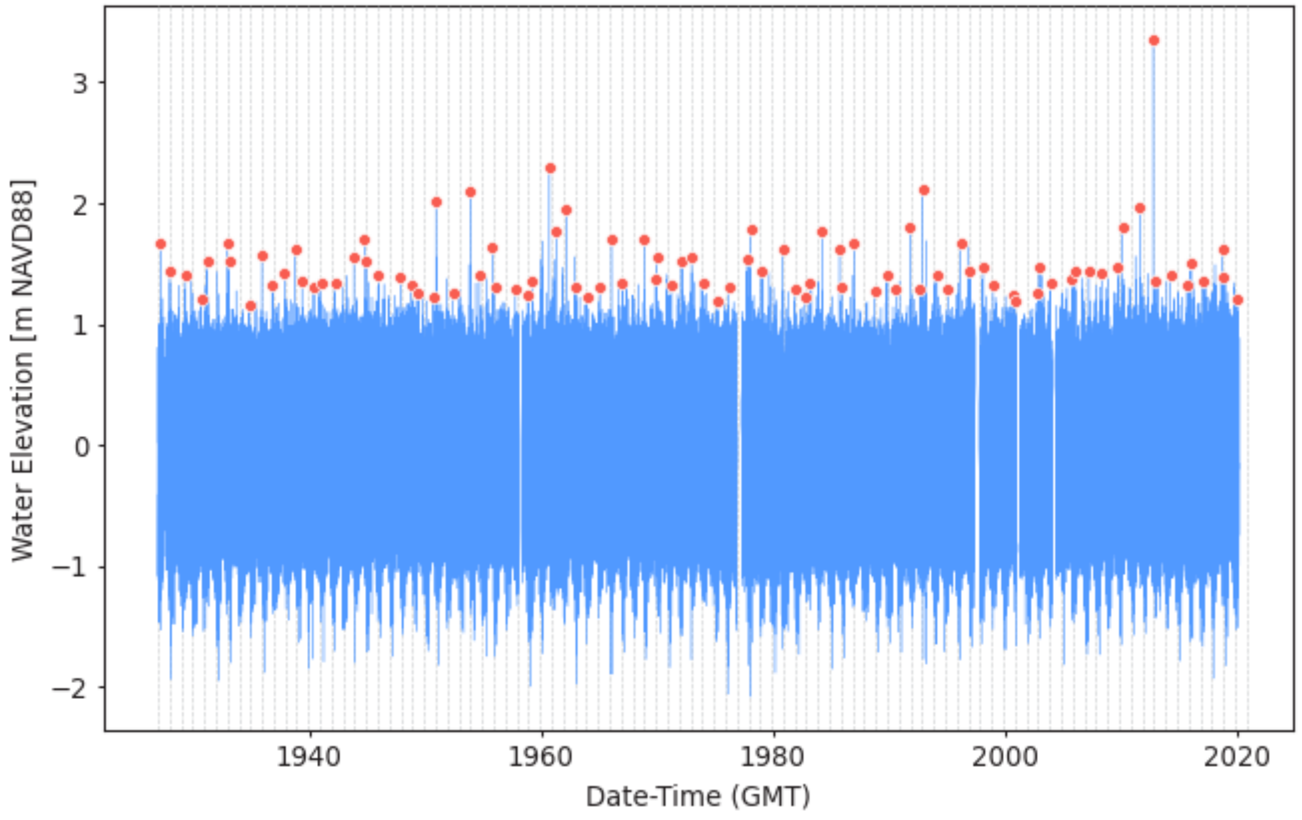
> model = EVA(data)

- EVA class provides interface to pyextremes library

> model.get_extremes(method="BM", block_size="365.2425D")

> model.plot_extremes()

Water Levels for the Battery Station in New York




See https://pypi.org/project/pyextremes/

# Example 2: pyextremes (Model Fit)

> model.fit_model()

```
                       Univariate Extreme Value Analysis
=========================================================================================
                                 Source Data
-----------------------------------------------------------------------------------------
Data label:    Water Elevation [m NAVD88]      Size:                             796,751
Start:                      November 1926      End:                           March 2020
=========================================================================================
                                Extreme Values
-----------------------------------------------------------------------------------------
Count:                                 94      Extraction method:                     BM
Type:                                high      Block size:           365 days 05:49:12
=========================================================================================
                                    Model
-----------------------------------------------------------------------------------------
Model:                                MLE      Distribution:                 genextreme
Log-likelihood:                    18.026      AIC:                             -29.786
-----------------------------------------------------------------------------------------
Free parameters:                 c=-0.266      Fixed parameters: All parameters are free
                                loc=1.353
                              scale=0.146
=========================================================================================
```

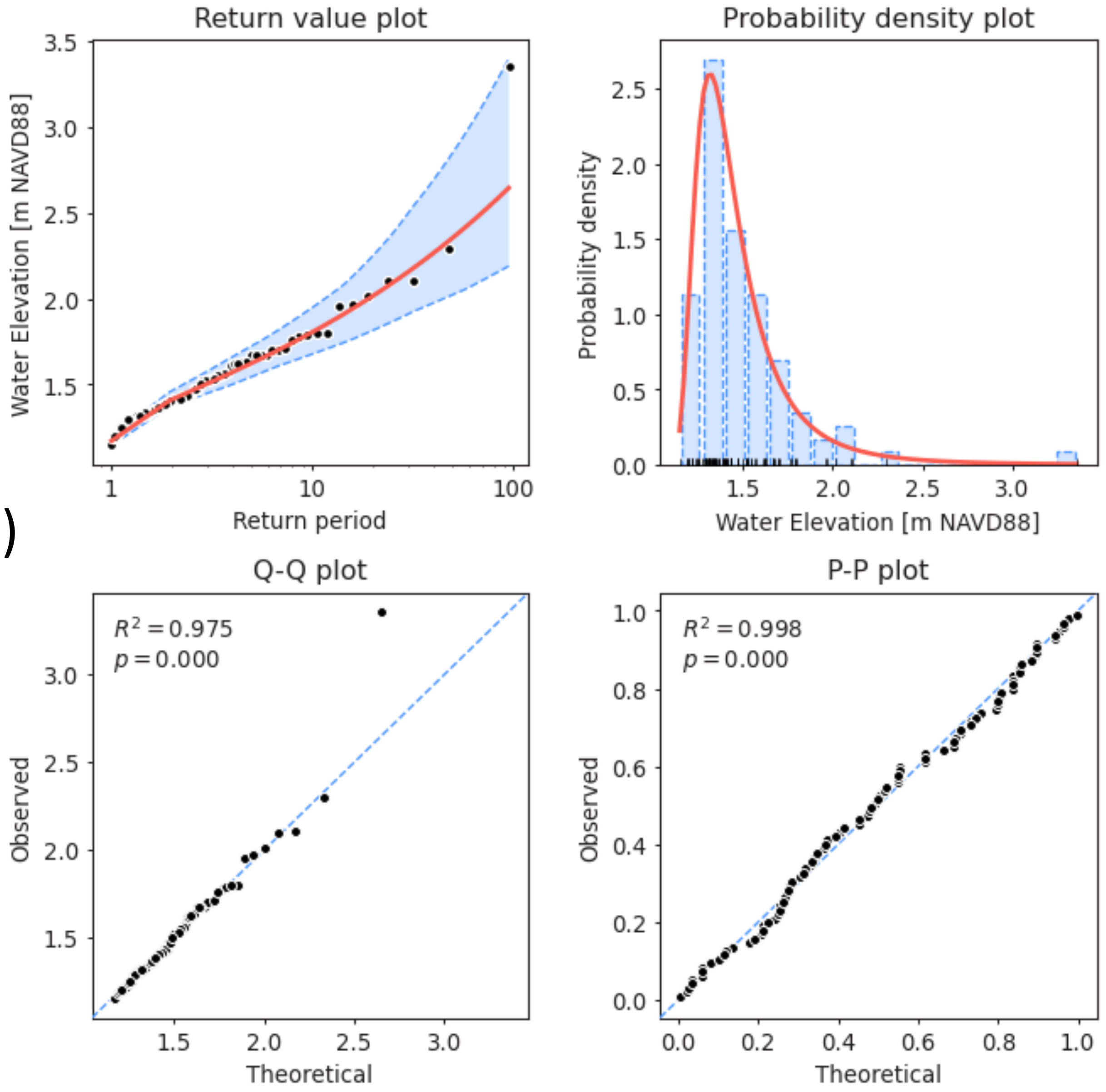
# Example 2: pyextremes (Model Summary)

```
summary = model.get_summary(
    return_period=[1, 2, 5, 10, 25, 50, 100, 250, 500, 1000],
    alpha=0.95,
    n_samples=1000,
)
summary
```

| return period | return value | lower ci | upper ci |
|---|---|---|---|
| **1.0** | 0.802610 | -0.313507 | 1.025702 |
| **2.0** | 1.409343 | 1.372263 | 1.453800 |
| **5.0** | 1.622565 | 1.547693 | 1.706435 |
| **10.0** | 1.803499 | 1.674898 | 1.951093 |
| **25.0** | 2.090267 | 1.854483 | 2.392612 |
| **50.0** | 2.354889 | 1.992968 | 2.875355 |
| **100.0** | 2.671313 | 2.139693 | 3.575801 |
| **250.0** | 3.188356 | 2.346309 | 4.843293 |
| **500.0** | 3.671580 | 2.522520 | 6.239443 |
| **1000.0** | 4.252220 | 2.704200 | 8.166698 |

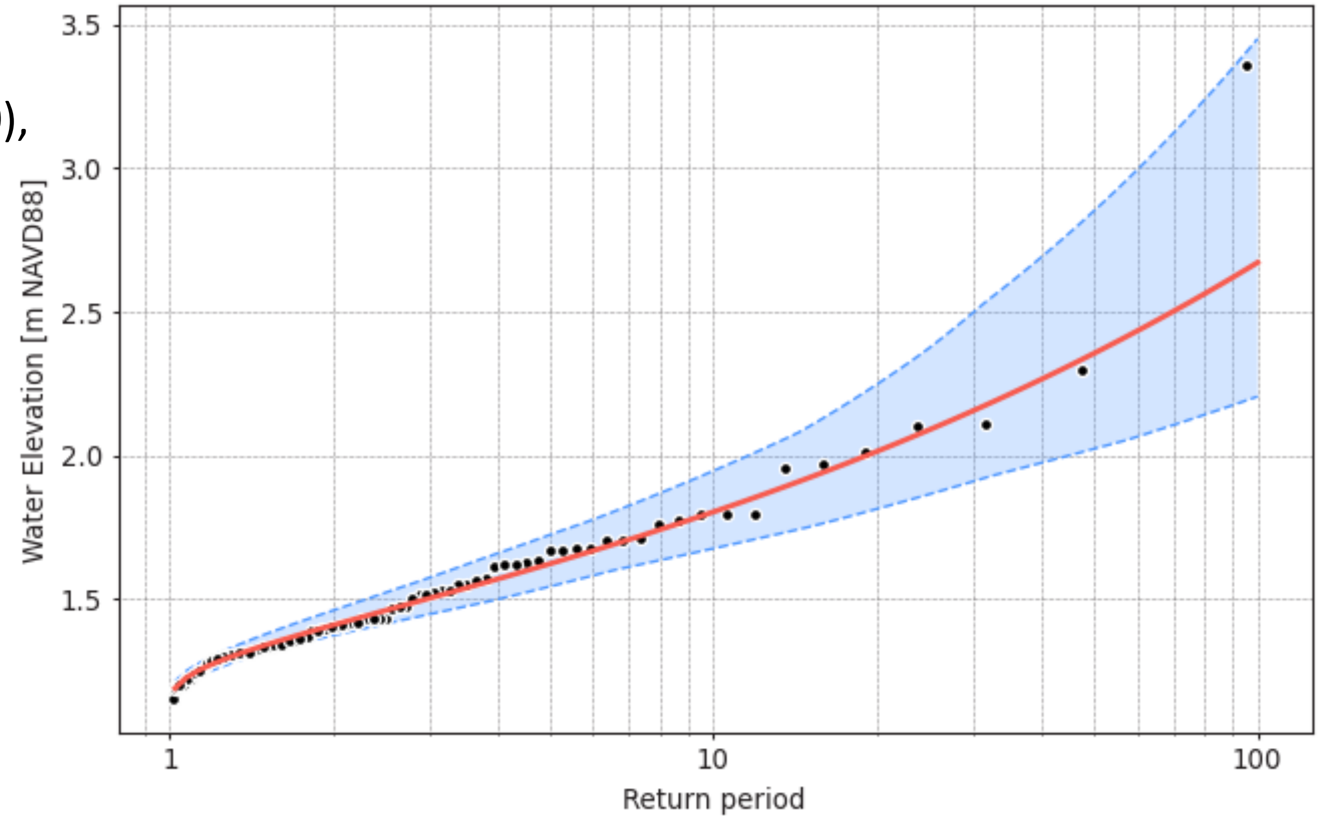
# Example 2: pyextremes (Diagnostic Plots)

> model.plot_diagnostic(alpha=0.95)



# Example 2: pyextremes (Return Values Plot)

```
> model.plot_return_values(
    return_period=np.logspace(0.01, 2, 100),
    return_period_size="365.2425D",
    alpha=0.95,
)
```



# BM or POT?

- We cannot say that one method is better than another
- Different models and approaches (correctly applied) should converge to the same answer (within reasonable limits)
- So, investigate both
- BM is a simpler and more stable model
  - Requires very little input from the user
  - Use BM with a reasonable block size to avoid capturing seasonality
  - Get the initial estimates and see how the extremes behave
- Use POT with a reasonable threshold and declustering
  - To see how well the model behaves near the target return periods
  - and to gain more confidence in the results

Source: PyExtremes User Guide