Practical 2

Melvil Deleage, Jeff Macaraeg

June 15, 2025

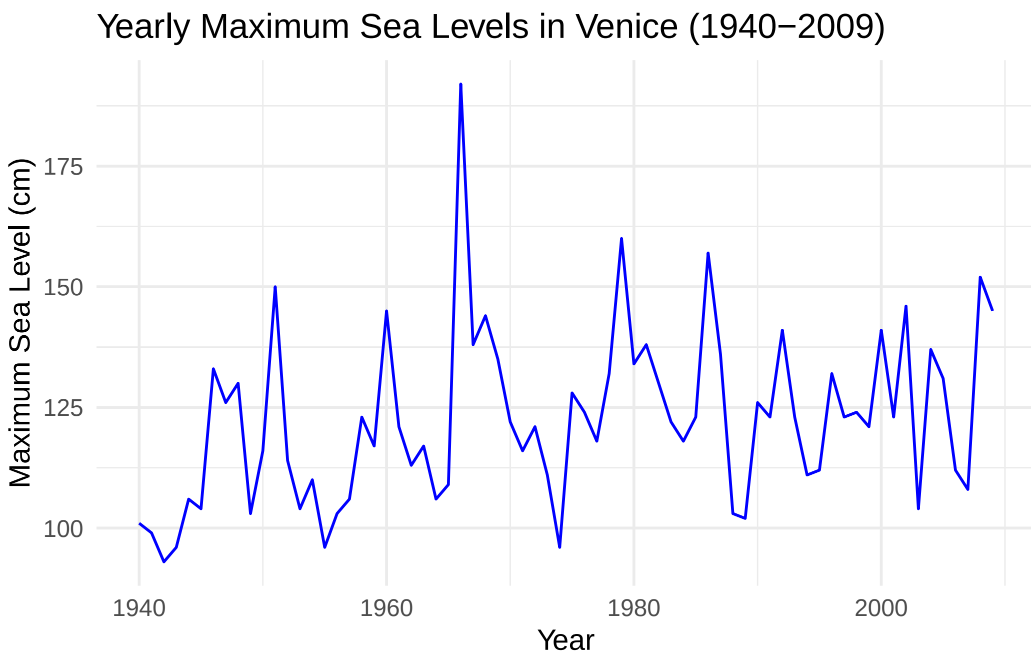
Table of contents

Part 1 - Venice

The `venice90` dataset can be found in the `VGAM` package.

Question a)

> Read in the data. Extract and represent the yearly max values from 1940 to 2009. What do you observe ?

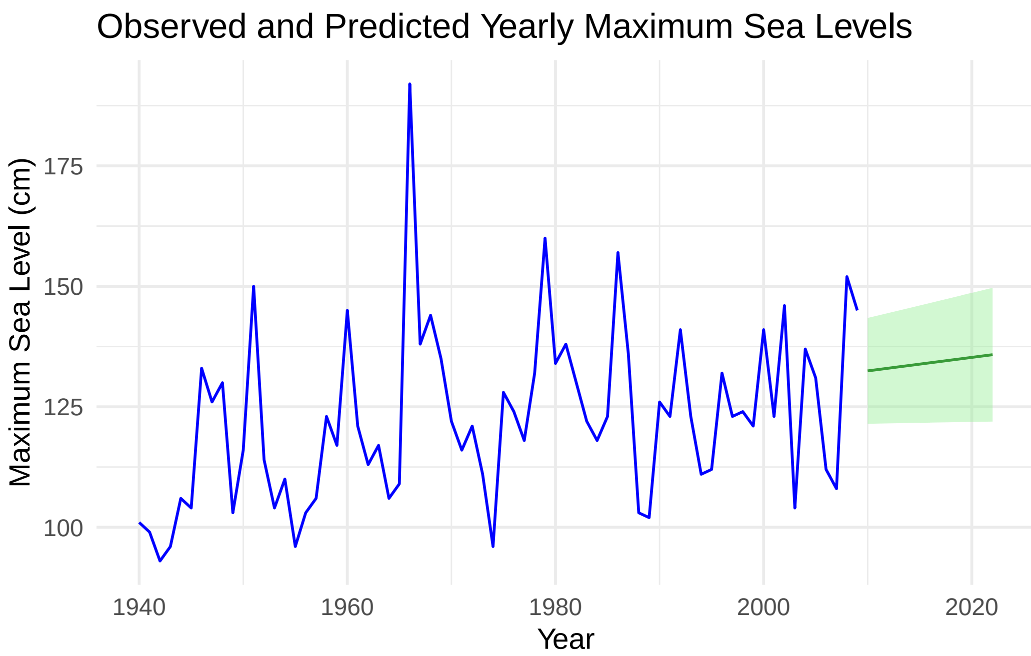


We can observe some variability over the years and a slight upward trend, so the maximum levels in Venice seem to be increasing.

Question b)

> We are end of 2009 and would like to predict the yearly maximum values over the next 13 years (from 2010 to 2022). A naive approach consists of fitting a linear model on the observed yearly maxima and predict their values for 2010–2022. Proceed to this prediction and provide confidence intervals.

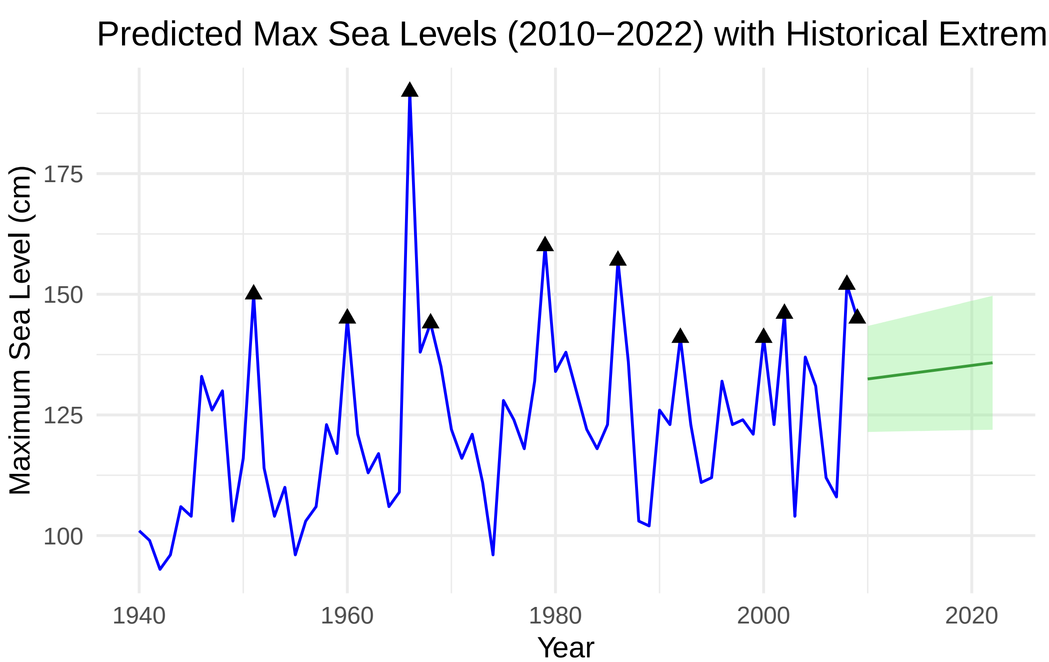
```
   year      fit      lwr      upr
1  2010 132.4522 121.4683 143.4361
2  2011 132.7321 121.5137 143.9505
3  2012 133.0121 121.5576 144.4665
4  2013 133.2920 121.6002 144.9838
5  2014 133.5719 121.6414 145.5025
6  2015 133.8519 121.6813 146.0225
7  2016 134.1318 121.7200 146.5436
8  2017 134.4118 121.7576 147.0659
9  2018 134.6917 121.7942 147.5892
10 2019 134.9716 121.8298 148.1135
11 2020 135.2516 121.8644 148.6387
12 2021 135.5315 121.8982 149.1649
13 2022 135.8115 121.9311 149.6919
```



We used a confidence interval of 99% to predict for the years 2010 to 2022.

Question c)

Represent in the same graph the predicted yearly max values for the period 2010–2022, their point-wise confidence bounds and the observed values greater than 140 cm from the table below.



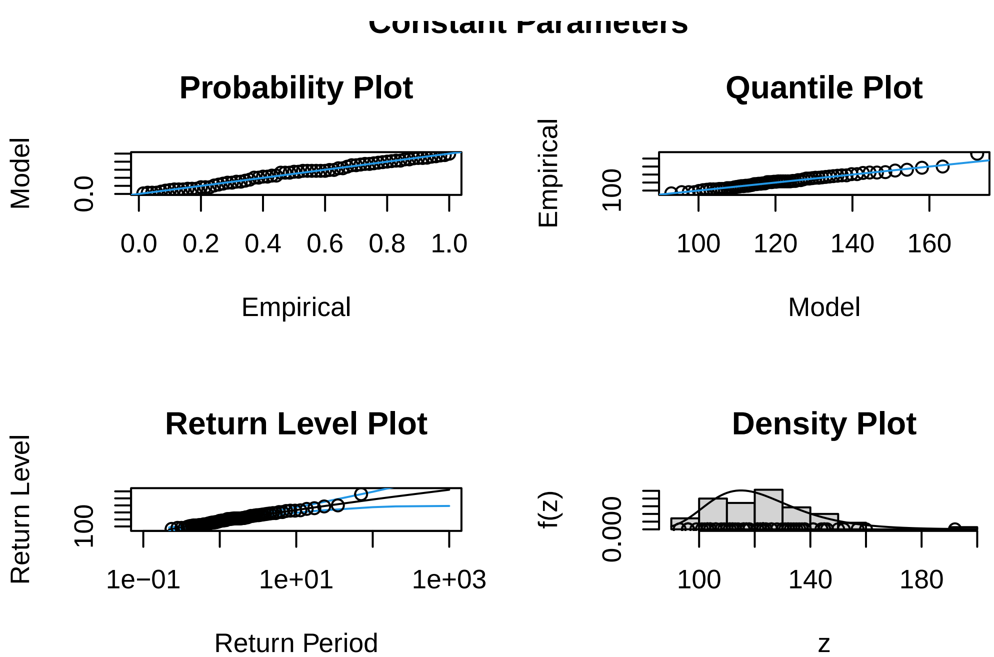
This plot provides all the necessary information, from the historical data in the blue line, to the yearly maximum values with the red points, the dark green line being the prediction for 2010 to 2022, the light green area being the confidence intervals and finally, the black triangles being the values greater than 140cm.

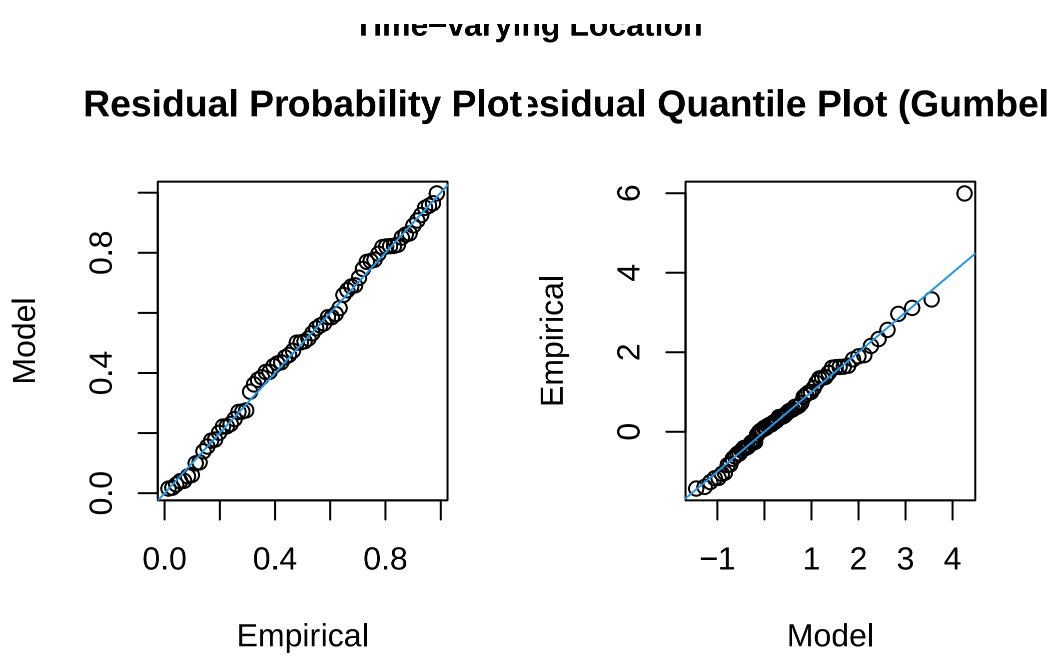
Now we perform a risk analysis and because we are interested in the period 2010–2022, we want to calculate the 13-years return level., for each year.

Question d)

Fit a GEV a with constant parameters to the historical yearly max values. Fit a GEV with time varying location parameter. Compare the two embedded models using likelihood ratio test (LRT). Show diagnostic plots.

```
--- d) Likelihood-ratio test ---
LRT = 11.62,  p = 0.001
Selected model: Time-varying Location
```





We fitted a constant and a time-varying model. The latter is better thanks to the low p-value and the log-likelihood of 11.62.The model looks overall okay, despite having some outliers which might influence it. There are no major pattern nor heteroskedasticity.

Question e)

> Add if necessary a time varying scale and or shape GEV parameter. Select the best model according to LRT.

```
# A tibble: 4 x 4
  comparison                                LR    df        p
  <chr>                                  <dbl> <dbl>    <dbl>
1 location vs const                     11.6       1 0.000651
2 location+scale vs location             0.892     1 0.345
3 location+shape vs location             5.03      1 0.0250
4 location+scale+shape vs location+scale  5.94      1 0.0148
```

```
Selected model: location+shape
```

The best model includes time-varying location and shape parameters. The addition of a time-varying scale is not necessary based on the LRT. This model provides the best fit and should be used for further analysis or prediction.

Question f) + g)

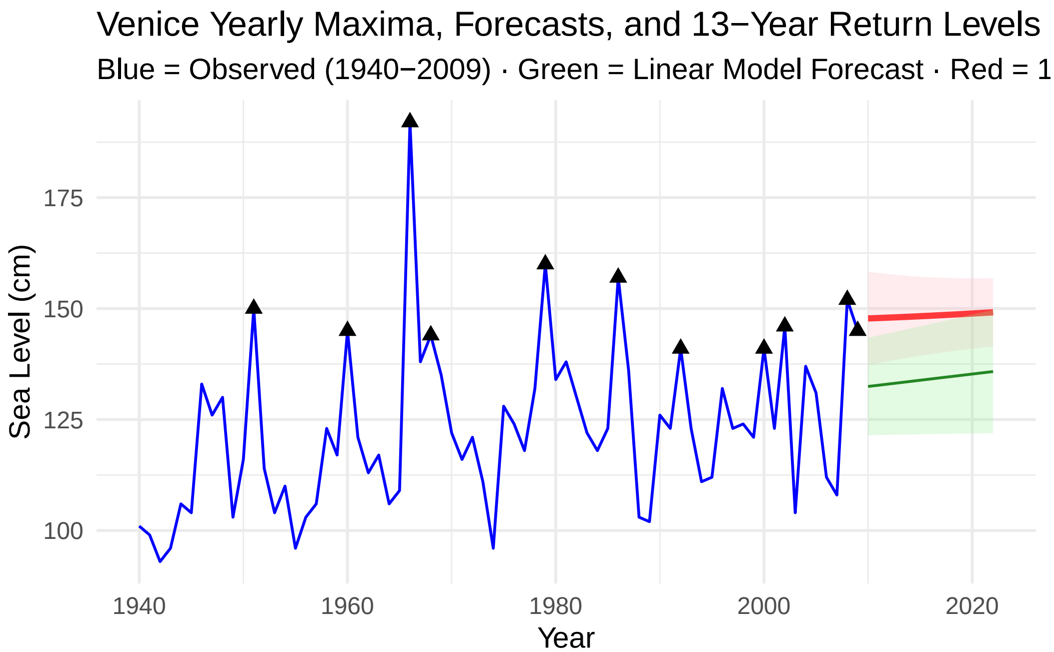
> f) Predict the 13-years return level, each year from 2010 to 2022.
> g) Calculate confidence bands for these predictions.

	year	return_level	lower_bound	upper_bound
1	2010	147.78	137.25	158.32
2	2011	147.87	137.72	158.02
3	2012	147.96	138.17	157.75
4	2013	148.06	138.60	157.52
5	2014	148.17	139.00	157.33
6	2015	148.27	139.38	157.17
7	2016	148.39	139.73	157.04
8	2017	148.51	140.07	156.94
9	2018	148.63	140.39	156.87
10	2019	148.75	140.68	156.82
11	2020	148.88	140.96	156.81
12	2021	149.02	141.22	156.82
13	2022	149.16	141.47	156.85

For each year from 2010 to 2022, the estimated 13-year return level gradually increases from approximately 147.78 cm to 149.16 cm. This indicates a slight upward trend in extreme sea level risk over time. The 95% confidence intervals range from about 137–158 cm in 2010 to 141–157 cm in 2022, showing that while uncertainty remains, the expected extremes are becoming higher. This trend supports the idea that extreme sea level events in Venice are becoming more likely and potentially more severe over time.

Question h)

> Represent in the same graph your predictions of the 13-years return levels, their pointwise confidence intervals, the predicted yearly max values from the linear model and the observed values greater than 140 cm from the table below.



Question i)

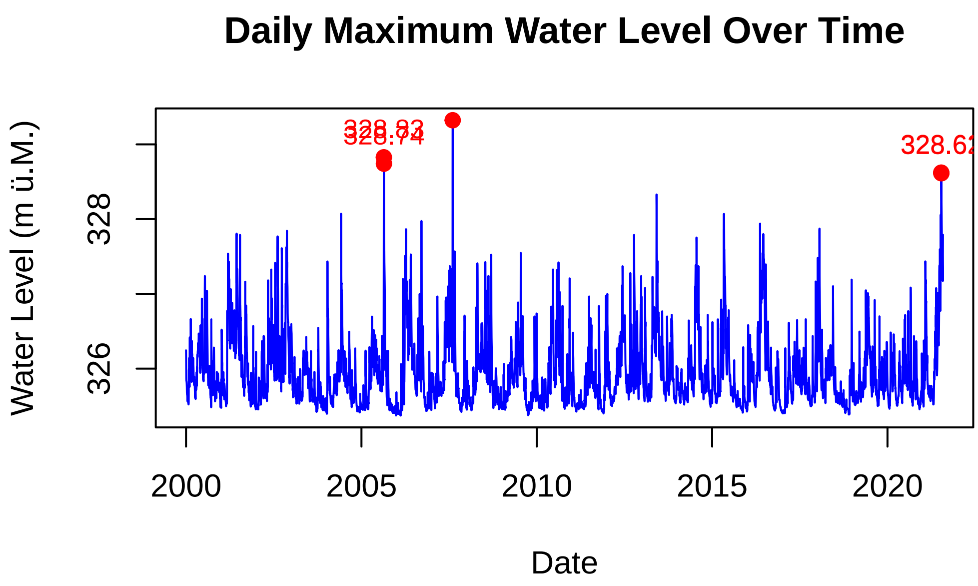
> Broadly speaking, each year, there is a chance of 1/13 that the observed value is above the 13-years return level. Comment the results for both the linear model prediction and GEV approach. Note that 12 of the 20 events occurred in the 21st century.

While both models provide useful insights, the linear model clearly underestimates extremes and provides overly narrow confidence intervals. The GEV approach, especially with time-varying parameters, is more suited for modeling extremes and gives a more realistic picture of sea level risk. However, even the GEV predictions fall short of the most recent high events, such as 2.04m in 2022, indicating that the system is non-stationary and that risk is increasing over time. This shift is emphasized by the concentration of extreme events in the 21st century, suggesting that return periods are shortening and that what was once a 13-year event may now be happening more frequently.

Part 2 - Nuclear Reactors

Question a)

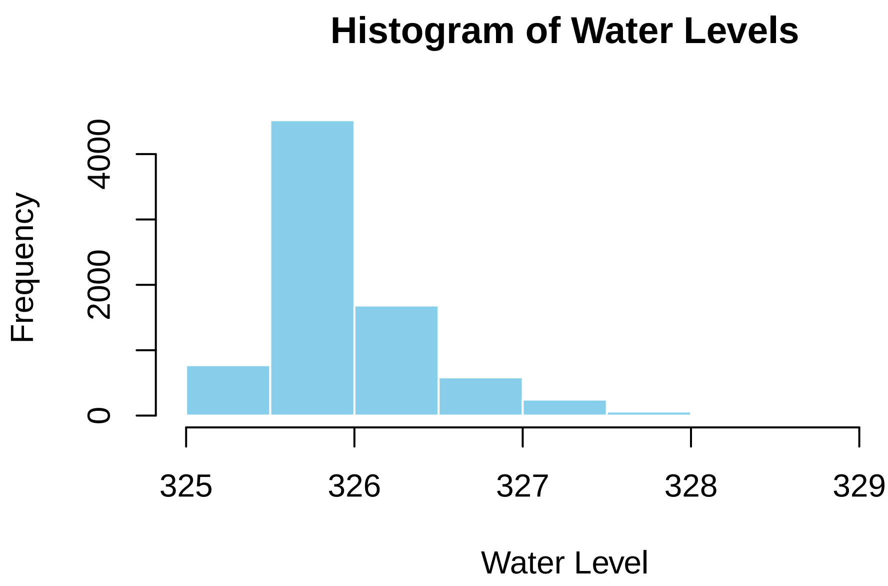
> Read in the data. Display a time series plot of the water level across the data range and try to identify times of highest levels.




```
                   Stationsname Stationsnummer Parameter   Zeitreihe
2778 Untersiggenthal, Stilli           2205     Pegel Tagesmaxima
2061 Untersiggenthal, Stilli           2205     Pegel Tagesmaxima
2062 Untersiggenthal, Stilli           2205     Pegel Tagesmaxima
7865 Untersiggenthal, Stilli           2205     Pegel Tagesmaxima
7866 Untersiggenthal, Stilli           2205     Pegel Tagesmaxima
     Parametereinheit Gewässer Zeitstempel Zeitpunkt_des_Auftretens    Wert
2778          m ü.M.     Aare  2007-08-09      2007-08-09 11:15:00 329.323
2061          m ü.M.     Aare  2005-08-22      2005-08-22 18:05:00 328.827
2062          m ü.M.     Aare  2005-08-23      2005-08-23 08:35:00 328.742
7865          m ü.M.     Aare  2021-07-14      2021-07-14 07:45:00 328.622
7866          m ü.M.     Aare  2021-07-15      2021-07-15 15:05:00 328.614
                       Freigabestatus
2778    Freigegeben, validierte Daten
2061    Freigegeben, validierte Daten
2062    Freigegeben, validierte Daten
7865 Freigegeben, provisorische Daten
7866 Freigegeben, provisorische Daten
```

Question b)

Now display a histogram of the water levels. What do you observe about the distribution?



The distribution is right-skewed. Most levels are concentrated between 325 and 326. Extreme levels such as above 327 are rare yet still present. These can represent potential flood events or unusual conditions.

> The FOEN plans for several degrees of risk. In this assignment, we focus on two risk levels: 50-year events and 100-year events.

Question c)

> Explain how you would model the high water levels using a peaks-over-threshold approach.

```
99% threshold: 327.5054
```

```
Number of exceedances: 79
```

Using a Peaks-over-Threshold approach, we set a threshold above which the values are considered extreme. This threshold should be high enough to focus only on rare exceedences, but not to high to avoid having too few exceedences. Here, the threshold is set at the 99th percentile, which is at 327.5054 (so around 327.51) meters. The exceedances are modeled using the Generalized Pareto Distribution, suitable for a skewed distribution. In this case, the POT approach is useful as there are a lot of non-extreme values. This approach thus focuses only on the extreme events to assess a better statistical efficiency, especially with daily data over many years.

Question d)

> Comment on the aspect of clustering of extremes. How do you propose to measure and deal with clustering of the daily water levels?

```
Number of exceedances before declustering: 79
```

```
Number of cluster maxima (after declustering): 28
```

Clustering extremes uses runs methods. We keep only one peak per cluster, which makes the exceedances more independent and suitable for modelling.

Question e)

> Perform the analysis you suggest in c) and d) and compute the 50- and 100-year return levels. Explain your choice of threshold and provide an estimate of uncertainty for the return levels. Note: take care to compute the return level in yearly terms.

Using the POT approach:

```
$threshold
     99%
327.5054

$nexc
[1] 28

$conv
[1] 0

$nllh
[1] 0.8853974

$mle
[1] 0.3366714 0.1202787

$rate
[1] 1

$se
[1] 0.1009116 0.2337588

50-year return level: 329.33

95% CI: 328.37 - 331.62

100-year return level: 329.73

95% CI: 328.43 - 333.53
```

The threshold is the 99th percentile to capture the extremes, have a balance between bias and variance and to have an adequate sample size to fit a GPD

Question f)

> Explain the drawbacks and advantages of using a block maxima method instead of the one used in c)-e).

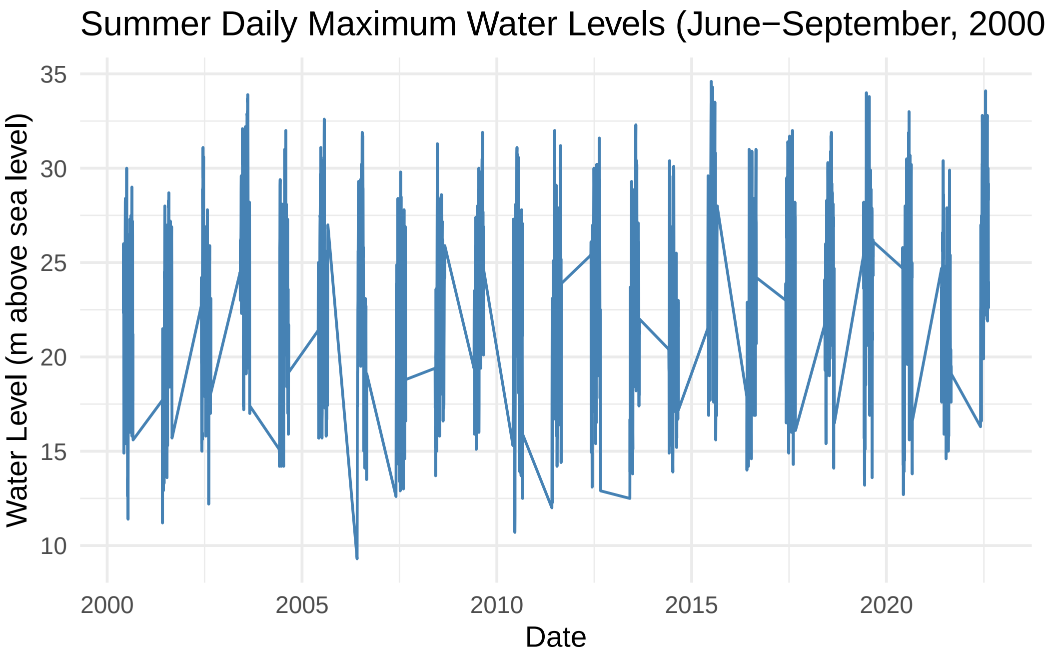
The Block Maxima method selects the maximum observation from a given time interval, but uses only one observation per block which leads to an inefficient use of the data. The POT approach uses all the values above the given threshold and handles clustering well via declustering. It is more efficient and flexible especially when extreme events happen in clusters. Thus, the POT approach is more precise and provides more information on the behavior of extreme events.

Part 3 - Night temperatures in Lausanne

Question a)

> Read in the data for the daily night maximum temperatures in Lausanne. Subset the summer months (June to September).

```
New names:
New names:
* `` -> `...1`
```



We are doing the same process for minimum to asnwer question e.

Question b)

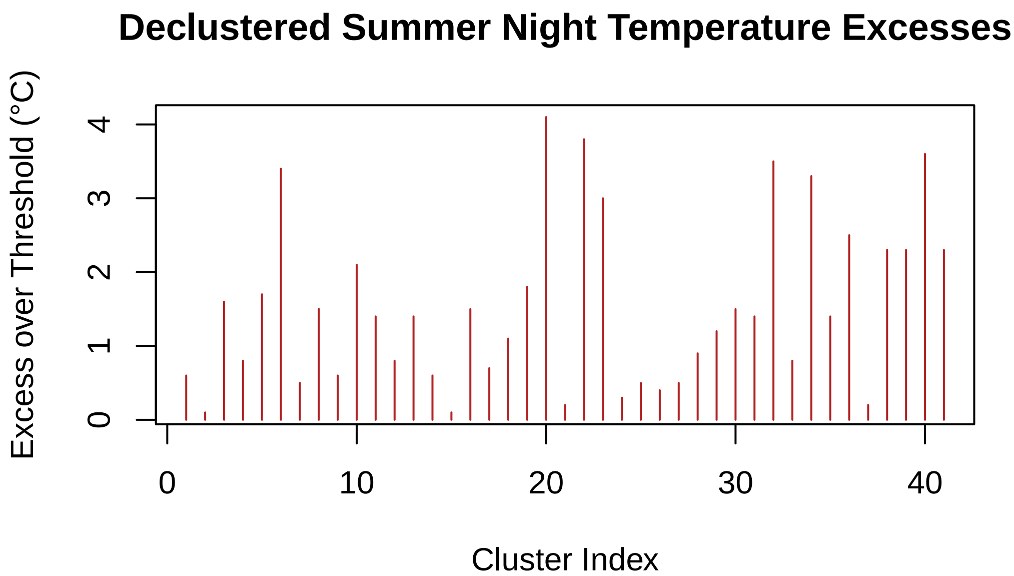
> Assess whether extremes of the subsetted series in (a) occur in cluster.

```
 Runs Estimator for the Extremal Index
    extremal.index number.of.clusters         run.length
         0.4019608         41.0000000          3.0000000

 Runs Estimator for the Extremal Index
    extremal.index number.of.clusters         run.length
         0.3673469         36.0000000          3.0000000
```

The obtained extremal index is 0.402, which is lower than 1. This suggests that extreme night temperatures during summer in Lausanne tend to occur in clusters rather than being isolated. This means that if you observe one extremely hot night, there is a higher chance that other extreme nights will follow shortly, such as during a heatwave for example.

Question c)

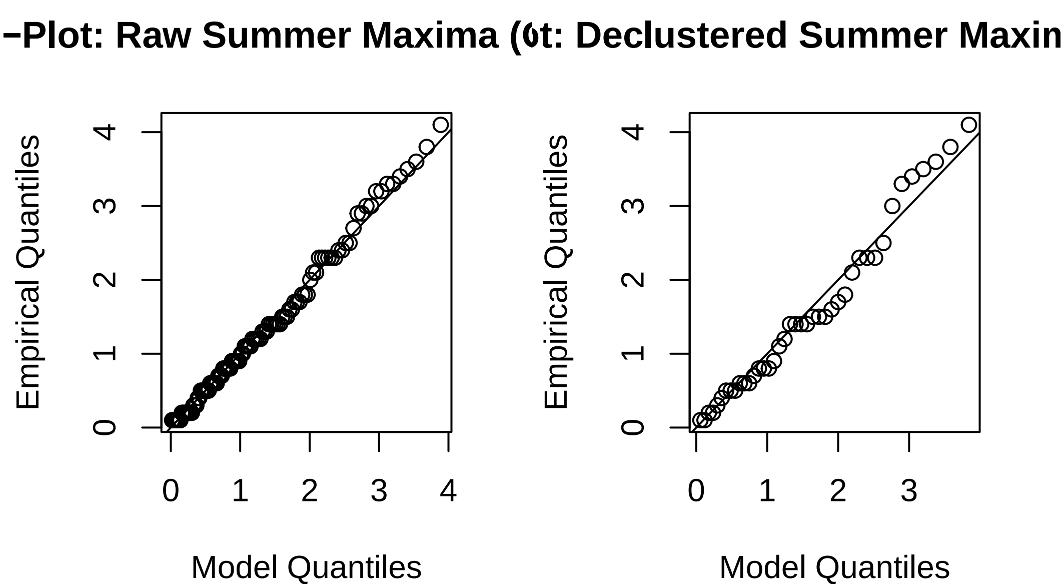
Decluster the data from (a) using a suitable threshold. Plot the resulting declustered data. (Hint: you may want to use the extRemes package.)

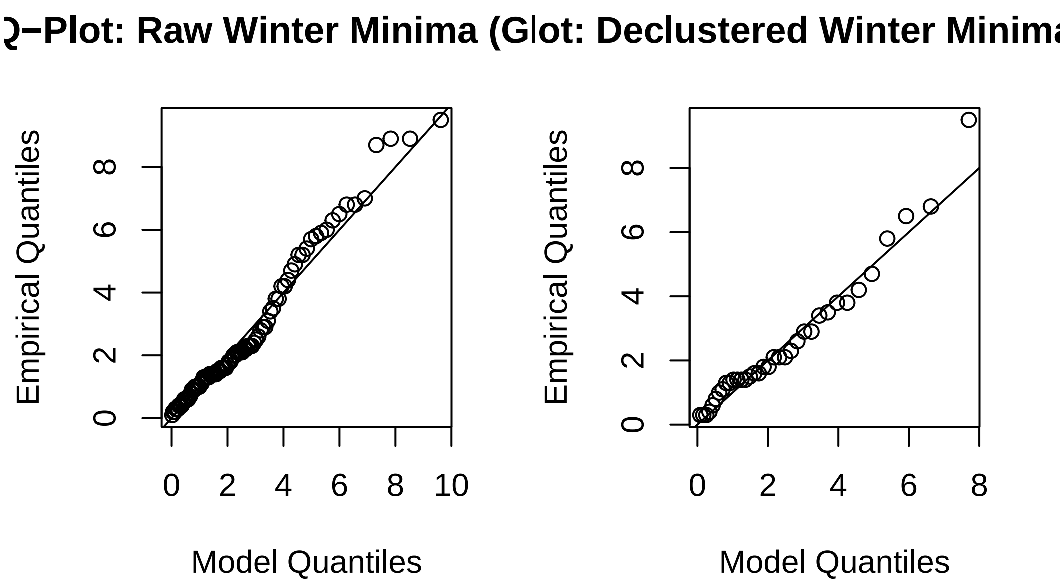


After declustering the extreme summer night temperatures using the 95th percentile threshold and a 3-day run length, we isolated 42 independent exceedances above the threshold. The resulting plot of declustered excesses reveals a wide range of magnitudes, with some cluster peaks exceeding 4°C above the threshold. This confirms the presence of significant and varied extreme temperature events, now stripped of temporal dependence.

Question d)

Fit a GPD to the data, both raw and declustered. Assess the quality of the fit.





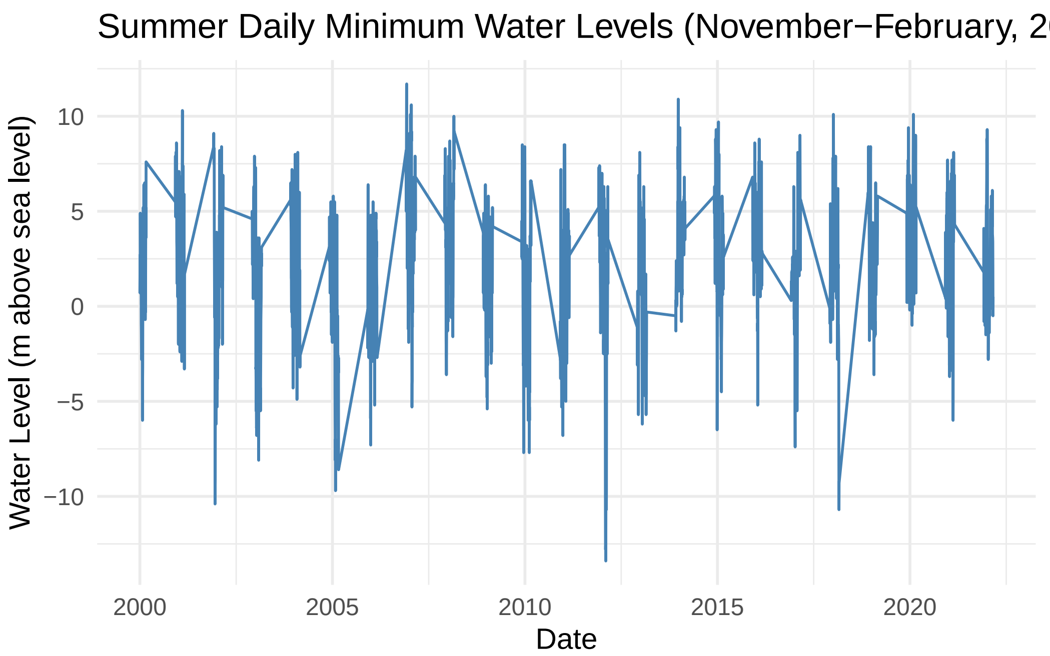
Despite the raw model having points more aligned in the QQ-plot, the declustered model is theoretically better due to the lower AIC. The deviations in the QQ-plot for the declustered can be explained due to the smaller sample than the raw data.

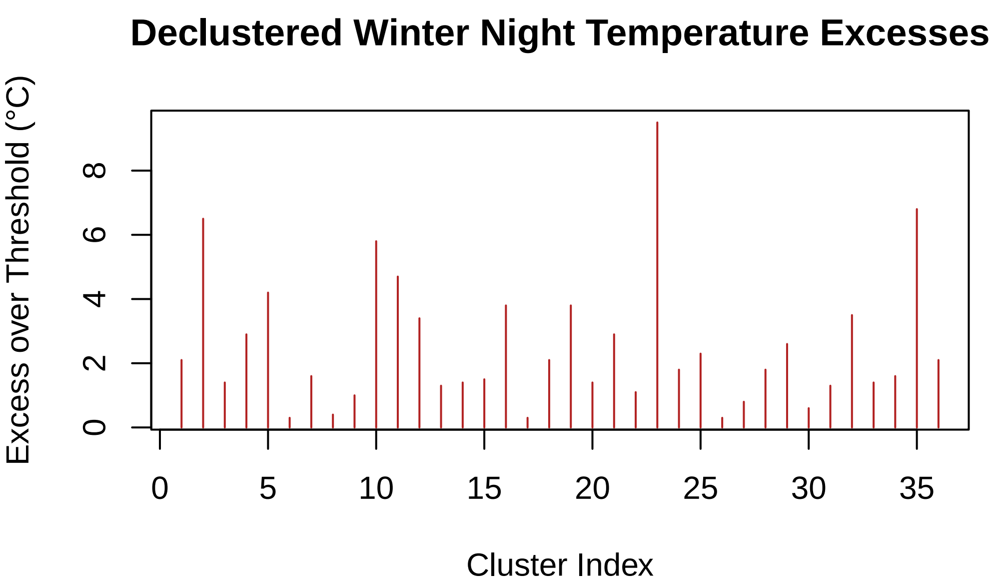
Question e)

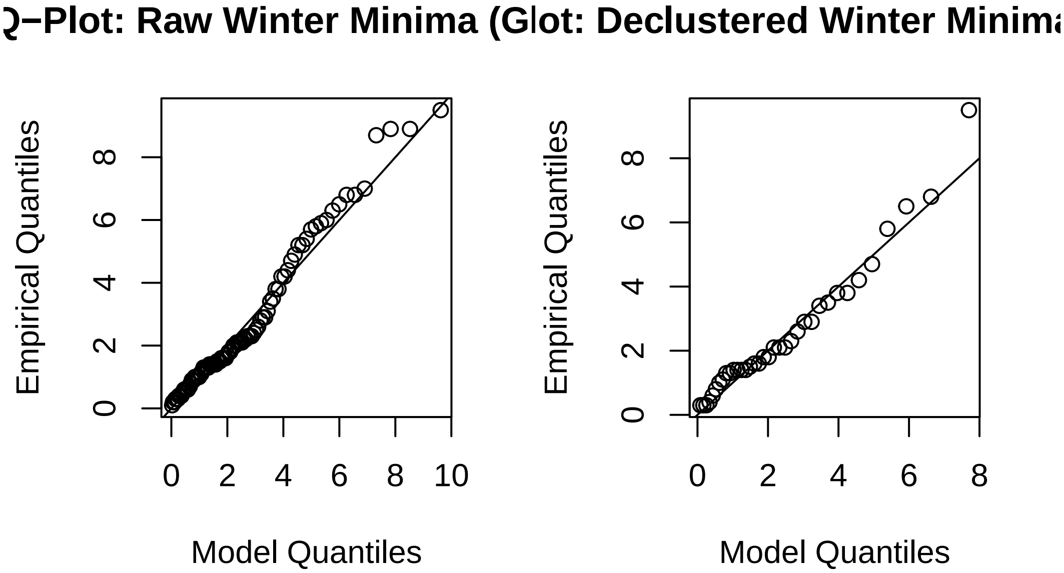
> Repeat the above analysis for the negatives of the daily nightly minimum temperatures for the winter months (November-February).

```
$winter_extremal_index

 Runs Estimator for the Extremal Index
    extremal.index number.of.clusters         run.length
         0.3673469         36.0000000          3.0000000

$gpd_winter_raw

fevd(x = (-winter_min$tmin)[-winter_min$tmin > u_min] - u_min,
    threshold = 0, type = "GP", method = "MLE")

[1] "Estimation Method used: MLE"


 Negative Log-Likelihood Value:  186.0393


 Estimated parameters:
     scale      shape
 2.8113093 -0.1352909

 Standard Error Estimates:
    scale     shape
```

```
0.4203671 0.1110884

 Estimated parameter covariance matrix.
            scale       shape
scale  0.17670849 -0.03804989
shape -0.03804989  0.01234063

 AIC = 376.0786

 BIC = 381.2486

$gpd_winter_dc

fevd(x = excess_min, threshold = 0, type = "GP", method = "MLE")

[1] "Estimation Method used: MLE"


 Negative Log-Likelihood Value:  68.36057


 Estimated parameters:
     scale      shape
 3.0113921 -0.2034978

 Standard Error Estimates:
    scale     shape
0.6471591 0.1398393

 Estimated parameter covariance matrix.
            scale       shape
scale  0.41881484 -0.07244737
shape -0.07244737  0.01955502

 AIC = 140.7211

 BIC = 143.8882
```


Summer Daily Minimum Water Levels (November−February, 2
Water Level (m above sea level)
10
5
0
−5
−10
2000
2005
2010
2015
2020
Date


Declustered Winter Night Temperature Excesses
Excess over Threshold (°C)
0
2
4
6
8
0
5
10
15
20
25
30
35
Cluster Index



We apply the negative to the winter values to treat the extremely low values as high values for modelling purposes. We then do an extremal index and we obtain 0.367, lower than 1, indicating clustering. We then declustered using the 95th percentile threshold to the negated temperatures and the plot shows that some peaks go even above 6 degrees. Fitting the model using GPD shows that the AIC for the declustered is again lower than raw, so a better fit.