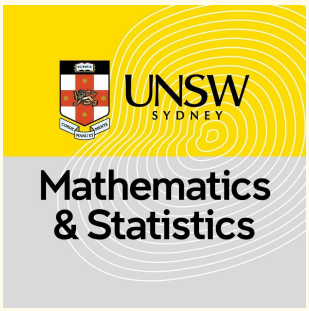


MATH5805
A Modern Introduction to Extreme Value Theory
Term 1, 2025

Dr. Boris Béranger

Week 3 - Lab Solutions

Exercise 1

The file `AUD-exchangeRates.csv` contains historical daily exchange rates of the Australian dollar against a number of other currencies between January 2010 to December 2013.

The dataset columns

`FXRUSD FXRCR FXRJY FXREUR FXRSKW FXRUKPS FXRSD FXRIRE FXRTB FXRNZD FXRNTD FXRMR FXRIR FXRVD FXRUAED FXRPNGK FXRHKD FXRCD FXRSARD FXRSARY FXRSF FXRSK FXRPHP FXRSDR`

respectively denote the AUD/x exchange rate with the

`USD CNY JPY EUR KRW GBP SGD INR THB NZD TWD MYR IDR VND AED PGK HKD CAD ZAR SAR CHF SEK PHP SDR`.

Column `FXRTWI` is the AUD trade-weighted index (May 1970=100).

a) Read the data into R using:

```
read.table("AUD-exchangeRates.csv", header=TRUE, sep=",")
```

Plot the AUD/USD exchange rate. These data are clearly non-stationary. If the series is $X_1, \ldots, X_n$ it is common to instead model log daily returns:

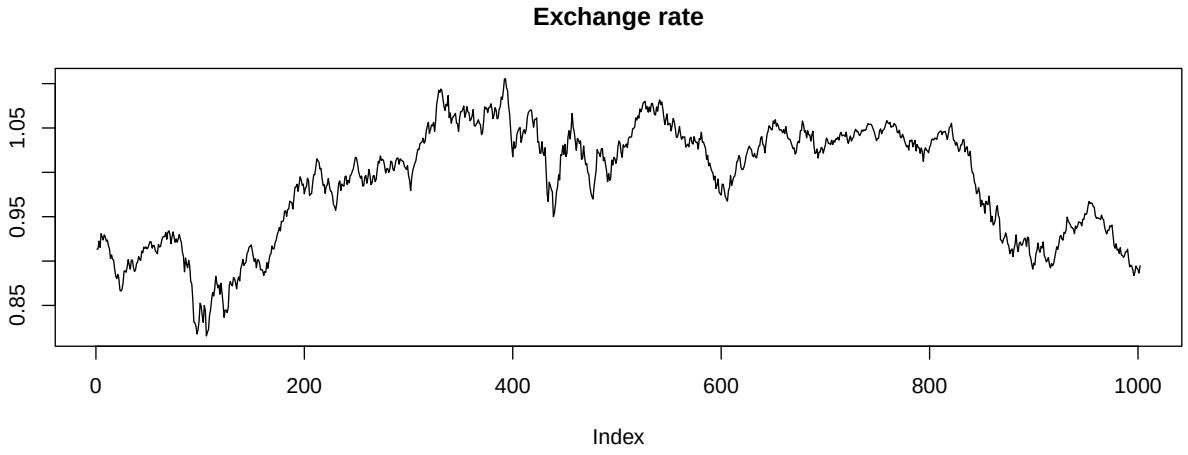
$$Y_i = \log(X_{i+1}) - \log(X_i) \qquad i = 1, \ldots, n-1.$$

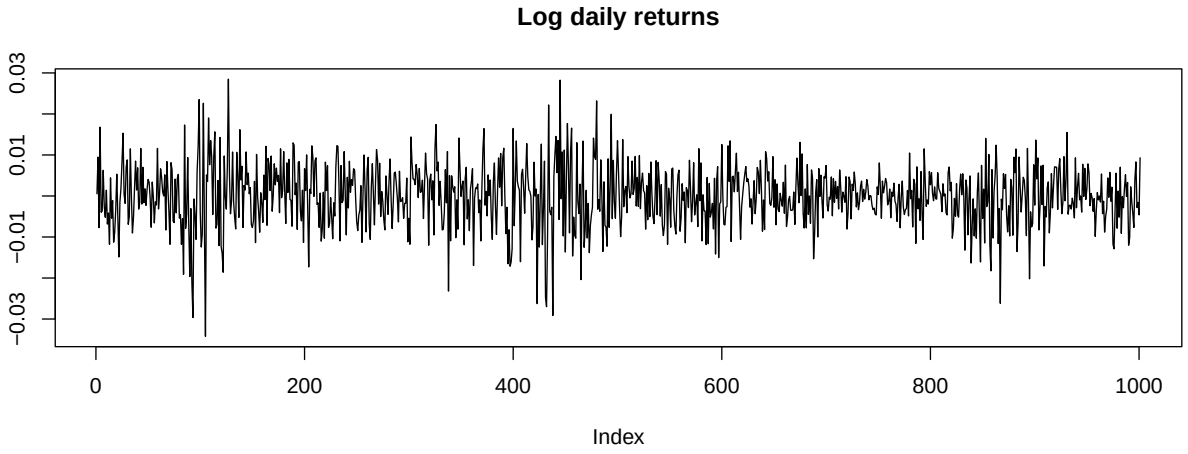
Construct and plot these data. Note that we have interest in both large positive and large negative values.

Solution:

```
# Modify path to dataset if needed
dat <- read.table("../data/AUD-exchangeRates.csv", header=TRUE, sep=",")
usaud <- dat$FXRUSD
par(mfrow=c(2,1))
plot(usaud, type='l', main='Exchange rate', ylab="")
L <- length(usaud)
```

```r
ldr <- log(usaud[2:L]/usaud[1:(L-1)])
plot(ldr, type='l', main="Log daily returns", ylab="")
```





b) There are 1,001 log daily return observations over 4 years. If we construct block maxima on a yearly scale we will have only 4 data points (resulting in an analysis with high variance). If we construct block maxima on a daily scale, the extreme value limiting distribution may not hold (high bias). Explore constructing block maxima over different scales including:

- daily
- weekly (note: only weekday exchange rates are recorded)
- monthly
- quarterly
- annually.

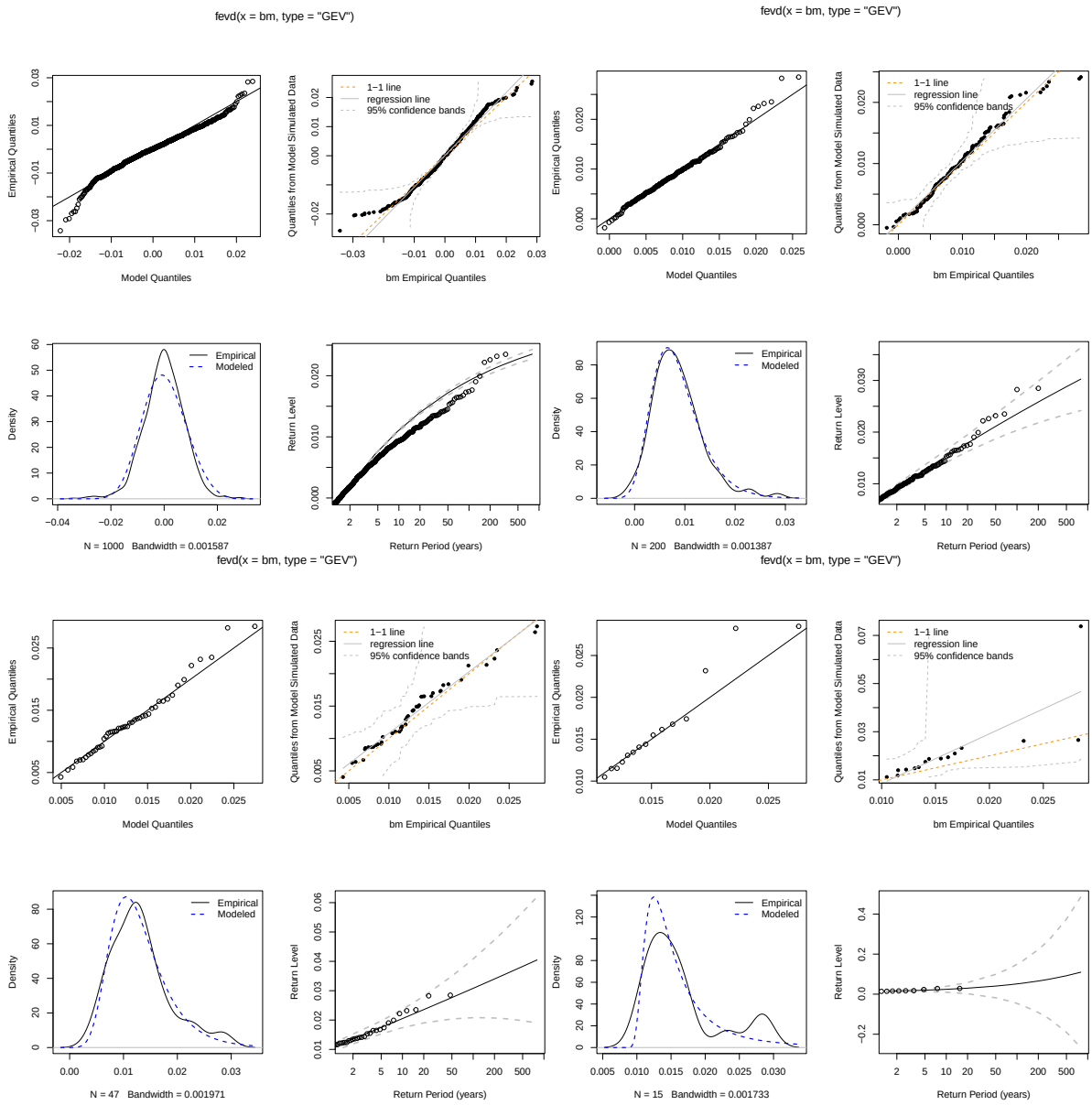
Which scales are we able to produce a viable analysis using GEV modelling of block maxima? You may assume that these correspond to $n = 1$ (daily), 5 (weekly), 21 (monthly), 65 (quarterly), 250 (annually) consecutive observations, for simplicity. Consider goodness-of-fit diagnostics to inform your choice.

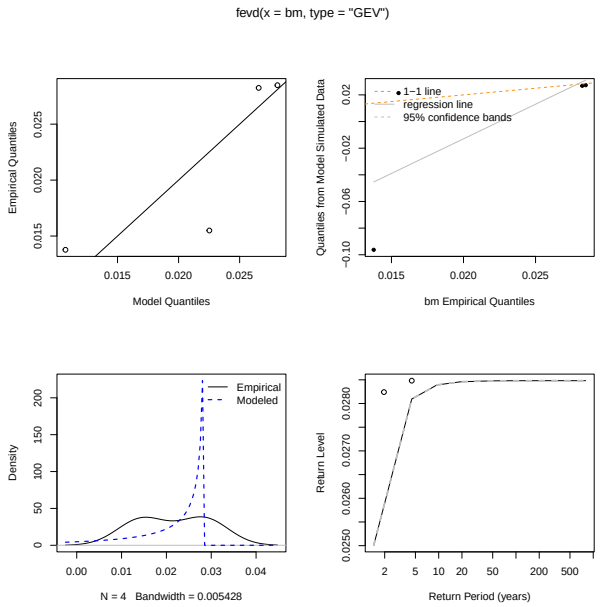
Solution:

Its a little fiddly to exactly put these data into calendar weeks/quarters/years etc, so we can

do this approximately by assuming that ever consecutive $n = 1, 5, 21, 65, 250$ observations corresponds to each block, and ignoring any remaining observations that don't completely fill a block.

```
library(extRemes)
nn <- c(1,5,21,65,250) # approximate no. of data per block
for(i in 1:length(nn)){
  temp <- matrix(ldr, ncol=nn[i], byrow=T)
  temp <- temp[-dim(temp)[1],] # need to remove last row
  bm <- apply(as.matrix(temp), 1, max)
  fit <- fevd(bm, type="GEV")
  plot(fit)
}
```





For $n = 1$ the GEV limit doesn't hold, with the qq-plot and return-level plot showing clear divergence from the GEV. For $n = 65$ and $n = 250$ we are encountering high variability of parameter estimates due to low amounts of data. So here the $n = 250$ model fit is unreliable, although the $n = 65$ fit doesn't look too bad, even given the lack of data. For $n = 5$ (weekly) and $n = 21$ (monthly) the fits look credible.

c) For each scale, what is the predicted return period of the largest observed log daily return? Do these predictions make sense given the length of the each observed block-maxima dataset?

Solution:

The maximum observed log daily return is 0.0285, which will appear in each dataset as the maximum of the block in which it resides. For each dataset the empirical estimate of the probability of observing greater than or equal to this value is $1/N$, where N is the number of block maxima ($\approx 1001/n$) in each dataset. This means that the estimated return period of this observation is $1/(1/N) = N$.

For the fitted GEV model, if $p = G(0.0285; \hat{\mu}, \hat{\sigma}, \hat{\xi})$ is the probability of observing a block maxima less than 0.0285, then $1/(1-p)$ is the estimated return period of observing 0.0285 (or above).

```
m <- max(ldr)
for(i in 1:length(nn)){
  temp <- matrix(ldr, ncol=nn[i], byrow=T)
  temp <- temp[-dim(temp)[1],] # need to remove last row
  bm <- apply(as.matrix(temp), 1, max)
  fit <- fevd(bm, type="GEV")
  emp <- length(bm)/sum(bm>=m)
  p <- fit$results$par
  pp <- pevd(m, loc=p[1], scale=p[2], shape=p[3])
  cat("Empirical RP =", emp, "; Model RP =", 1/(1-pp), "\n")
}
```

```
## Empirical RP = 1000 ; Model RP = 188341.9
```

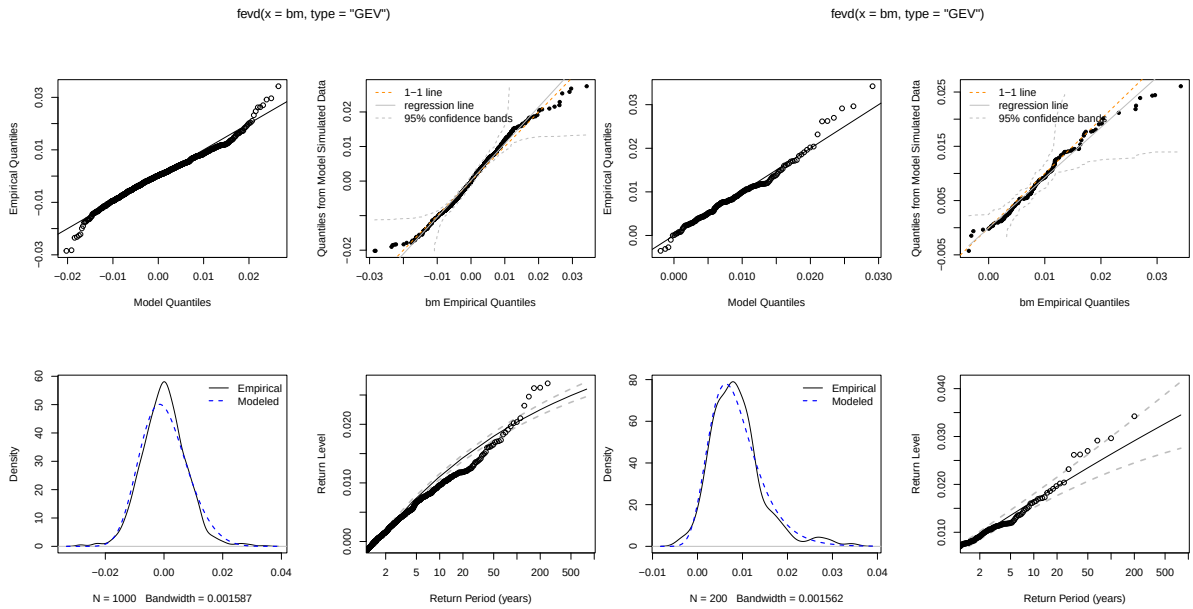
```
## Empirical RP = 200 ; Model RP = 454.5392
## Empirical RP = 47 ; Model RP = 60.39598
## Empirical RP = 15 ; Model RP = 17.66933
## Empirical RP = 4 ; Model RP = 4893533
```

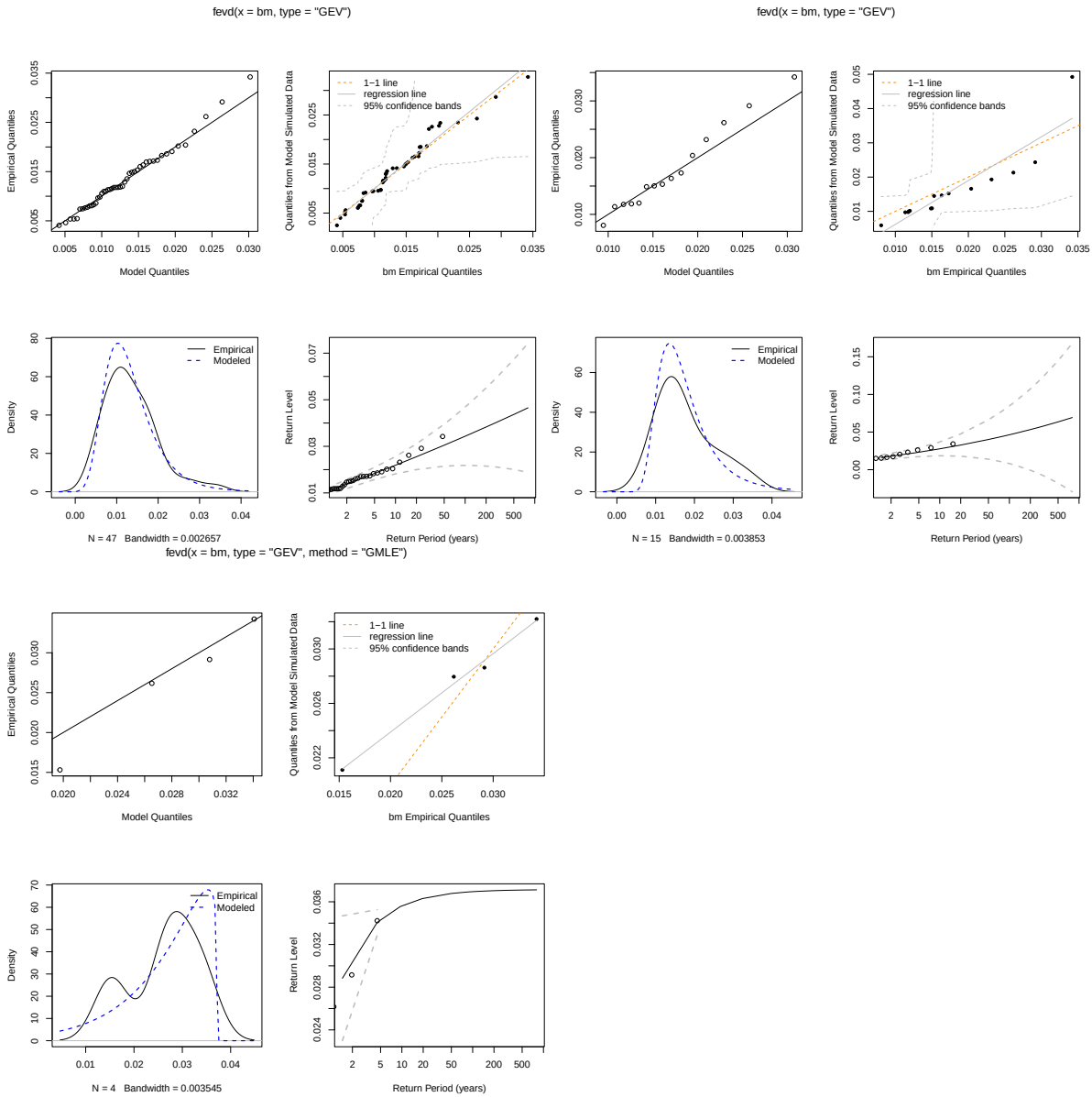
If the model fits the data well, then the empirical and model-based return period will be approximately consistent with each other. For $n = 1$ and $n = 260$ these estimates are clearly far apart. The estimates are reasonably close (a factor of 2) for $n = 5$ and really close for $n = 21$ and $n = 65$. Taken together with the previous diagnostics, we might be reasonably happy with the model fits for $n = 21$ and $n = 65$, perhaps with a preference for $n = 21$ so that more data (and inferential precision) is available.

d) Repeat questions b) and c) but considering the extreme low log daily returns via $Z_i = -Y_i$. Do your conclusions change?

Solution:
For b):

```
m <- max(-ldr)
for(i in 1:length(nn)){
    temp <- matrix(-ldr, ncol=nn[i], byrow=T)
    temp <- temp[-dim(temp)[1],] # need to remove last row
    bm <- apply(as.matrix(temp), 1, max)
    if(i==5){
      fit <- fevd(bm, type="GEV", method="GMLE")  # Otherwise Error
    }else{
      fit <- fevd(bm, type="GEV")
    }
    plot(fit)
}
```





For c):

```
for(i in 1:length(nn)){
  temp <- matrix(-ldr, ncol=nn[i], byrow=T)
  temp <- temp[-dim(temp)[1],] # need to remove last row
  bm <- apply(as.matrix(temp), 1, max)
  fit <- fevd(bm, type="GEV")
  emp <- length(bm)/sum(bm>=m)
  p <- fit$results$par
  pp <- pevd(m, loc=p[1], scale=p[2], shape=p[3])
  cat("Empirical RP =",emp,"; Model RP =",1/(1-pp), "\n")
}
```

```
## Empirical RP = 1000 ; Model RP = 337972
## Empirical RP = 200 ; Model RP = 745.3884
```

```
## Empirical RP = 47 ; Model RP = 98.57419
## Empirical RP = 15 ; Model RP = 24.93447
## Empirical RP = 4 ; Model RP = Inf
```

From the model fit diagnostics (plots) we have similar results to those for the upper tail behaviour. $n = 1$ and $n = 260$ are either poor or unreliable fits. The GEV fit for $n = 5$ looks questionable in the upper tail for the 6 largest observations. The fit looks reasonable for $n = 21$ and $n = 65$. This assessment is largely supported by the empirical and model-based return periods of 0.0285. So again, we might choose $n = 21$ given the reasonable fit, and larger number of observations.

e) Should we expect any relationship between positive and negative tail behaviour in general?

Solution:
There is no reason in general to expect that the lower tail of a distribution should behave the same as the upper tail. This should be obvious when considering a few examples. When considering log daily returns of stock prices, trader behaviour will be different if there have been (relatively) large price increases or decreases the previous day, producing different distributional tail behaviour for log daily returns. When considering rainfall distributions, upper and lower rainfall tail behavour will be different as it can be driven by different rain generating processes e.g. convective or frontal rainfall (not to mention that rainfall amount is bounded below by zero).

f) This analysis has assumed that the log daily returns are stationary. Is this assumption valid?

Solution:
Visual inspection of the log daily return plot suggests that there may be some scale heterogeneity. This would imply that larger extremes (maxima or minima) become more likely at some times compared to others.

Exercise 2

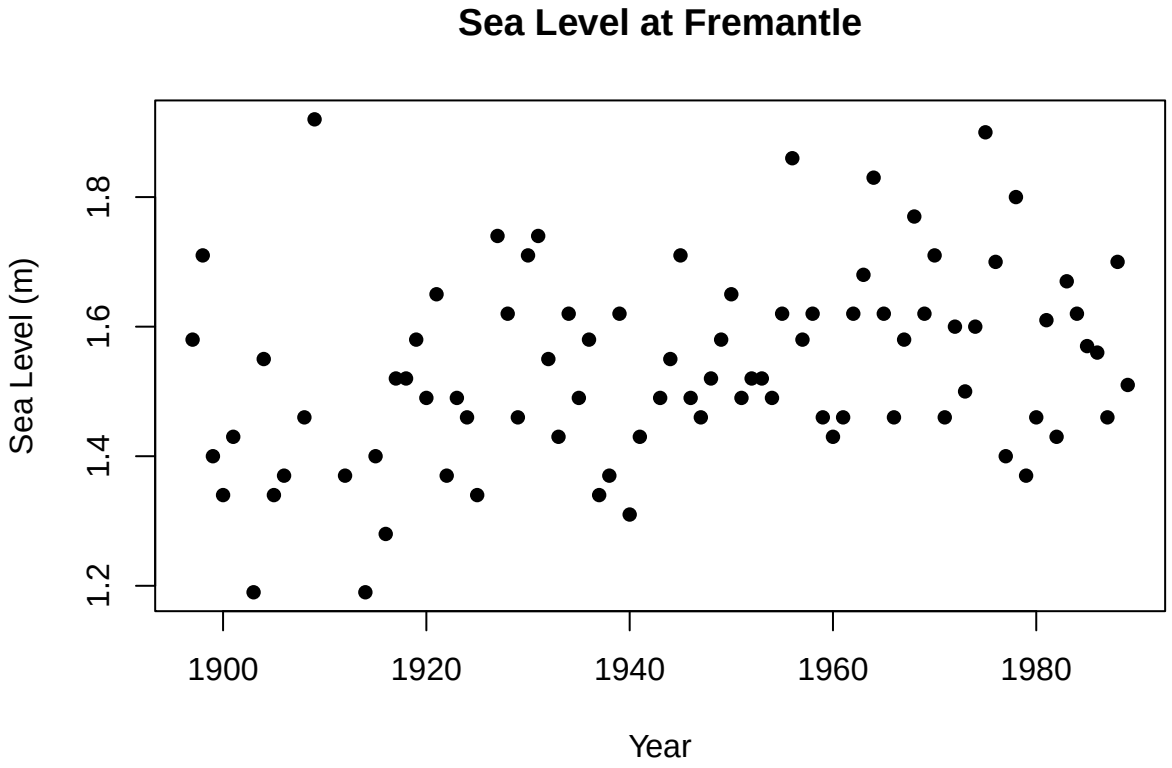
The dataset freemantle in the `R` package `ismev` contains annual maximum sea levels in metres at Fremantle in Western Australia from 1837–1989, with a few observations missing in the early years. Also included is the annual mean value of the Southern Oscillation Index (SOI) for each year there is a recorded observation. The data is loaded as follows

```
library(ismev)
data(fremantle)
```

a) Plot the annual maxima data. Do the data look stationary?

Solution:

```
library(ismev)
data(fremantle)
fre <- fremantle
plot(fre$Year, fre$SeaLevel, pch=16, xlab="Year", ylab="Sea Level (m)",
     main="Sea Level at Fremantle")
```



Despite the largest annual maxima being observed in 1909, and even allowing for the possibility of more variable data records prior to 1930, it does appear that there is a steady increase in annual maximum daily sea levels over the observation period. Clearly fitting a GEV(μ,σ,ξ) with constant parameters would be in appropriate here.

b) One way to account for this non-stationarity is to allow the GEV model parameters μ, σ, ξ to vary with time: $\mu(t), \sigma(t), \xi(t)$. This can be implemented within the `fevd` function by supplying a data frame (`data`) and writing a formula for each model parameter. E.g.~the following code will fit a GEV with $\mu(t) = \mu_0 + \mu_1(\text{Year}_t - \overline{\text{Year}})$:

```
X <- data.frame(time=fre$Year-mean(fre$Year))
fit <- fevd(fre$SeaLevel, data=X, location.fun=~time)
```

Explore whether the GEV model parameters can be expressed as linear or quadratic models over time. Likelihood ratio tests can be performed using e.g.~`lr.test` in the `extRemes` package. Is there any evidence for a non-constant trend in the GEV model parameters?

Solution:

```
fit.const <- fevd(fre$SeaLevel)
X <- data.frame(time=fre$Year-mean(fre$Year),
                time2=(fre$Year-mean(fre$Year))^2)
fit.1 <- fevd(fre$SeaLevel, data=X, location.fun=~time)
lr.test(fit.const, fit.1) # support for loc linear in time
```

```
##
##  Likelihood-ratio Test
##
```

```
## data:  fre$SeaLevelfre$SeaLevel
## Likelihood-ratio = 12.692, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.0003672
## alternative hypothesis: greater
```

```
fit.2 <- fevd(fre$SeaLevel, data=X, location.fun=~time+time2)
lr.test(fit.1, fit.2) # no support for quadratic over linear
```

```
##
##  Likelihood-ratio Test
##
## data:  fre$SeaLevelfre$SeaLevel
## Likelihood-ratio = 1.4839, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.2232
## alternative hypothesis: greater
```

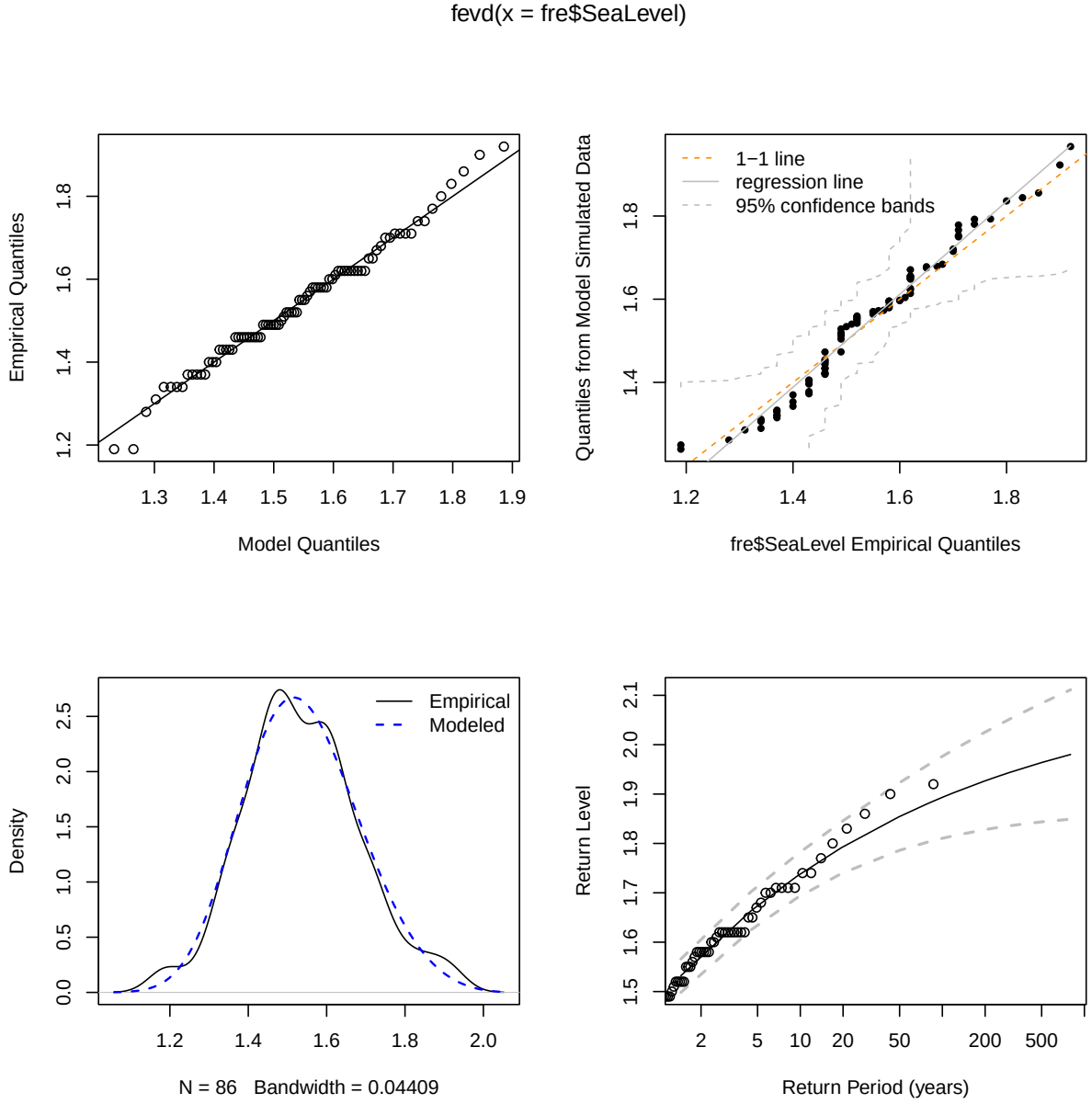
```
fit.3 <- fevd(fre$SeaLevel, data=X, location.fun=~time,scale.fun=~time)
lr.test(fit.1, fit.3) # no support for linear scale trend
```

```
##
##  Likelihood-ratio Test
##
## data:  fre$SeaLevelfre$SeaLevel
## Likelihood-ratio = 1.5808, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.2087
## alternative hypothesis: greater
```

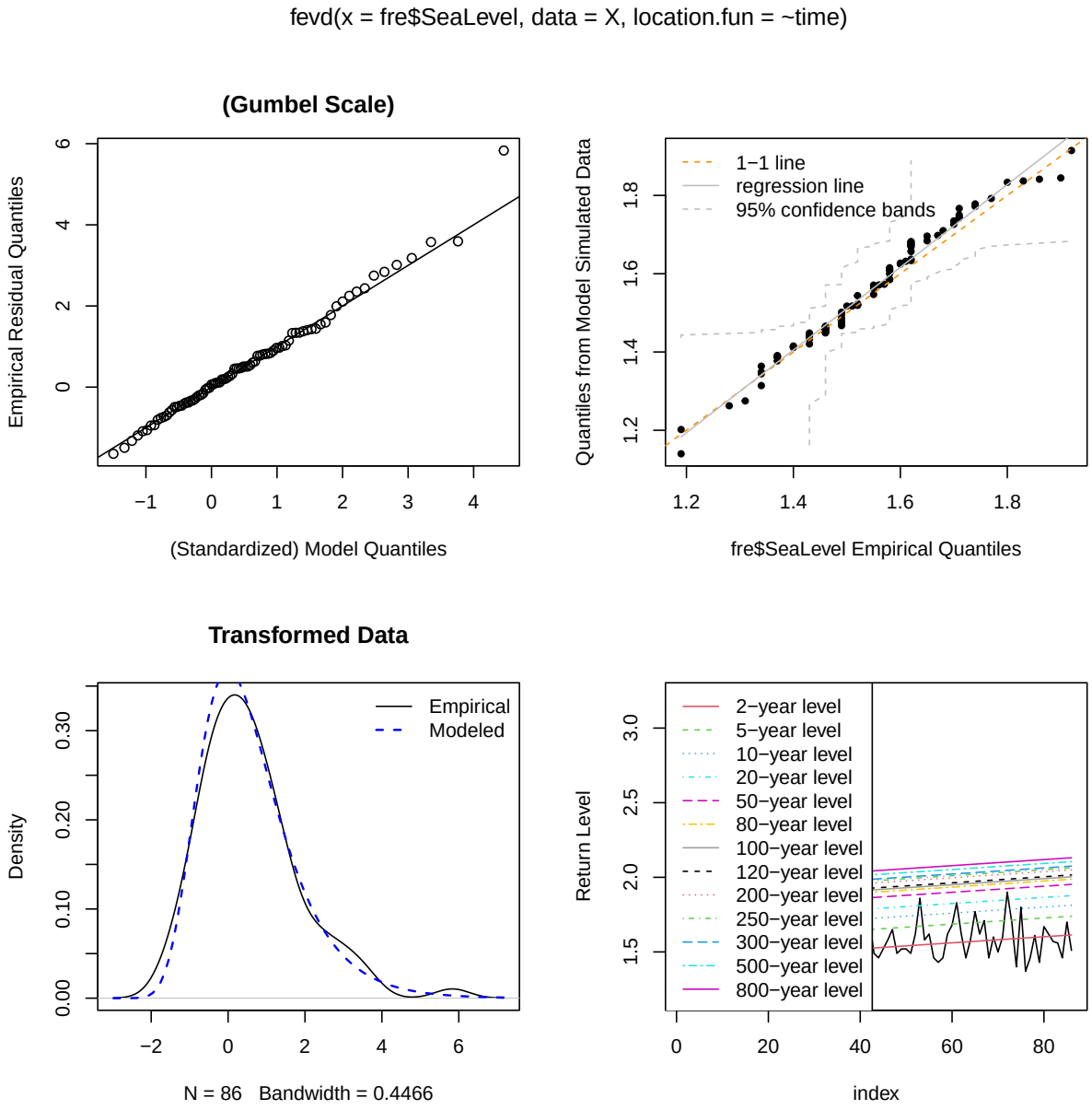
```
fit.4 <- fevd(fre$SeaLevel, data=X, location.fun=~time,shape.fun=~time)
lr.test(fit.1, fit.4) # no support for linear shape trend
```

```
##
##  Likelihood-ratio Test
##
## data:  fre$SeaLevelfre$SeaLevel
## Likelihood-ratio = 0.47067, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.4927
## alternative hypothesis: greater
```

```
plot(fit.const)
```


fevd(x = fre$SeaLevel)
Empirical Quantiles
Model Quantiles
Quantiles from Model Simulated Data
fre$SeaLevel Empirical Quantiles
1−1 line
regression line
95% confidence bands
Density
Empirical
Modeled
N = 86 Bandwidth = 0.04409
Return Level
Return Period (years)

```
plot(fit.1)
```

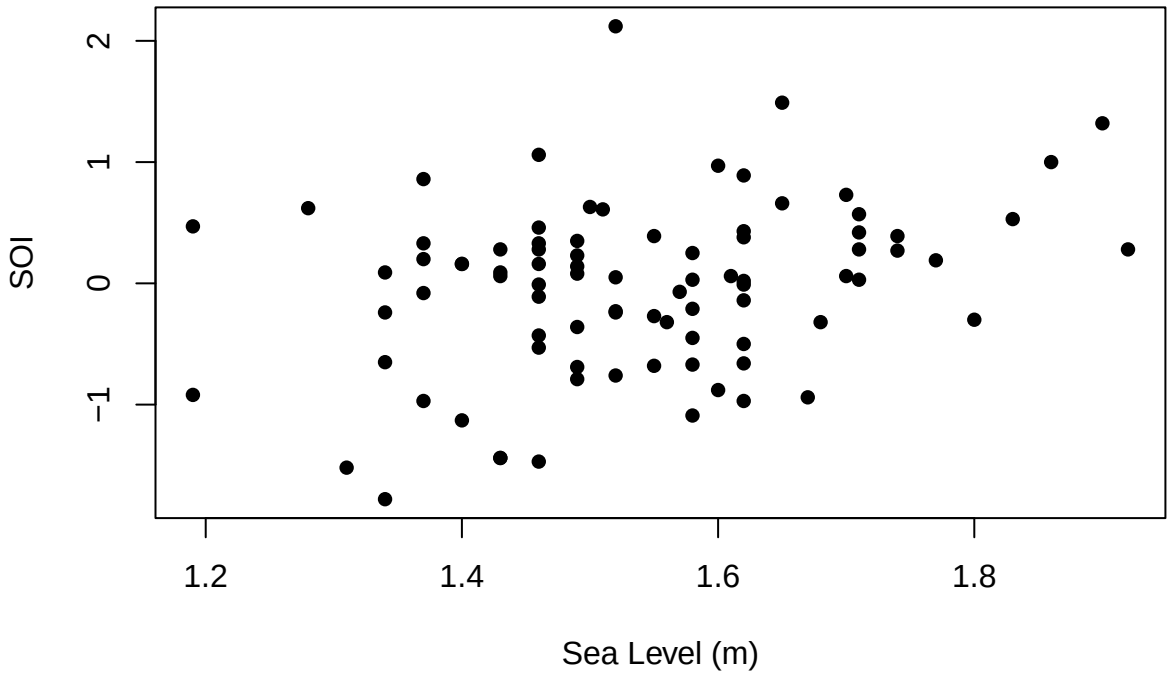


Using forward selection and likelihood ratio tests, there is support for a linear (but not quadratic) trend in μ, and constant σ and ξ. From the qq-plots of the constant and linear-μ model fits, the latter appears to be the best fit, while noting the potential outlier. (Note that the return level diagnostic output has changed to reflect that the parameters of the GEV are now changing over time.)

c) An alternative explanation to the non-stationarity in these data is that the annual maximum sea level is influenced by abnormal large scale meteorological activity. The SOI is a proxy for certain aspects of this activity, such as El Niño. Plot the annual maximum sea levels against the annual mean SOI – is there any visual relationship?

Solution:

```
plot(fre$SeaLevel, fre$SOI, pch=16, xlab="Sea Level (m)", ylab="SOI")
```



While it does not appear to be a strong relationship, it is credible that there is a weak positive linear relationship between annual maximum sea levels and SOI.

d) Take the best model for $\mu(t)$ that you found above and include SOI(t) as an additional predictor. Does the SOI contribute significantly?

Solution:

```
X <- data.frame(time=fre$Year-mean(fre$Year), SOI=fre$SOI)
fit.5 <- fevd(fre$SeaLevel, data=X, location.fun=~time+SOI)
lr.test(fit.1, fit.5) # support for loc linear in SOI
```

```
##
##  Likelihood-ratio Test
##
## data:  fre$SeaLevelfre$SeaLevel
## Likelihood-ratio = 7.9721, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.00475
## alternative hypothesis: greater
```

```
fit.6 <- fevd(fre$SeaLevel, data=X, location.fun=~time+SOI, scale.fun=~SOI)
lr.test(fit.5, fit.6) # support for scale linear in SOI
```

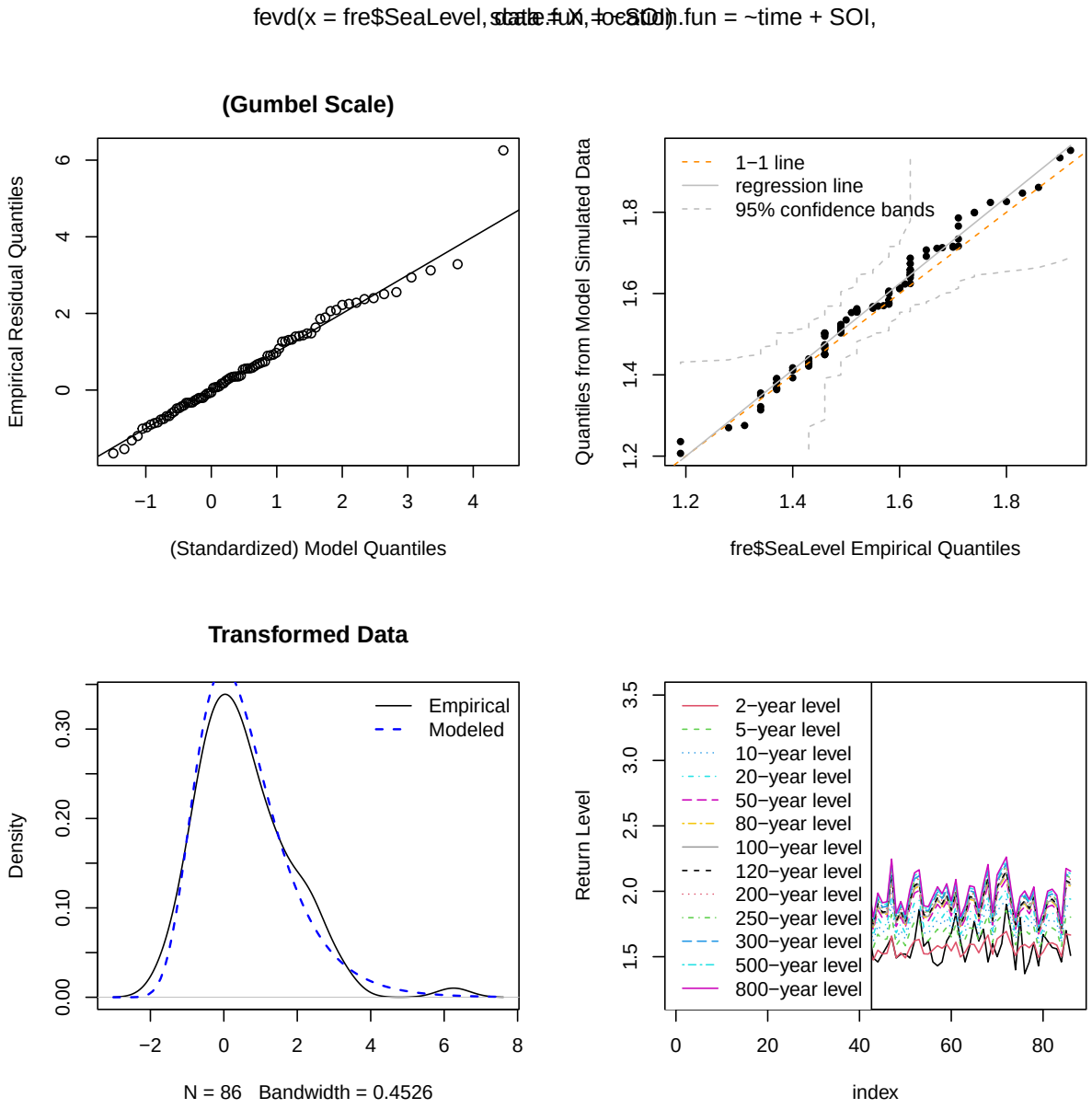
```
##
##  Likelihood-ratio Test
##
## data:  fre$SeaLevelfre$SeaLevel
## Likelihood-ratio = 5.2074, chi-square critical value = 3.8415, alpha =
```

```
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.02249
## alternative hypothesis: greater
```

```
fit.7 <- fevd(fre$SeaLevel, data=X, location.fun=~time+SOI, scale.fun=~SOI,
              shape.fun=~SOI)
lr.test(fit.6, fit.7) # no support for shape linear in SOI
```

```
##
##  Likelihood-ratio Test
##
## data:  fre$SeaLevelfre$SeaLevel
## Likelihood-ratio = 0.78313, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.3762
## alternative hypothesis: greater
```

```
plot(fit.6)
```



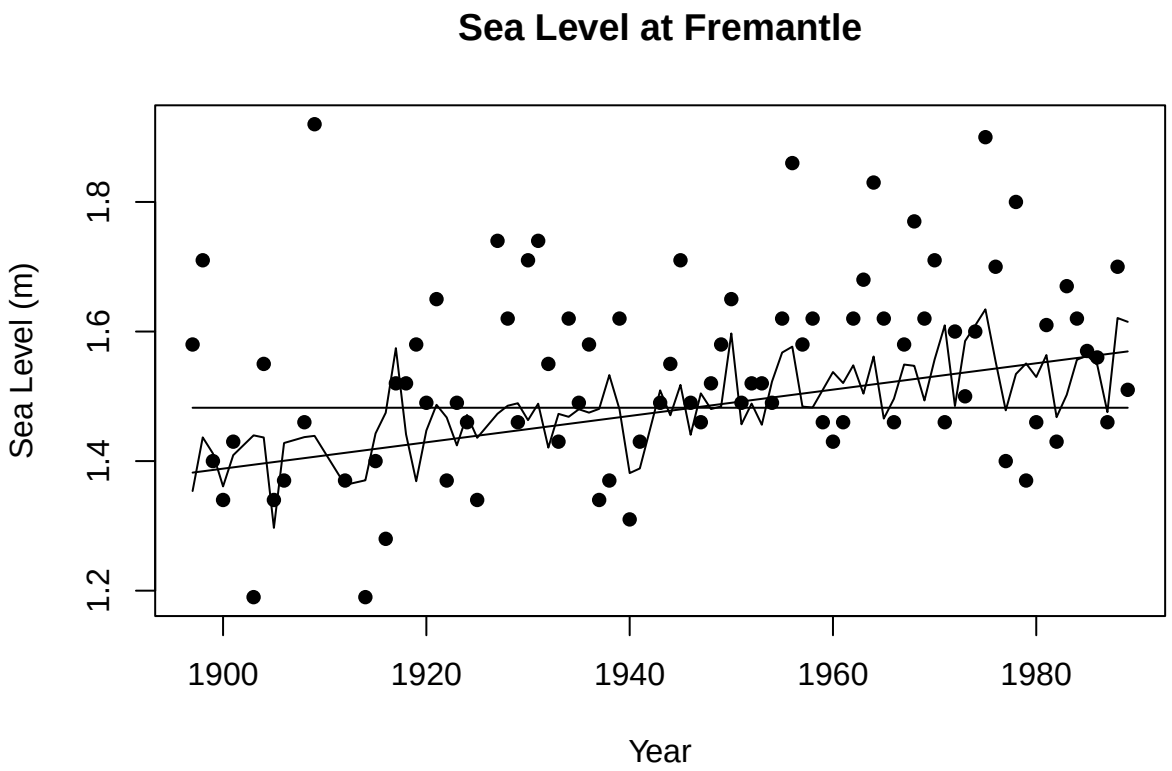
Again using forward selection and likelihood ratio tests, the best performing model is:

$$\begin{aligned} \mu(t) &= \mu_0 + \mu_1(\text{Year}_t - \overline{Year}) + \mu_2 \text{SOI}_t \\ \sigma(t) &= \sigma_0 + \sigma_1 \text{SOI}_t \\ \xi(t) &= \xi_0. \end{aligned}$$

Superimposing $\mu(t)$ on the plot of the data for the constant, linear in time, and linear in time and SOI models, we have:

```
plot(fre$Year,fre$SeaLevel, pch=16, xlab="Year", ylab="Sea Level (m)",
     main="Sea Level at Fremantle")
p <- fit.const$results$par
r <- range(fre$Year)
lines(r, rep(p[1],2))
p2 <- fit.1$results$par
```

```
lines(r, p2[1]+p2[2]*(r-mean(fre$Year)))
p3 <- fit.6$results$par
lines(fre$Year, p3[1]+p3[2]*(fre$Year-mean(fre$Year))+p3[3]*fre$SOI)
```



The linear trend in the location parameter seems clearly more appropriate than a constant value. When SOI is included it seems visually as though a reduction/increase in location for low/high SOI years is broadly supported by the data.

e) How much does the SOI contribute to the annual maximum sea levels per unit of SOI? Is this greater or less than any remaining time trend?

Solution:

```
fit.6$results$par
```

```
##          mu0          mu1          mu2       sigma0       sigma1        shape
##  1.490633739  0.001931625  0.065188483  0.123444535  0.033521530 -0.194131566
```

The MLEs for the fitted model include: $\hat{\mu}_0 = 1.490633739, \hat{\mu}_1 = 0.001931625$ and $\hat{\mu}_2 = 0.065188483$. This means that sea levels rise in Fremantle on average 6.5cm for every unit increase in SOI. This is above and beyond any increase in sea levels due to longer term trend, which corresponds to around 2mm per year.

Exercise 3

The files `100m-women.csv` and `100m-men.csv` record the times of the fastest official 100 metres race times in the world from 1972–2021. These data can be read in via e.g.

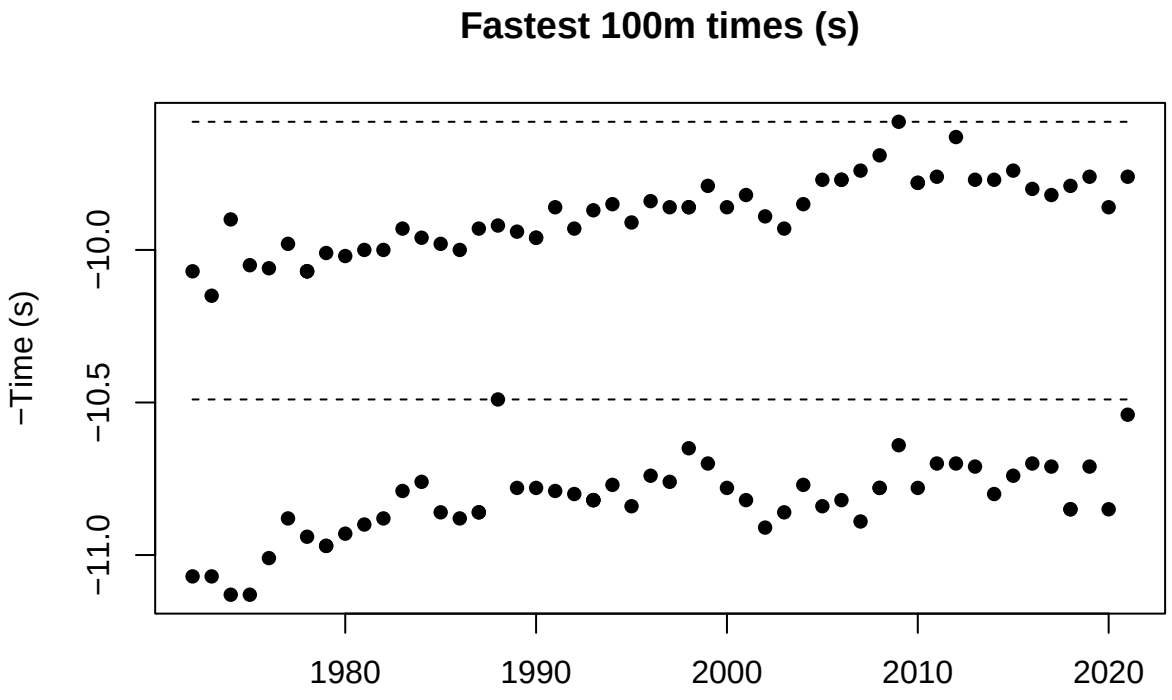
```
read.csv("100m-women.csv", header=T, sep=",")
```

The current world records were set during these years for both women and men. We are interested in whether the world record will be broken in future years.

a) Plot these data. Do they look stationary?

Solution:

```
women <- read.csv("../data/100m-women.csv",header=T,sep=",")
men <- read.csv("../data/100m-men.csv"  ,header=T,sep=",")
all <- c(women$Time,men$Time)
plot(women$Year, -women$Time, ylim=c(-max(all),-min(all)),xlab="",
    ylab="-Time (s)", pch=16, main="Fastest 100m times (s)")
points(men$Year, -men$Time, pch=16)
wrec <- min(women$Time)
mrec <- min(men$Time)
lines(range(women$Year), -rep(wrec,2), lty=2)
lines(range(men$Year), -rep(mrec,2), lty=2)
```



There appears to be a clear increasing trend in negative race times (that is, race times seem to be getting lower over time). So we might expect the GEV location parameter to vary over time. There may be some visual evidence of an increase in variability over time. Note that there does appear to be some temporal dependence of successive observations. The GEV limits hold with assumptions of no, or weak dependence, so we need to be wary here of whether we are violating any assumptions for our models to hold.

b) For the moment, let's focus on the men's dataset (see below for women's). Explore fitting a GEV model (for *negative* times) that has parameters that are constant, linear, quadratic or

cubic in time.

Solution:

```
Xm <- data.frame(year=men$Year-mean(men$Year),
                 year2=(men$Year-mean(men$Year))^2,
                 year3=(men$Year-mean(men$Year))^3)
fit.1 <- fevd(-men$Time)
fit.2 <- fevd(-men$Time,data=Xm, location.fun=~year)
lr.test(fit.1, fit.2) # support for loc linear in time
```

```
##
##  Likelihood-ratio Test
##
## data:  -men$Time-men$Time
## Likelihood-ratio = 78.116, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
fit.3 <- fevd(-men$Time,data=Xm, location.fun=~year+year2)
lr.test(fit.2, fit.3) # support for loc quadratic in time
```

```
##
##  Likelihood-ratio Test
##
## data:  -men$Time-men$Time
## Likelihood-ratio = 13.648, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.0002205
## alternative hypothesis: greater
```

```
fit.4 <- fevd(-men$Time,data=Xm, location.fun=~year+year2+year3)
lr.test(fit.3, fit.4) # no support for loc cubic in time
```

```
##
##  Likelihood-ratio Test
##
## data:  -men$Time-men$Time
## Likelihood-ratio = 2.2774, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.1313
## alternative hypothesis: greater
```

```
fit.5 <- fevd(-men$Time,data=Xm, location.fun=~year+year2, scale.fun=~year)
lr.test(fit.4, fit.5) # support for scale linear in time
```

```
##
##  Likelihood-ratio Test
##
## data:  -men$Time-men$Time
## Likelihood-ratio = 0.12867, chi-square critical value = 0.00, alpha =
## 0.05, Degrees of Freedom = 0.00, p-value < 2.2e-16
```

```
## alternative hypothesis: greater
```

```
fit.6 <- fevd(-men$Time,data=Xm, location.fun=~year+year2,
              scale.fun=~year+year2)
lr.test(fit.5, fit.6) # no real support for scale quadratic in time
```

```
##
##  Likelihood-ratio Test
##
## data:  -men$Time-men$Time
## Likelihood-ratio = 3.2664, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.07071
## alternative hypothesis: greater
```

```
fit.7 <- fevd(-men$Time,data=Xm, location.fun=~year+year2,
              scale.fun=~year,shape.fun=~year)
lr.test(fit.5, fit.7) # no support for shape linear in time
```

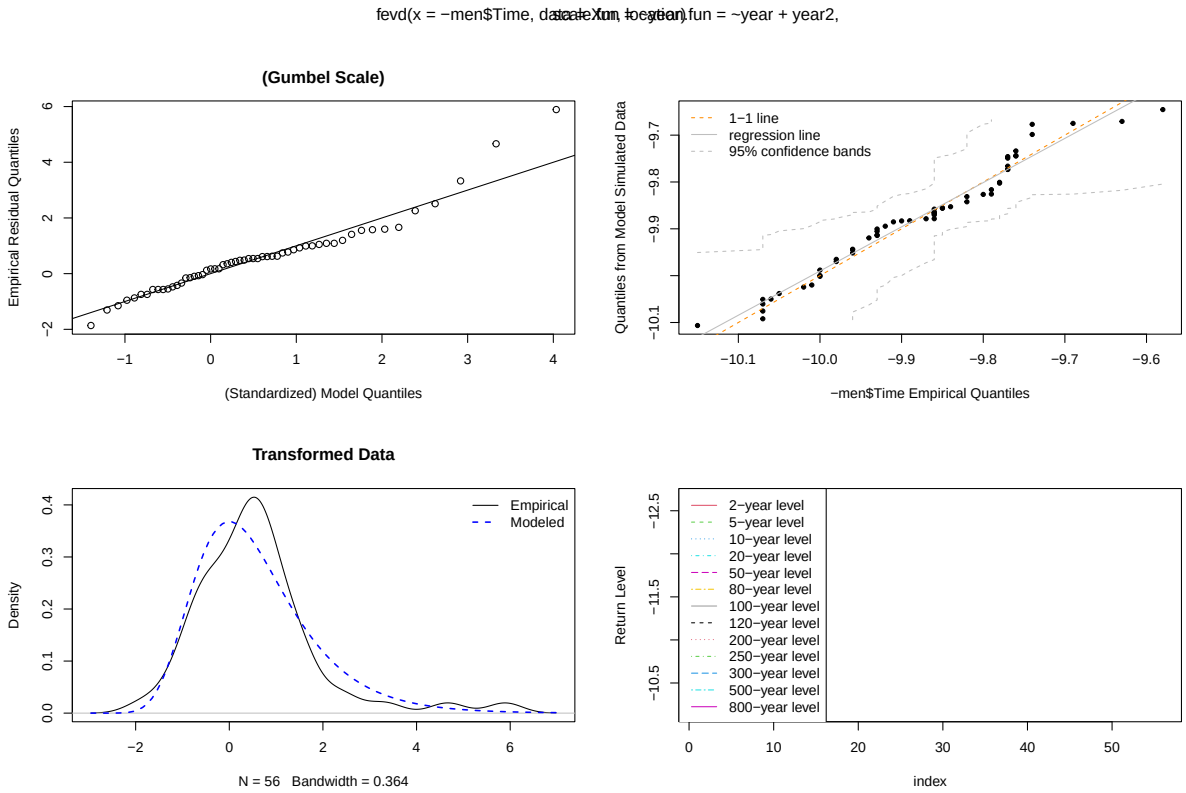
```
##
##  Likelihood-ratio Test
##
## data:  -men$Time-men$Time
## Likelihood-ratio = 1.7079, chi-square critical value = 3.8415, alpha =
## 0.0500, Degrees of Freedom = 1.0000, p-value = 0.1913
## alternative hypothesis: greater
```

Forward selection and likelihood ratio tests support a quadratic trend in $\mu(t)$, a linear trend in $\sigma(t)$ (with a p-value of 0.07 for inclusion of a quadratic term), and a constant value of ξ.

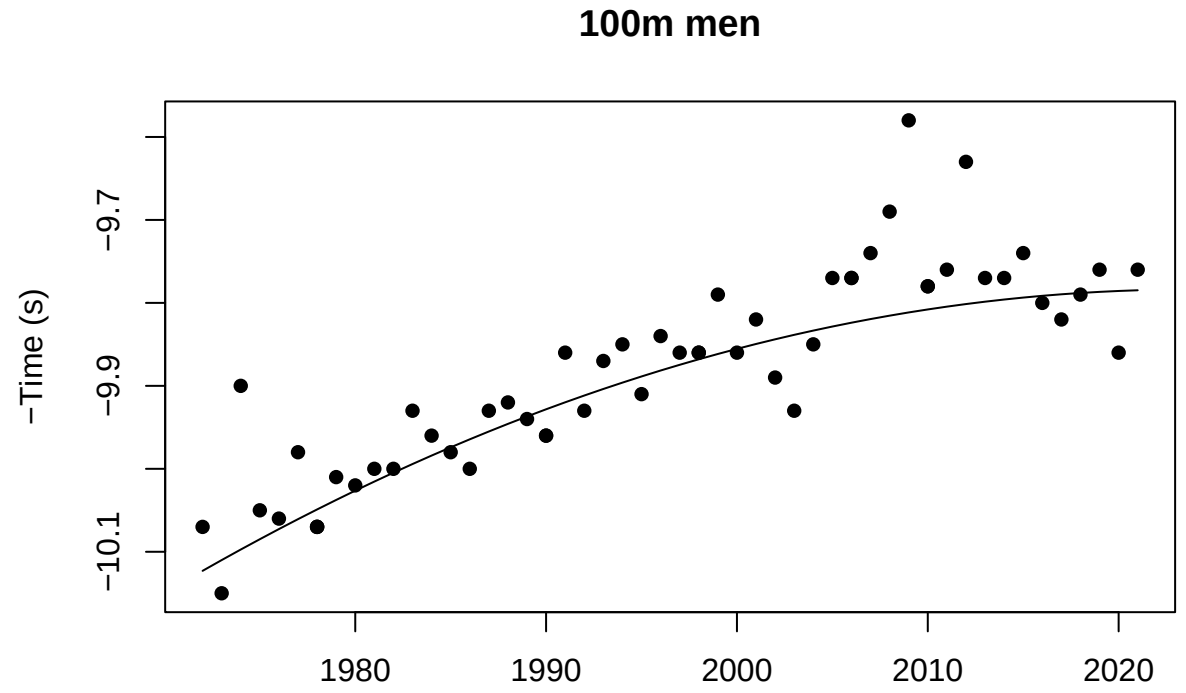
c) Assess goodness-of-fit diagnostics for your chosen model, and superimpose your model for $\mu(t)$ on a plot of the data.

Solution:

```
plot(fit.5)
```




```
p <- fit.5$results$par
plot(men$Year, -men$Time, pch=16, xlab="", ylab="-Time (s)", main="100m men")
lines(men$Year, p[1] + p[2]*Xm$year + p[3]*Xm$year2)
```



The fit for the best model is not perfect, but it's not unreasonable. The fitted trend for $\mu(t)$ appears to follow the direction of the data fairly well. Note that as with any statistical model,

whether $\mu(t)$ and $\sigma(t)$ can be reliably extrapolated into the future is something that needs careful consideration.

d) What is the probability that:

- the world record will be broken in 2022?

  Solution:

```
pyear <- 2022:2041
ploc <- p[1]+p[2]*(pyear-mean(men$Year))+p[3]*(pyear-mean(men$Year))^2
pscale <- p[4]+p[5]*(pyear-mean(men$Year))
pr <- rep(NA,length=20)
for (i in 1:length(pyear)) {
  pr[i]=pevd(-mrec,loc=ploc[i],scale=pscale[i],shape=p[6],type="GEV")
}
1-pr[1] # Prob of world record in 2022
```

```
## [1] 0.02563161
```

  The probability of breaking the world record in 2022 is 0.025.

- the world record will hold for the next 20 years (2022-2041, inclusive)?

  Solution:

```
prod(pr) # Prob of no world record in next 20 years
```

```
## [1] 0.5692057
```

  The probability of the world record not being broken in the next 20 years is 0.569 (assuming independence of years, and correct GEM parameter extrapolation).

e) Repeat your analysis for the women's dataset.

  Note: The women's dataset seems to be causing some problems to fit even a basic GEV. If anyone can figure out why this is, please let us know!

f) Which world record (women's or men's) is more likely to be broken in 2022, or in the next 20 years?