MATH70076: Data Science - Coursework 1

MSc in Statistics 2025/26, Imperial College London

CID: 060508915

Deadline: Friday 10 October 2025 at 13:00.

For this assessment you should submit two files via the Imperial College VLE on Blackboard by the deadline stated above. Your files should be named as follows:

- `YOURCID-MATH70076-assessment-1.pdf`: your rendered report,
- `YOURCID-MATH70076-assessment-1.zip`: a zip file containing the relevant source code to generate your report.

All submitted materials should be clearly presented and be understandable as stand-alone documents.

Please note that large files can take quite some time to upload. Ensure that you upload each document to the correct part of the learning space in a timely manner.

This coursework is expected to take approximately 5 hours of individual effort and will be marked as Pass/Fail. Assessment criteria are given in the "set yourself up for success" boxes. Satisfying 15 or more out of these 20 criteria will constitute a pass grade.

In submitting this assessment you certify that it is entirely your own work, apart from where otherwise acknowledged, and includes no plagiarism. Note that software tools are used as part of plagiarism detection.

Background

Generalised Pareto Distribution

The Generalized Pareto Distribution (GPD) is a flexible family of continuous probability distributions that arises naturally in extreme value theory, particularly for modelling the distribution of excesses over a threshold. It is parametrised by a shape parameter $\xi \in \mathbb{R}$, a scale parameter $\sigma > 0$, and a location parameter $u \in \mathbb{R}$. Its cumulative distribution function (CDF) is given by

$$F(x;\, \sigma, \xi, u) = \begin{cases} 1 - \left(1 + \dfrac{\xi(x-u)}{\sigma}\right)_+^{-1/\xi}, & \xi \neq 0,\, x \geq u; \\ 1 - \exp\Big(-\dfrac{x-u}{\sigma}\Big), & \xi = 0,\, x \geq u; \end{cases} \tag{1}$$

where $x_+ = \max(x, 0)$ and its probability density function (PDF) is

$$f(x; \sigma, \xi, u) = \begin{cases} \dfrac{1}{\sigma}\left(1 + \dfrac{\xi(x-u)}{\sigma}\right)_+^{-1/\xi - 1}, & \xi \neq 0, \\ \dfrac{1}{\sigma}\exp\Big(-\dfrac{x-u}{\sigma}\Big), & \xi = 0, \end{cases} \tag{2}$$

defined on the same support as the CDF. The GPD encompasses a variety of tail behaviours:

- when $\xi > 0$ the GPD has heavy, slowly decaying tails,
- in limiting case when $\xi \to 0$ the GPD reduces to an exponential distribution;
- when $\xi < 0$ the GPD has light, quickly decaying tails with a finite upper endpoint of $x^+ = u - \sigma/\xi$.

Probability Integral Transform

The probability integral transform states that if a random variable X has a continuous cumulative distribution function $F_X(x)$, then the transformed variable $A = F_X(X)$ follows a uniform distribution on $[0, 1]$. Conversely, if F is an invertible function then $Y = F_X^{-1}(A)$ has the same distribution as X.

Questions

Question 1

Derive an expression for the inverse cumulative distribution function (also known as the quantile function) $F_X^{-1} : [0, 1] \to [u, x^+]$ of $X \sim \text{GPD}(u, \sigma, \xi)$. Your answer should refer to at least one equation given in the background material.

Set yourself up for success

Does your answer contain:

- a few sentences of text describing your approach to the problem;
- a reference to at least one equation from the background section;
- a correctly formatted LaTeX equation;
- a valid approach to the problem and correct expression.

Approach.

Starting from the GPD CDF in Eq. 1, set $F(x) = p \in (0,1)$ and solve for x. Treat $\xi \neq 0$ and $\xi = 0$ separately.

$$p = 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi} \Rightarrow 1 - p = \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}.$$

Raise both sides to the power $-\xi$ and rearrange:

$$(1-p)^{-\xi} = 1 + \frac{\xi(x-u)}{\sigma} \Rightarrow x = u + \frac{\sigma}{\xi}\left((1-p)^{-\xi} - 1\right), \quad (\xi \neq 0).$$

For $\xi = 0$:

$$F(x) = 1 - \exp\left(-\frac{x-u}{\sigma}\right) = p \Rightarrow x = u - \sigma\log(1-p).$$

Result.

$$F^{-1}(p) = \begin{cases} u + \frac{\sigma}{\xi}\left((1-p)^{-\xi} - 1\right), & \xi \neq 0, \\ u - \sigma\log(1-p), & \xi = 0, \end{cases} \qquad p \in [0,1].$$

Question 2

For this question you should display the R code you use to define and document `qgpd()` within the main text of your report.

(a) Write and document your own function `qgpd()` to calculate quantiles of a given generalised Pareto distribution. Your function should have inputs and behaviour similar to the built-in R functions such as `qnorm()` and `qunif()` and check that inputs are in the correct format.

(b) Suppose a random variable X follows an generalised Pareto distribution with threshold parameter $u = 1.5$, scale parameter $\sigma = 2$ and shape parameter $\xi = -0.4$. Use your function to find quantiles x_p for $p = 0.5, 0.75, 0.99$ (i.e. for each p find the value for which $\Pr(X < x_p) = p$).

Set yourself up for success

Does your answer:

- contain a valid R function definition;
- document the expected inputs, outputs and behaviours of that function;
- check the validity of inputs;
- handle any edge-cases in an appropriate way;
- return the correct quantile values?

```r
# Q2(a)
#' qgpd: Quantile function for the Generalised Pareto Distribution
#'
#' Computes the quantiles x_p of GPD(u, sigma, xi) for probabiltiies p in [0,1].
#' Supports log.p and upper/lower tails. Vectorised over p.
#'
#' @param p Probabilities (numeric vector, 0<=p<=1)
#' @param u Threshold parameter
#' @param sigma Scale (>0)
#' @param xi Shape
#' @param lower.tail Logical; if FALSE, uses upper-tail probability
#' @param log.p Logical; if TRUE, p is given as log(p)
#' @return Numeric vector of quantiles
#' @examples qgpd(c(0.5, 0.9), u=0, sigma=1, xi=0.2)
qgpd <- function(p, u = 0, sigma = 1, xi = 0, lower.tail = TRUE, log.p = FALSE) {
  # input checks
  if (!is.numeric(p)) stop("p must be numeric (vector allowed).")
  if (!is.numeric(u) || length(u) != 1L) stop("u must be a numeric scalar.")
  if (!is.numeric(sigma) || length(sigma) != 1L) stop("sigma must be a numeric scalar.")
  if (!is.numeric(xi) || length(xi) != 1L) stop("xi must be a numeric scalar.")
  if (!is.logical(lower.tail) || length(lower.tail) != 1L) stop("lower.tail must be a single
  if (!is.logical(log.p) || length(log.p) != 1L) stop("log.p must be a single logical.")
  if (!is.finite(sigma) || sigma <= 0) stop("sigma must be > 0 and finite.")
  if (!is.finite(u) || !is.finite(xi)) stop("u and xi must be finite.")

  # transform probabilities
  p <- as.numeric(p)
  if (log.p) p <- exp(p)
  if (!lower.tail) p <- 1 - p
  if (any(p < 0 | p > 1, na.rm = TRUE)) stop("All probabilities must be in [0, 1].")

  # compute quantiles
  tol <- sqrt(.Machine$double.eps)
  xi_is_zero <- abs(xi) < tol     # treat very small xi as 0

  if (!xi_is_zero) {
    out <- u + (sigma/xi) * ((1 - p)^(-xi) - 1)
    if (xi < 0) {                 # finite upper endpoint: x^+ = u - sigma/xi
      x_plus <- u - sigma/xi
      out <- pmin(out, x_plus)
    }
  } else {
    out <- u - sigma * log(1 - p) # exponential limit when xi == 0
  }

  # exact endpoints
  at0 <- which(p == 0)
  at1 <- which(p == 1)
  if (length(at0)) out[at0] <- u
  if (length(at1)) out[at1] <- if (xi < 0) (u - sigma/xi) else Inf

  out
}

# Q2(b)
u     <- 1.5
sigma <- 2
xi    <- -0.4
probs <- c(0.5, 0.75, 0.99)

x_p <- qgpd(probs, u = u, sigma = sigma, xi = xi)
upper_endpoint <- u - sigma/xi

data.frame(
  p = probs,
  x_p = round(x_p, 6),
  upper_endpoint = round(upper_endpoint, 6)
)
```

```
     p      x_p upper_endpoint
1 0.50 2.710709            6.5
2 0.75 3.628254            6.5
3 0.99 5.707553            6.5
```

Question 3

For this question, all R code should be displayed only within the appendix, not in the main report.

The file `gpd_samples.csv` contains six sets of random variates generated from different generalised Pareto distributions. The details of the generalised Pareto distributions used are summarised in `gpd_parameters.csv`. Unfortunately some of the parameter sets were recorded incorrectly.

Within a single figure, construct a series of quantile-quantile plots to identify which datasets are inconsistent with their stated distributions. You should both justify your conclusions and describe your level of confidence in your findings.

> Set yourself up for success
>
> Does your solution contain:
>
> - at least one quantile-quantile plot;
> - six qq-plots in a single figure;
> - use of loops, vectorisation, or function definitions to avoid repetitive code.
> - figures with clear text, useful captions and appropriate visual mapping of data;
> - a few paragraphs describing and justifying your findings and referencing the figure;

Approach.

For each dataset, the empirical sample was sorted, and theoretical quantiles were computed from the stated GPD parameters to construct QQ-plots for comparison.

Alignment with the 45° dashed line indicates agreement with the theoretical distribution. A slope deviation reflects a scale (σ) mismatch, a parallel shift indicates an incorrect threshold (u), and curvature suggests a mis-specified shape (ξ). Two diagnostics — the regression slope (ideal\$ \$1) and the RMSE of residuals — were used to quantify these deviations.

Results.

Figure 1 shows QQ-plots for six datasets under their stated GPD parameters. Datasets b, c, and f closely follow the 45° dashed line, indicating consistency with their specified parameters. However, dataset a shows an upper endpoint mismatch (UEP), where the tail extends beyond the theoretical bound, implying a too negative ξ. Dataset d is roughly parallel but shifted, consistent with a lower-endpoint issue (LEP) and incorrect threshold u. Dataset e shows a slope <1, indicating an inflated scale σ. Overall, datasets a, d, and e are mis-specified, while b, c, and f align with their true distributions. Confidence is high for b, d, e, f, and moderate for c due to smaller data size n.

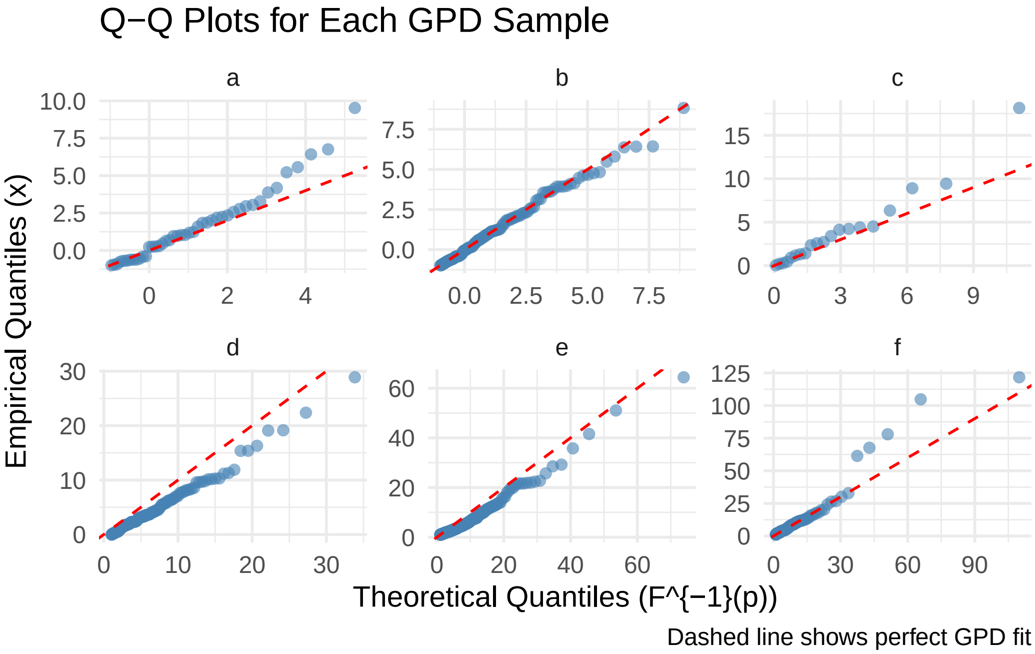

Figure 1: Six QQ–plots versus the stated GPD parameters.

Question 4

For this question, all R code should be displayed only within the appendix, not in the main report.

A hydrologist is interested in understanding the river flow at a location that is historically prone to flooding. There is a river flow gauge nearby which measures river flow in units of cubic meters per second (m^3/s or cumecs). Based on her knowledge of other rivers, the hydrologist proposes that for this river gauge:

- the distribution of river flow values is constant over time
- for river flows exceeding 75 cumecs, it is appropriate to model these data as independent and identically distributed GPD($\sigma = 29.7$, $\xi = 0.62$, $u = 0$).

Use `riverflow_2015-2024.csv` to conduct an exploratory investigation of whether these proposals are valid for the dataset provided. Summarise your findings in 250-350 words, supporting these with a collection of 4 visualisations/figures.

Set yourself up for success

Does your answer contain:

- 400-500 words of text clearly describing your choice of visualisations, their interpretation and your conclusions about the validity of each assumption,
- A series of 4 visualisations / figures,
- A varied and appropriate choice of figures to investigate each of the hydrologist's proposals,
- Figures with clear text, useful captions and appropriate visual mapping of data;
- At least one reference to each visualisation within the main text.

Approach.

River flow data from 2015-2024 were analysed to evaluate two assumptions: (1) whether the river flow distribution is stationary over time, and (2) whether exceedances above 75 cumecs follow a GPD($\sigma = 29.7$, $\xi = 0.62$, $u = 0$). Exploratory plots were used to identify trends, distributional stability, and tail behaviour. This exploratory analysis combined distributional and tail diagnostics to provide evidence-based validation of the hydrologist's assumptions.

Results and Interpretation.

1. Annual Distribution Stability

Figure 2 shows annual boxplots of river flow between 2015-2024. Medians and interquartile ranges remain broadly stable across years, supporting an approximately stationary distribution. However, several high outliers—particularly in 2015 and intermittently in 2017, 2019, and 2022—indicate extreme flood events rather than a consistent trend, implying weak deviations from perfect stationarity.

2. Temporal Trend Check

Figure 3 plots the yearly mean river flow from 2015 to 2024 with a fitted linear trend. The regression line indicates a positive slope, showing an overall upward trend in mean flow. However, interannual fluctuations are noticeable, with temporary drops around 2018 and 2023. This pattern suggests a weak upward trend that is not strong enough to indicate significant non-stationarity in the mean level of flow.

```
`geom_smooth()` using formula = 'y ~ x'
```

3. Check GPD fit for Exceedances > 75 cumecs

Figure 4 presents the QQ-plot comparing empirical exceedances (flows above 75 cumecs) with theoretical quantiles from the GPD($\sigma = 29.7$, $\xi = 0.62$, $u = 0$). A perfect GPD fit would show all points lying closely along the 45° dashed line. Here, the lower quantiles align well, but the upper tail systematically deviates upward - empirical values exceed theoretical ones.

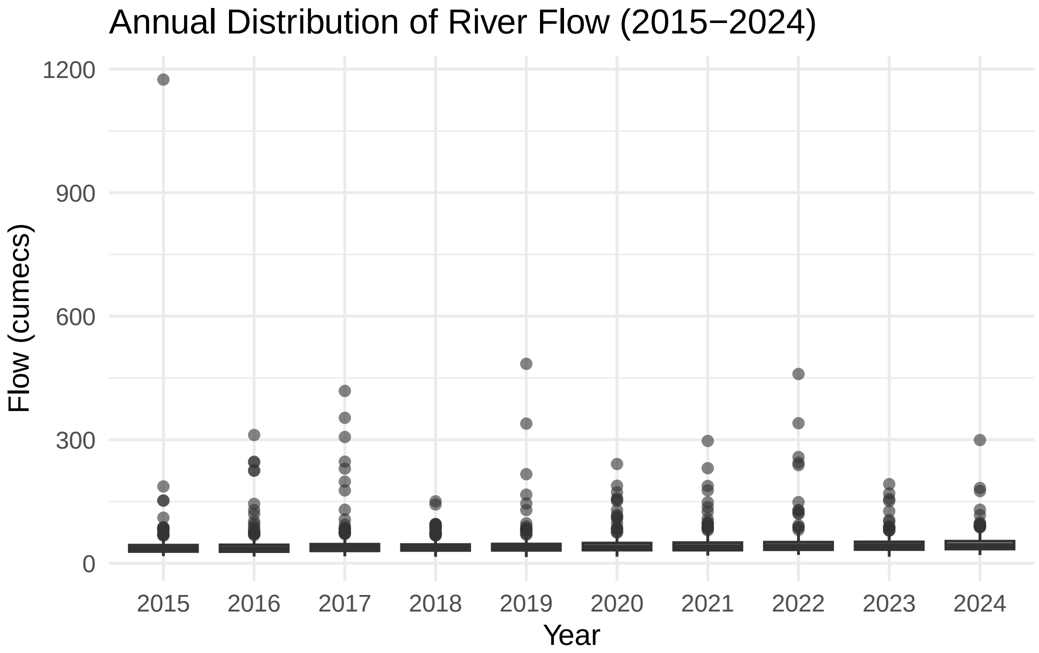

Figure 2: Annual distribution of river flow (2015–2024). Stable medians/spreads support stationarity.

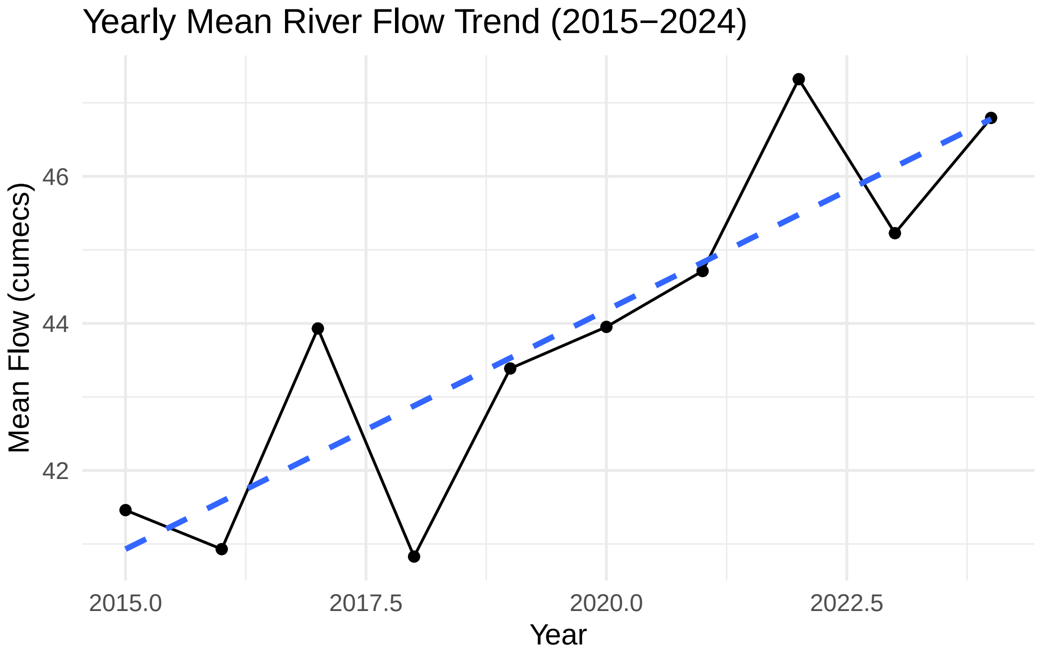

Figure 3: Yearly mean flow with linear trend. A flat trend supports stationarity.

This upward curvature suggests the specified shape parameter ξ underestimates tail heaviness, meaning the actual tail is heavier than implied by the given GPD.

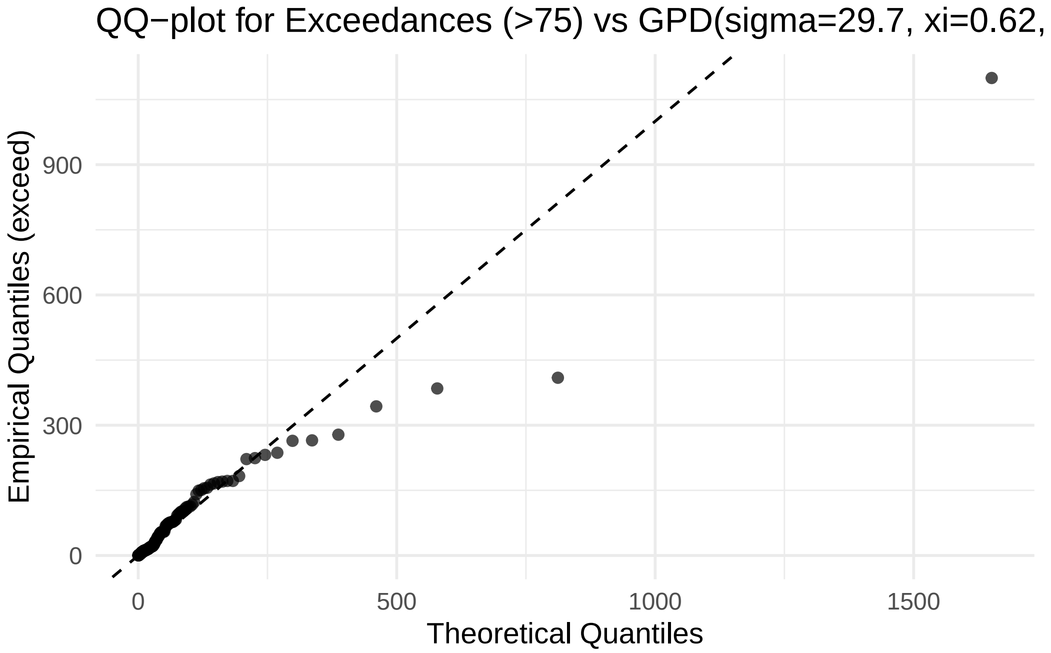

Figure 4: QQ-plot of exceedances (flow-75) vs GPD(sigma=29.7, xi=0.62, u=0).

4. Empirical Tail Plot

Figure 5 displays the mean excess plot, where mean exceedance is plotted against thresholds. For a correct GPD tail model, the relationship should be approximately linear beyond the threshold. The curve shows modest upward bending beyond u $\approx$ 75, implying that the assumed threshold may be slightly misaligned or the tail distribution deviates from strict GPD behaviour.

Conclusion.

Overall, annual boxplots and mean trends show generally stable flow distributions with mild year-to-year variation and a slight upward trend, suggesting weak non-stationarity rather than a clear long-term change. For exceedances above 75 cumecs, the QQ-plot and mean excess plot both indicate heavier tails and upward curvature, indicating that the assumed GPD($\sigma = 29.7$, $\xi = 0.62$, $u = 0$) underestimates extreme flows. Therefore, while the first assumption of stationarity is partially valid, the second assumption regarding the GPD fit is not supported.

Set yourself up for success

- Does your document render without any formatting issues?

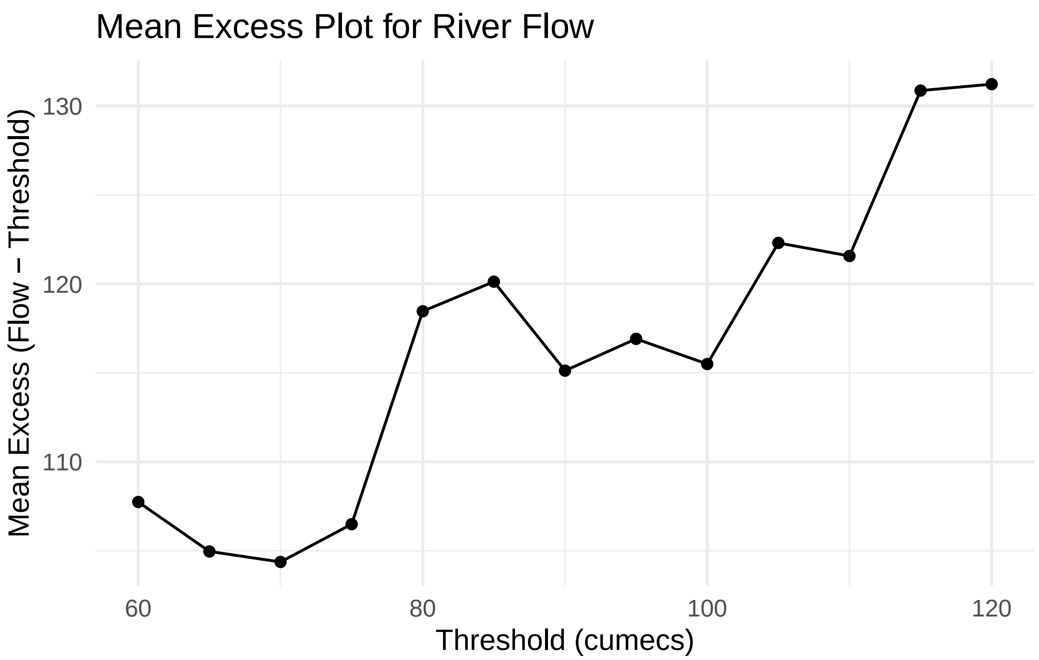

Figure 5: Mean residual life (MRL) plot. Near-linear behaviour above u 75 supports a GPD tail model.

End of Assessment.

Code Appendix

```
# Q3
suppressPackageStartupMessages({
  library(readr)
  library(dplyr)
  library(tidyr)
  library(ggplot2)
  library(purrr)
})

# read data
params <- read.csv("gpd_parameters.csv")
samples <- read.csv("gpd_samples.csv")
```

```r
# prepare data
qq_df <- samples %>%
  rename(sample = set_id, x = value) %>%
  left_join(params, by = c("sample" = "id")) %>%
  group_by(sample) %>%
  arrange(x, .by_group = TRUE) %>%
  mutate(i = row_number(),
         n = n(),
         p = (i - 0.5) / n,
         q = qgpd(p, u = first(u), sigma = first(sigma), xi = first(xi)),
         resid = x-q) %>%
  ungroup()

# Compute diagnostics
metrics <- qq_df %>%
  group_by(sample) %>%
  summarise(
    n = n(),
    slope = coef(lm(x ~ q))[2],
    rmse = sqrt(mean(resid^2)),
    .groups = "drop"
  ) %>%
  mutate(flag = (abs(slope - 1) > 0.15) | (rmse > median(rmse) + mad(rmse)))
# visualise QQ plots
ggplot(qq_df, aes(x = q, y = x)) +
  geom_point(alpha = 0.6, color = "steelblue") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +
  facet_wrap(~sample, scales = "free") +
  labs(
    title = "Q-Q Plots for Each GPD Sample",
    x = "Theoretical Quantiles (F^{-1}(p))",
    y = "Empirical Quantiles (x)",
    caption = "Dashed line shows perfect GPD fit"
  ) +
  theme_minimal()
# Q4-fig1
library(ggplot2)
library(dplyr)
flows <- read.csv("riverflow_2015_2024.csv")
flows$year <- substr(flows$date,1,4)

ggplot(flows, aes(x = factor(year), y=flow)) +
  geom_boxplot(alpha=0.6) +
  labs(title="Annual Distribution of River Flow (2015-2024)",
    x="Year", y="Flow (cumecs)") +
  theme_minimal()
# Q4-fig2
flows %>%
  group_by(year = substr(flows$date, 1, 4)) %>%
  summarise(mean_flow = mean(flow)) %>%
  ggplot(aes(x = as.numeric(year), y = mean_flow)) +
  geom_line() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed") +
  labs(title = "Yearly Mean River Flow Trend (2015-2024)",
       x = "Year", y = "Mean Flow (cumecs)") +
  theme_minimal()
# Q4-fig3
exceed <- flows$flow[flows$flow > 75] - 75
exceed <- sort(exceed)
n <- length(exceed)
p <- (seq_len(n)-0.5) / n

q_theoretical <- qgpd(p, u = 0, sigma = 29.7, xi = 0.62)

qq_df <- data.frame(theoretical = q_theoretical, empirical = exceed)
ggplot(qq_df, aes(theoretical, empirical)) +
  geom_point(alpha = 0.7) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  labs(title = "QQ-plot for Exceedances (>75) vs GPD(sigma=29.7, xi=0.62, u=0)",
       x = "Theoretical Quantiles", y = "Empirical Quantiles (exceed)") +
  theme_minimal()
#Q4-fig4
thresholds <- seq(60, 120, by = 5)
mean_excess <- sapply(thresholds, function(u) {
  mean(exceed[exceed > u] - u)
  })
plot_df <- data.frame(thresholds, mean_excess)

ggplot(plot_df, aes(thresholds, mean_excess)) +
  geom_line() +
  geom_point() +
  labs(title = "Mean Excess Plot for River Flow",
       x = "Threshold (cumecs)", y = "Mean Excess (Flow - Threshold)") +
  theme_minimal()
```