Step-by-step Inference for Extreme Value Theory on Real Data: Block Maxima (GEV) and Peaks Over Threshold (GPD)

with Worked Numerical Calculations

1 Goal and Setup

Let $\{X_t\}_{t=1}^n$ be a real dataset observed over time (e.g., daily rainfall, hourly river flow, daily maximum temperature, financial losses). We want to model *extremes* and estimate tail quantities such as:

- **Return level** z_T: a level exceeded on average once every T time units (e.g., years).
- **Tail probability** $\mathbb{P}(X > x)$ for large x.

EVT provides two standard approaches:

1. **Block Maxima** $\Rightarrow$ fit a **GEV** distribution to maxima in blocks.
2. **Peaks Over Threshold (POT)** $\Rightarrow$ fit a **GPD** distribution to exceedances above a high threshold.

2 Common Pre-Processing Steps (Do This First)

C1. Define the extreme direction. If extremes are large values, proceed with X_t. If extremes are small values, transform e.g. $Y_t = -X_t$ and analyze maxima of Y_t.

C2. Clean the data. Handle missing values, obvious sensor errors, duplicates, unit changes, and outliers due to known measurement faults.

C3. Check (non-)stationarity. Plot X_t over time, seasonal cycles, and potential trends. If strong seasonality/trend exists, consider:

- restricting to homogeneous seasons (e.g., analyze summer only), or
- modeling parameters as functions of covariates (Section 8).

C4. Assess dependence. EVT is simplest under (approximate) independence. Time series often have clustering of extremes (storms, heat waves). If dependence is strong, use declustering (especially for POT; see Section 6).

3 Method A: Block Maxima Inference (GEV)

3.1 Model

Choose a block size m (e.g., monthly blocks or yearly blocks) and form block maxima:

$$M_j = \max\{X_{(j-1)m+1}, \dots, X_{jm}\}, \qquad j = 1, \dots, k, \quad k = \left\lfloor \frac{n}{m} \right\rfloor.$$

Under EVT, for large m, M_j is approximately **GEV**:

$$\mathbb{P}(M \le x) \approx G(x; \mu, \sigma, \xi) = \exp\left\{ -\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-1/\xi} \right\},$$

defined for $1 + \xi(x-\mu)/\sigma > 0$, with parameters:

$$\mu \in \mathbb{R} \quad \text{(location)}, \qquad \sigma > 0 \quad \text{(scale)}, \qquad \xi \in \mathbb{R} \quad \text{(shape)}.$$

For $\xi \to 0$, interpret the limit as the Gumbel case:

$$G(x) = \exp\left\{ -\exp\left(-\frac{x-\mu}{\sigma} \right) \right\}.$$

3.2 Step-by-step workflow on real data

BM1. Choose candidate block sizes. Common choices:

- daily data $\to$ yearly maxima (one maximum per year),
- hourly data $\to$ monthly or seasonal maxima,
- financial daily losses $\to$ monthly maxima (if appropriate).

Trade-off: larger blocks $\Rightarrow$ better EVT approximation but fewer maxima (larger variance).

BM2. Extract block maxima $\{M_j\}_{j=1}^k$ and keep the block timestamps.

BM3. Fit the GEV parameters (μ, σ, ξ). Most common: **Maximum Likelihood Estimation (MLE)** by maximizing

$$\ell(\mu, \sigma, \xi) = \sum_{j=1}^{k} \log g(M_j; \mu, \sigma, \xi),$$

where g is the GEV density (the derivative of G).

BM4. Diagnostics (must do).

- **GEV QQ-plot** of maxima vs fitted quantiles.
- **GEV PP-plot** (empirical vs fitted probabilities).
- **Return level plot** (observed maxima vs fitted return levels).
- Check if changing block size changes $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ drastically.

BM5. Compute return levels. For a return period T in *blocks* (e.g., $T = 100$ years when blocks are years), the return level z_T solves

$$\mathbb{P}(M > z_T) = \frac{1}{T} \quad \Longleftrightarrow \quad G(z_T) = 1 - \frac{1}{T}.$$

Thus,

$$z_T = \begin{cases} \mu + \dfrac{\sigma}{\xi}\left[\left\{ -\log\left(1 - \dfrac{1}{T}\right) \right\}^{-\xi} - 1 \right], & \xi \ne 0, \\ \mu - \sigma \log\left\{ -\log\left(1 - \dfrac{1}{T}\right) \right\}, & \xi = 0. \end{cases}$$

BM6. Quantify uncertainty. Common options:

- **Asymptotic SEs** from the observed information (Hessian of $-\ell$).
- **Profile likelihood** for (ξ) or for z_T.
- **Bootstrap** by resampling blocks (preserves within-block dependence).

4 Method B: Peaks Over Threshold Inference (POT)

4.1 Model

Choose a high threshold u and define exceedances:

$$Y_i = X_{t_i} - u \quad \text{for those } t_i \text{ with } X_{t_i} > u.$$

Let N_u be the number of exceedances.

EVT implies that for high u, the conditional excess distribution is approximately **GPD**:

$$\mathbb{P}(Y \le y \mid X > u) \approx H(y; \beta, \xi) = 1 - \left(1 + \xi \frac{y}{\beta}\right)^{-1/\xi},$$

defined for $y \ge 0$ and $1 + \xi y/\beta > 0$, with $\beta > 0$ and shape ξ. For $\xi \to 0$,

$$H(y) = 1 - \exp\left(-\frac{y}{\beta}\right).$$

4.2 Step-by-step workflow on real data

POT1. Pick candidate thresholds u. Use exploratory tools:

- **Mean Residual Life (MRL) plot:**

  $$e(u) = \mathbb{E}[X - u \mid X > u] \approx \frac{1}{N_u} \sum_{i:X_i > u} (X_i - u).$$

  Look for an approximately linear region.
- **Parameter stability plots:** fit GPD for many u values and look for stable $\hat{\xi}$ and adjusted scale.
- **Exceedance rate:** ensure enough exceedances (rule-of-thumb: at least 50–100, context-dependent).

POT2. Handle dependence / clustering (often necessary). If exceedances cluster in time, decluster first (Section 6).

POT3. Form exceedances $Y_i = X_{t_i} - u$, $i = 1, \ldots, N_u$.

POT4. Fit the GPD parameters (β, ξ) via MLE.

POT5. Diagnostics.

- **GPD QQ-plot** of exceedances vs fitted quantiles.
- **GPD PP-plot**.
- Check threshold sensitivity: small changes in u should not drastically change $\hat{\xi}$.

POT6. Estimate tail probabilities and return levels on original scale. Let λ_u be the exceedance rate per observation:

$$\hat{\lambda}_u = \frac{N_u}{n}.$$

For $x > u$, the tail approximation is

$$\mathbb{P}(X > x) \approx \hat{\lambda}_u \left(1 + \hat{\xi} \frac{x - u}{\hat{\beta}}\right)^{-1/\hat{\xi}}.$$

If observations are made at frequency r per year (e.g. $r = 365$ for daily), and you define a T-year return level z_T by

$$\mathbb{P}(X > z_T) \approx \frac{1}{Tr},$$

then solve:

$$\frac{1}{Tr} \approx \hat{\lambda}_u \left(1 + \hat{\xi}\frac{z_T - u}{\hat{\beta}}\right)^{-1/\hat{\xi}},$$

which yields

$$z_T = \begin{cases} u + \dfrac{\hat{\beta}}{\hat{\xi}}\left[(Tr\,\hat{\lambda}_u)^{\hat{\xi}} - 1\right], & \hat{\xi} \neq 0, \\ u + \hat{\beta}\log(Tr\,\hat{\lambda}_u), & \hat{\xi} = 0. \end{cases}$$

5 Worked Numerical Calculations (Concrete Examples)

This section shows how the formulas look with real numbers. Treat the numbers as an example output from software (MLE).

5.1 Worked Example 3 (100 observed values shown as an $n \times n$ table; BOTH methods)

Let $n = 10$ so we have a 10×10 table (total 100 observations). Assume these are **100 daily observations** (so total sample size is $N = 100$ and $r = 1$ observation/day).

(A) The raw data: 100 numbers in a 10×10 table

Entry in row i, column j equals

$$X_{10(i-1)+j}, \qquad i = 1, \ldots, 10, \quad j = 1, \ldots, 10.$$

Row	1	2	3	4	5	6	7	8	9	10
1	35.8	23.6	22.1	22.1	65.1	42.0	18.6	36.8	30.5	4.7
2	11.9	32.0	61.4	22.2	17.5	28.0	85.0	21.2	25.8	11.4
3	43.4	9.8	8.9	29.6	41.3	29.1	24.0	21.0	18.7	49.5
4	32.2	19.8	92.0	38.4	48.8	46.2	13.8	20.9	18.4	22.9
5	43.2	58.1	32.9	16.2	25.4	64.6	27.6	21.1	22.3	32.8
6	14.2	18.1	45.8	32.2	27.8	47.1	15.5	75.0	31.6	30.9
7	8.2	19.3	23.3	33.8	30.8	24.7	4.7	25.5	27.1	98.3
8	22.7	31.7	51.9	41.7	59.5	8.3	88.0	17.2	7.1	27.2
9	12.9	64.1	14.5	20.7	43.2	9.8	6.7	29.4	30.9	42.4
10	36.4	31.6	15.7	95.0	31.5	15.3	74.8	35.3	12.3	42.5

Table 1: The 100 observations arranged as a 10×10 table (daily data).

(B) Block Maxima (GEV) using block size $m = 10$

Choose block size $m = 10$ days. Since $N = 100$, we have $k = 100/10 = 10$ blocks. Each block corresponds exactly to one table row, so the block maxima are the row-wise maxima:

$$M_j = \max\{X_{10(j-1)+1}, \ldots, X_{10j}\}, \qquad j = 1, \ldots, 10.$$

Block j	M_j
1	65.1
2	85.0
3	49.5
4	92.0
5	64.6
6	75.0
7	98.3
8	88.0
9	64.1
10	95.0

Table 2: Block maxima for $m = 10$ (one maximum per 10-day block).

Assume a GEV fit (from software/MLE) to $\{M_j\}_{j=1}^{10}$ gives:

$$\hat{\mu} = 76, \qquad \hat{\sigma} = 11, \qquad \hat{\xi} = 0.06.$$

Numerical calculation: 100-block return level Here 1 block = 10 days, so 100 blocks = 1000 days. The 100-block return level z_{100} satisfies $G(z_{100}) = 1 - \frac{1}{100}$:

$$z_{100} = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}} \left[\left\{ -\log\left(1 - \frac{1}{100}\right) \right\}^{-\hat{\xi}} - 1 \right].$$

Compute:

$$a = -\log(0.99) \approx 0.0100503, \qquad a^{-\hat{\xi}} = a^{-0.06} = \exp\bigl(-0.06 \ln a\bigr).$$

Since $\ln(0.0100503) \approx -4.59948$:

$$-0.06 \ln a \approx 0.27597, \qquad a^{-0.06} \approx e^{0.27597} \approx 1.3176.$$

Therefore:

$$z_{100} \approx 76 + \frac{11}{0.06}(1.3176 - 1) = 76 + 183.333(0.3176) \approx 76 + 58.274 \approx 134.274,$$

$$\boxed{z_{100} \approx 134.3 \text{ (100 blocks } = 1000 \text{ days)}}$$

(C) POT (GPD) on the same 100 numbers with threshold $u = 70$

Choose threshold $u = 70$. The exceedances ($X_t > 70$) from the table are:

Thus $N_u = 7$ and the exceedance rate is

$$\hat{\lambda}_u = \frac{N_u}{N} = \frac{7}{100} = 0.07.$$

Assume a GPD fit (from software/MLE) to the excesses gives:

$$\hat{\beta} = 10, \qquad \hat{\xi} = 0.20.$$

Index t	X_t	Excess $Y = X_t - u$
17	85.0	15.0
33	92.0	22.0
58	75.0	5.0
70	98.3	28.3
77	88.0	18.0
94	95.0	25.0
97	74.8	4.8

Table 3: POT exceedances above $u = 70$ and excesses $Y = X - u$.

Numerical calculation: 1000-day return level Because the data are daily ($r = 1/\text{day}$), define z_{1000} by

$$\mathbb{P}(X > z_{1000}) \approx \frac{1}{1000}.$$

Using the POT return-level formula:

$$z_{1000} = u + \frac{\hat{\beta}}{\hat{\xi}} \left[(1000\hat{\lambda}_u)^{\hat{\xi}} - 1 \right].$$

Compute $1000\hat{\lambda}_u = 1000(0.07) = 70$:

$$z_{1000} = 70 + \frac{10}{0.20} \left(70^{0.20} - 1\right) = 70 + 50 \left(70^{0.20} - 1\right).$$

Now

$$70^{0.20} = \exp\big(0.20 \ln 70\big), \qquad \ln 70 \approx 4.24850,$$

$$0.20 \ln 70 \approx 0.84970, \qquad \exp(0.84970) \approx 2.33894.$$

So

$$z_{1000} \approx 70 + 50(2.33894 - 1) = 70 + 50(1.33894) = 70 + 66.947 = 136.947,$$

$$\boxed{z_{1000} \approx 136.9 \text{ (1000-day return level)}}$$

5.2 Worked Example 1 (POT/GPD) — Daily rainfall

Assume:

- Data are **daily** for **30 years**: $n = 30 \times 365 = 10{,}950$ observations, $r = 365$ days/year.
- Choose threshold $u = 50$ mm.
- Number of exceedances above u: $N_u = 180$.
- Fitted GPD parameters (MLE): $\hat{\beta} = 10$ mm, $\hat{\xi} = 0.12$.

Step 1: Exceedance rate

$$\hat{\lambda}_u = \frac{N_u}{n} = \frac{180}{10{,}950} \approx 0.016438.$$

Step 2: Compute a T-year return level (take $T = 50$ years) First compute

$$T r \hat{\lambda}_u = 50 \times 365 \times \frac{180}{10{,}950}.$$

Since $10{,}950 = 30 \times 365$, the 365 cancels:

$$T r \hat{\lambda}_u = 50 \times \frac{180}{30} = 50 \times 6 = 300.$$

Now apply the POT return level formula (because $\hat{\xi} \neq 0$):

$$z_{50} = u + \frac{\hat{\beta}}{\hat{\xi}} \left[(T r \hat{\lambda}_u)^{\hat{\xi}} - 1 \right] = 50 + \frac{10}{0.12} \left(300^{0.12} - 1 \right).$$

Compute $300^{0.12}$ using exponentials:

$$300^{0.12} = \exp\bigl(0.12 \ln 300\bigr).$$

With $\ln 300 \approx 5.703782$, we have

$$0.12 \ln 300 \approx 0.684454, \qquad \exp(0.684454) \approx 1.982689.$$

So

$$z_{50} = 50 + \frac{10}{0.12}(1.982689 - 1) = 50 + 83.3333 \times 0.982689 \approx 50 + 81.8907 \approx 131.8907.$$

$$\boxed{z_{50} \approx 131.9 \text{ mm (50-year return level, daily series)}}$$

Optional: Tail probability at a specific large value Estimate $\mathbb{P}(X > 120)$ for $x = 120 > u$:

$$\mathbb{P}(X > 120) \approx \hat{\lambda}_u \left(1 + \hat{\xi} \frac{120 - u}{\hat{\beta}} \right)^{-1/\hat{\xi}} = 0.016438 \left(1 + 0.12 \frac{70}{10} \right)^{-1/0.12}.$$

Compute inside:

$$1 + 0.12 \frac{70}{10} = 1 + 0.84 = 1.84, \quad \Rightarrow \quad \mathbb{P}(X > 120) \approx 0.016438 \times 1.84^{-8.3333} \approx 0.000102.$$

$$\boxed{\mathbb{P}(X > 120) \approx 1.02 \times 10^{-4} \text{ per day}}$$

5.3 Worked Example 2 (Block Maxima/GEV) — Annual maxima

Assume we take **annual maxima** (one max per year) for $k = 30$ years and fit a stationary GEV:

$$\hat{\mu} = 80, \qquad \hat{\sigma} = 15, \qquad \hat{\xi} = 0.10.$$

Compute 50-year return level (in years, because blocks are years) The return level formula (for $\xi \neq 0$) is

$$z_T = \mu + \frac{\sigma}{\xi}\left[\left\{-\log\left(1-\frac{1}{T}\right)\right\}^{-\xi} - 1\right].$$

For $T = 50$:

$$-\log\left(1-\frac{1}{50}\right) = -\log\left(\frac{49}{50}\right) = -\log(0.98) \approx 0.0202027.$$

Now compute the power term:

$$(0.0202027)^{-0.10} = \exp\Big(-0.10\ln(0.0202027)\Big).$$

With $\ln(0.0202027) \approx -3.8990$, we get

$$-0.10\ln(0.0202027) \approx 0.38990, \qquad \exp(0.38990) \approx 1.47727.$$

Plugging into z_{50}:

$$z_{50} = 80 + \frac{15}{0.10}(1.47727 - 1) = 80 + 150 \times 0.47727 \approx 80 + 71.590 \approx 151.590.$$

$$\boxed{z_{50} \approx 151.6 \text{ (50-year return level of annual maxima)}}$$

Quick extra: 20-year and 100-year (same parameters)

$$z_{20} \approx 131.9, \qquad z_{100} \approx 167.6.$$

(Computed by the same steps with $T = 20$ and $T = 100$.)

6 Declustering for POT (if needed)

If exceedances cluster (common in environmental time series), a standard approach is:

D1. Choose a threshold u and a **run length** r_0 (time gap).

D2. Define clusters: consecutive exceedances separated by gaps shorter than r_0 belong to the same cluster.

D3. Reduce each cluster to a single representative, typically the **cluster maximum**.

D4. Fit the GPD to cluster maxima exceedances (now closer to independence).

D5. Optionally estimate the **extremal index** $\theta \in (0, 1]$; effective exceedance rate becomes $\theta\lambda_u$.

7 Choosing Between Block Maxima and POT

- **POT is usually more data-efficient.** It uses all exceedances above u, not just 1 per block.
- **Block maxima is simpler conceptually,** but can waste information and yield wide intervals if few blocks exist.
- If you have only a short record (few years), POT often performs better (if thresholding is done carefully).

8 Nonstationary Extensions (Real Data Often Needs This)

If extremes change over time or with covariates (season, climate index, etc.), allow parameters to depend on t or covariates c_t.

8.1 Nonstationary GEV (Block Maxima)

Example:

$$\mu(t) = \mu_0 + \mu_1 t, \qquad \log \sigma(t) = \sigma_0 + \sigma_1 t, \qquad \xi(t) = \xi_0 \text{ (often kept constant).}$$

8.2 Nonstationary POT (Threshold Exceedances)

Example:

$$\beta(t) = \exp(b_0 + b_1 t), \qquad \xi(t) = \xi_0 \text{ or } \xi(t) = \xi_0 + \xi_1 t.$$

9 Typical Interpretation of the Shape Parameter ξ

- $\xi > 0$ (Fréchet-type): heavy tail, no finite upper bound (very large extremes possible).
- $\xi = 0$ (Gumbel-type): exponential-like tail.
- $\xi < 0$ (Weibull-type): finite upper endpoint.