University of Glasgow

Statistical Modelling of Environmental Extremes

– Session 2 –

Dr Daniela Castro-Camilo

SCAN FOR COURSE MATERIAL

Institut Henri Poincaré, Sorbonne University – March 2022

What is this session about? – Trends and seasonality

- We will learn how to apply the asymptotic results for block-maxima to environmental data, facing very common challenges in environmental problems: trends and seasonality.

- Classical models for block maxima or threshold exceedances assume that observations are independent and identically distributed (iid).

- Specifically, for the block-maxima case, we know that if $X_1, \ldots, X_n$ are iid then, under certain conditions, we can approximate the distribution of $M_n = \max\{X_1, \ldots, X_n\}$ by the generalised extreme value (GEV) distribution

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

  defined of $\{z : 1 + \xi(z-\mu)/\sigma\} > 0\}$, where $\mu \in \mathbb{R}$ is the location, $\sigma > 0$ is the scale and $\xi \in \mathbb{R}$ is the shape parameter.

- But iid (and more generally, stationarity) rarely holds in environmental applications.

- Non-stationarity in environmental process arises due to, e.g., different seasons having different climate patterns, long-term trends owing to climate change or due to the influence of some covariate, among others.

Trends

- An example of trends in block-maxima can be seen in Figure 1, which shows annual maximum concentration of ozone (O_3 ppb) measured in Santiago, Chile.

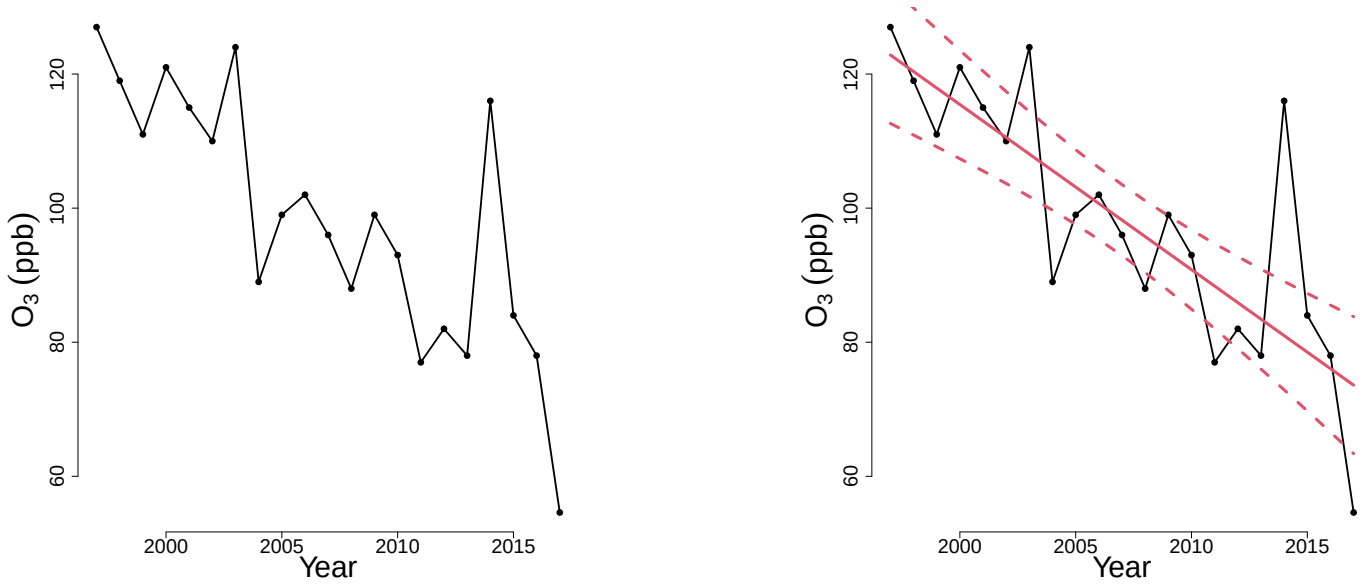

Figure 1: Left: time series plot showing annual maximum concentration of ozone, 1997-2017. Right: same as left plot but with a fitted linear regression line and 95% CI.

- The RHS plot shows a rather strong evidence for a negative trend over the years.
- It is likely that a substantial part of the variability that we observed in the data is actually explained by the systematic variation in ozone over time.

Seasonality

- An example of seasonality in block-maxima can be seen in Figure 2, which shows monthly maximum precipitation (inches) based on daily observations in Fort Collins, Colorado, US.

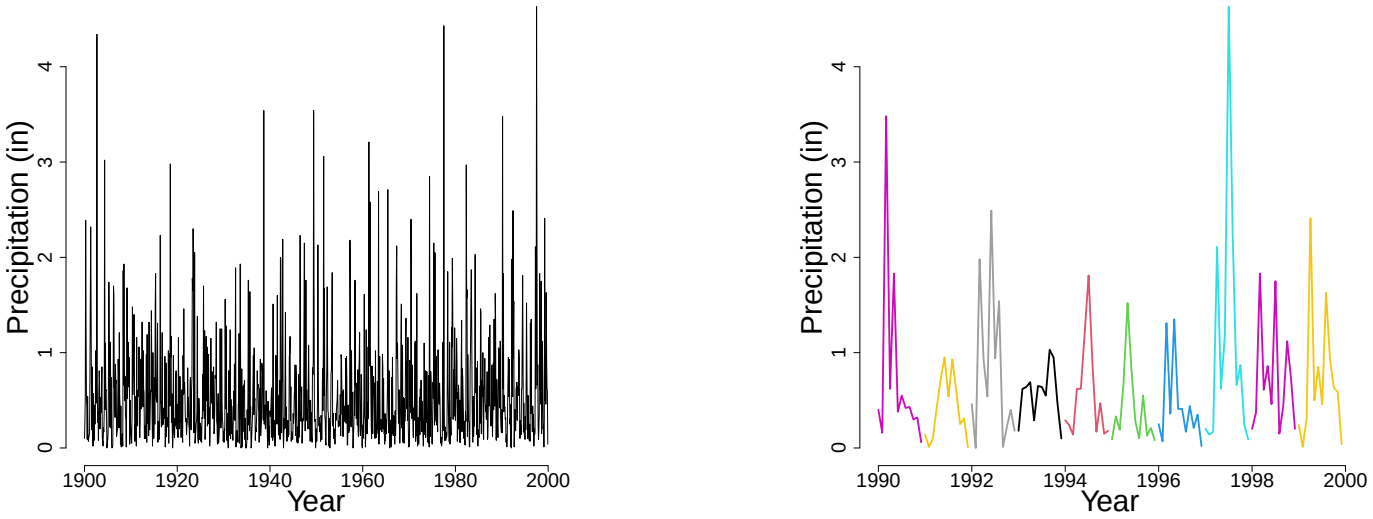

Figure 2: Left: time series plot showing monthly maximum precipitation (inches) measured in Fort Collins, 1990-1999. Right: same as left plot but for the sub-period 1990-1999 with different colours per year.

- Currently, there is no general theory for non-stationary extremes, but alternative solutions can be adopted based on the type of non-stationarity observed.
- In this session, we will use the Ozone and the precipitation data to show ways in which we can deal with non-stationarity.

Trends in GEV models - Ozone data

- We start by showing what happens when we try to fit a GEV distribution without accounting for the non-stationarity observed in Figure 1. We call this model $\mathscr{M}_0$.

- In R, $\mathrm{fit0} = \mathrm{ismev::gev.fit(y)}$ fits a stationary GEV model to data y.

- Task 1: fit a stationary GEV to the ozone data using the ismev package.

Note: whenever you see the R logo, it means we should open the lab sheet and work with R/RStudio.

Trends in GEV models - Ozone data

- The stationary analysis of the Ozone data shows an unsatisfactory fit, which was expected due to the strong negative trend observed in the annual maxima (Figure 1).
- One way to capture the trend we observe in Figure 1 is by allowing the GEV parameters to vary across time.
- Specifically, if Y_t is the maximum Ozone concentration at time t, then we can assume

$$\begin{aligned} Y_t &\sim \text{GEV}(\mu(t), \sigma(t), \xi) \\ \mu(t) = \sum_{j=0}^{J} \beta_j x_j^{\mu}(t), \quad \log\{\sigma(t)\} &= \sum_{k=0}^{K} \gamma_j x_k^{\sigma}(t), \qquad t \in \mathscr{T}, \end{aligned} \tag{1}$$

  where $\mathscr{T} = \{1997, \ldots, 2017\}$ and $x_j^{\mu}(t)$ and $x_k^{\sigma}(t)$ are time-varying covariates and we assume $x_0^{\mu}(t) = x_0^{\sigma}(t) = 1$ (in this way, β_0 and γ_0 are intercepts). The covariates can take any form:
  - Polynomial terms: $x_k^{\sigma}(t) = t, t^2, t^3$, etc.
  - Harmonic terms: $x_k^{\sigma}(t) = \sin(2\pi t/T), \cos(2\pi t/T)$.
  - More flexible terms: GAMs.
- In principle, we could also allow ξ to vary with time, but, due to practical (easier fit) and philosophical(!) reasons, we will assume ξ to be constant.

Trends in GEV models - Ozone data

- Since the trend in Figure 1 looks almost linear and no seasonality patterns are observed (expected since we have annual data), we can simplify the model in (1) and assume only a linear form in μ and σ:

$$\mu(t) = \beta_0 + \beta_1 t, \quad \log\{\sigma(t)\} = \gamma_0 + \gamma_1 t, \quad t \in \mathscr{T}. \tag{2}$$

- Note that they are various nested models of (2). We will fit and compare three candidates:

$$\begin{aligned} \mathscr{M}_1&: \quad \beta_1 \neq 0, \gamma_1 = 0, \text{ i.e., trend in } \mu \text{ but not in } \sigma. \\ \mathscr{M}_2&: \quad \beta_1 = 0, \gamma_1 \neq 0, \text{ i.e., trend in } \sigma \text{ but not in } \mu. \\ \mathscr{M}_3&: \quad \beta_1 \neq 0, \gamma_1 \neq 0, \text{ i.e., trend in } \mu \text{ and } \sigma. \end{aligned}$$

- Model choice: deviance statistic for nested models ($\mathscr{M}_0 \subset \mathscr{M}_1$) or AIC/BIC:

$$D = 2\{\ell(\mathscr{M}_1) - \ell(\mathscr{M}_0)\}, \quad \text{AIC}(\mathscr{M}_k) = -2\ell(\mathscr{M}_k) + 2p_k, \quad \text{BIC}(\mathscr{M}_k) = -2\ell(\mathscr{M}_k) + \log(N)p_k.$$

  $\mathscr{M}_0$ is rejected if $D > c_\alpha$ where c_α is the $(1-\alpha)-$quantile of the χ^2_{df} and df is the difference in the dimensionality of $\mathscr{M}_1$ and $\mathscr{M}_0$. For AIC/BIC, we want the model(s) that minimise them.

- Model diagnostics will be based on probability plots and qq-plots.

- Task 2: fit $\mathscr{M}_{1,2,3}$ using the `ismev` package.

Stationary case

- As you saw in the Risk Analytics course, quantiles of the GEV are called *return levels*.
- Based on a GEV fitted to annual maxima, the $r-$year return level z_r associated with the return period r can be computed by setting $G(z_r) = 1 - 1/r$ and solving for z_r.
- Suppose that authorities require an estimate of z_{50}, the ozone concentration we might expect to be exceeded once in $r = 50$ years (this could help to create regulations to prevent high levels of ozone, even higher than the ones we have observed).
- We can translate the above into a probability statement:

$$\Pr(\text{annual max} > z_{50}) = \frac{1}{50} \quad \Rightarrow \quad \Pr(\text{annual max} \leq z_{50}) = 1 - \frac{1}{50} = 0.98$$

- Provided we "trust" our model,

$$\Pr(\text{annual max} \leq z_{50}) = G(z_{50}; \hat{\mu}, \hat{\sigma}, \hat{\xi}) \quad \Rightarrow \quad \exp\left\{ -\left[1 + \hat{\xi}\left(\frac{z_{50} - \hat{\mu}}{\hat{\sigma}} \right) \right]^{-1/\hat{\xi}} \right\} = 0.98$$

- And we solve for z_{50} (technically for $\hat{z}_{50}$ as the return level depends on the estimated parameters):

$$\hat{z}_{50} = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}} \left[\{ -\log(1 - r^{-1}) \}^{-\hat{\xi}} - 1 \right] = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}} \left[\{ -\log(0.98) \}^{-\hat{\xi}} - 1 \right]$$

- Note that we are estimating beyond our range of observations: we are computing the 50-years return level with only 21 years of data.

Non-stationary case

- For the ozone data, $\mu = \mu(t)$, therefore the above equation becomes

$$\exp\left\{-\left[1+\hat{\xi}\left(\frac{z_{50}-\hat{\mu}(t)}{\hat{\sigma}}\right)\right]^{-1/\hat{\xi}}\right\} = 0.98 \quad \Rightarrow \quad \hat{z}_{50}(t) = \hat{\mu}(t) + \frac{\hat{\sigma}}{\hat{\xi}}\left[\{-\log(0.98)\}^{-\hat{\xi}} - 1\right]$$

- Figure 3 show how we might expect this estimate to vary for $t = 22, 23, \ldots$ i.e., for the years 2018, 2019,...

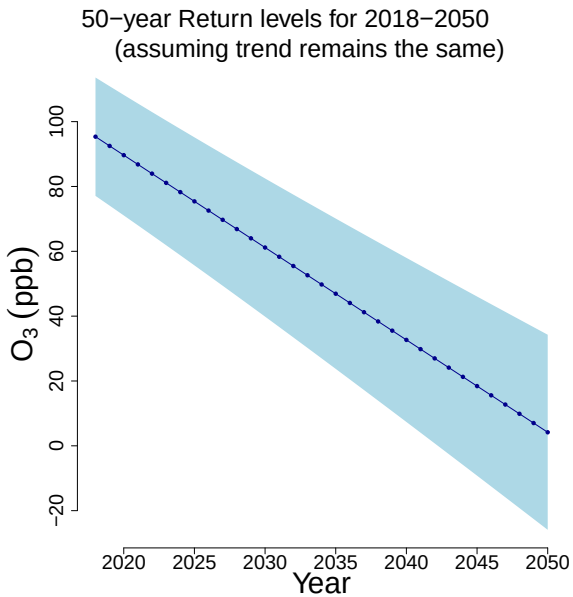

Figure 3: Forecasted 50-year return levels for the Ozone data, for the years 2018-2050 with 95% CI (based on normal approximation and delta method).

- We could treat these as forecasts of the 50-year return levels as we move through time.
- Such forecasts will assume the linear trend for μ continues beyond the range of data we have observed.

Return levels for non-stationary GEV models

- !! As it is well-known in statistics, it is very dangerous to extrapolate a linear regression outside the sample support.
- Hence, extra caution is required when interpreting $\hat{z}_{50}(t)$ for t beyond the sample support because it means that we are assuming the trend will remain the same until time t.
- Within the sample support ($t \in \{1, \ldots, 21\}$), we can compute the "effective" return levels.
- Effective return levels (ERLs) are the return levels obtained for given parameter values of a non-stationary model. Figure 4 show the 2-, 50- and 100-year ERLs.

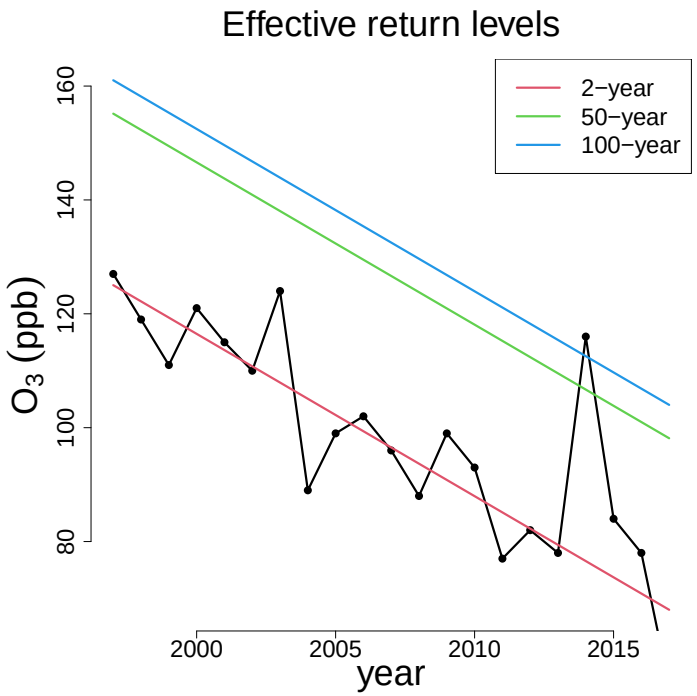

Figure 4: Effective 2-, 50- and 100-year return levels for the Ozone data.

Task 3: Reproduce Figure 4 in R.

Question: Is there another name for the 2-year return level?

Seasonality in GEV models - Precipitation data

- Figure 2 ⏩ here shows no clear trend and a strong seasonal pattern (expected since we have monthly data). We can try to capture this form of non-stationarity by including harmonic terms in the GEV parameters:

$$\mu(t) = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{T}\right) + \beta_2 \cos\left(\frac{2\pi t}{T}\right), \quad \log\{\sigma(t)\} = \gamma_0 + \gamma_1 \sin\left(\frac{2\pi t}{T}\right) + \gamma_2 \cos\left(\frac{2\pi t}{T}\right). \quad (3)$$

- In this case, $t = 1, \ldots, 12 \times 100 = 1200$ and $T = 12$ is the oscillation period.
- As before, there are various nested models of (3) but in this session we will focus on fitting model (3) and computing the associated 2- and 100-year return levels (see Figure 5).

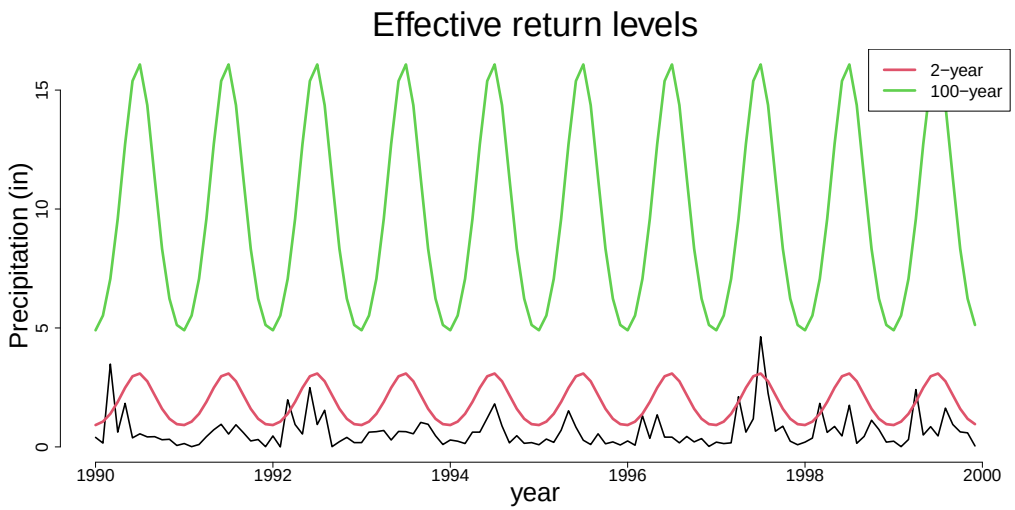

Figure 5: Effective 2- and 100-year return levels for the prec data.

Task 4: Fit model (3) and Reproduce Figure 5 in R. Note that the data is on daily scale, therefore monthly max need to be computed.

HW: fit and compare all the nested models of (3).

Trends and seasonality in GEV models - Challenge

The dataset `co2.txt` contains 468 (simulated) monthly maximum concentrations (in parts per million) of CO_2 from 1959 to 1997. Fit a GEV distribution to the data (accounting for possible non-stationarity) and plot the effective 2- and 100-year return levels (see Figure 6).

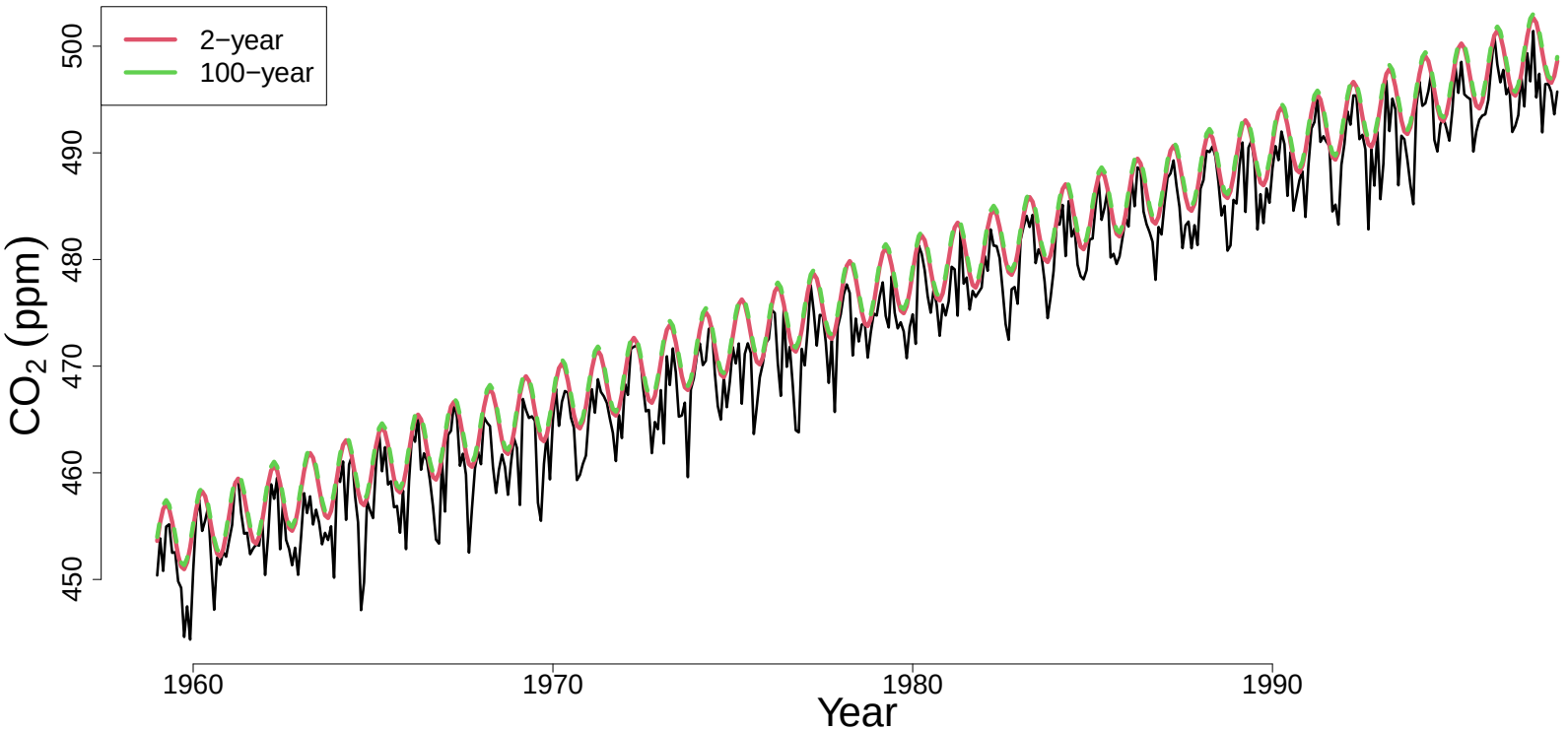

Figure 6: Simulated monthly maximum concentration of CO_2, 1959–1997 (black) with fitted 2- and 100-yer return levels (red and green, respectively).