Changes in extremes

Detection and consequences

Ilaria Prosdocimi - Ca' Foscari University of Venice

ilaria.prosdocimi@unive.it

13 July 2020

Change (?)

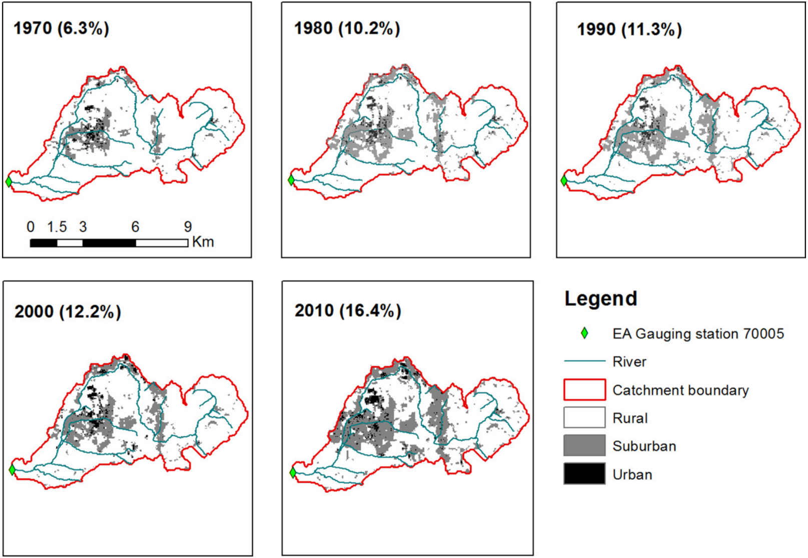
Increasing interest in assessing changes in extremes related to natural hazards.

Many studies investigate changes in extreme rainfall and extreme flows.

Changes in magnitude/frequencies: infrastructures are designed to withstand extreme events of some magnitude.

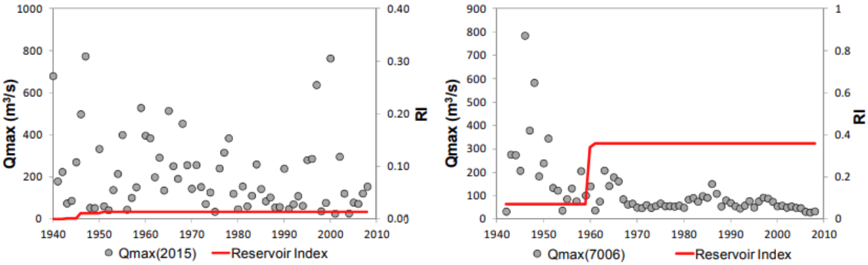
Problematic if these become more (or less!) frequent.

What causes change



from Prosdocimi et al. (2015), WRR, doi:10.1002/2015WR017065

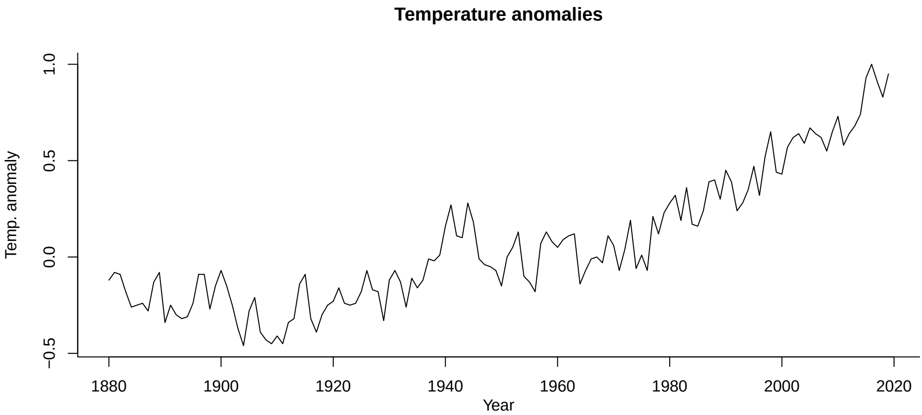
What causes change



from Lopez Frances (2013), HESS, doi:10.5194/hess-17-3189-2013

What causes change

Implicit assumption:



NOAA National Centers for Environmental information, Climate at a Glance:
Global Time Series, published June 2020, retrieved on July 5, 2020 from https://www.ncdc.noaa.gov/cag/

Why study change?

- Understand if process of interest (river flow, rainfall, etc) is evolving in time
- Understand how process of interest is affected by external drivers
- Assess risk connected to a certain hazard and its evolution
- If this is changing, how to account for this

Detection,

Why study change?

- Understand if process of interest (river flow, rainfall, etc) is evolving in time
- Understand how process of interest is affected by external drivers
- Assess risk connected to a certain hazard and its evolution
- If this is changing, how to account for this

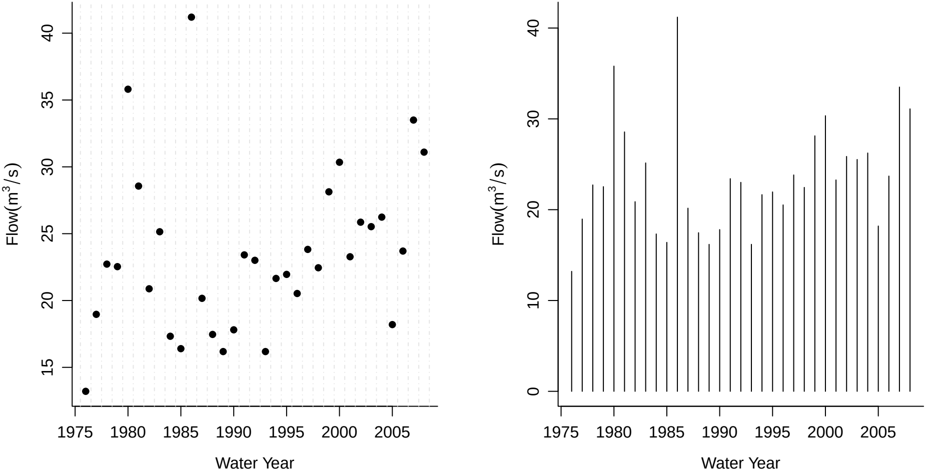
Detection, attribution

Why study change?

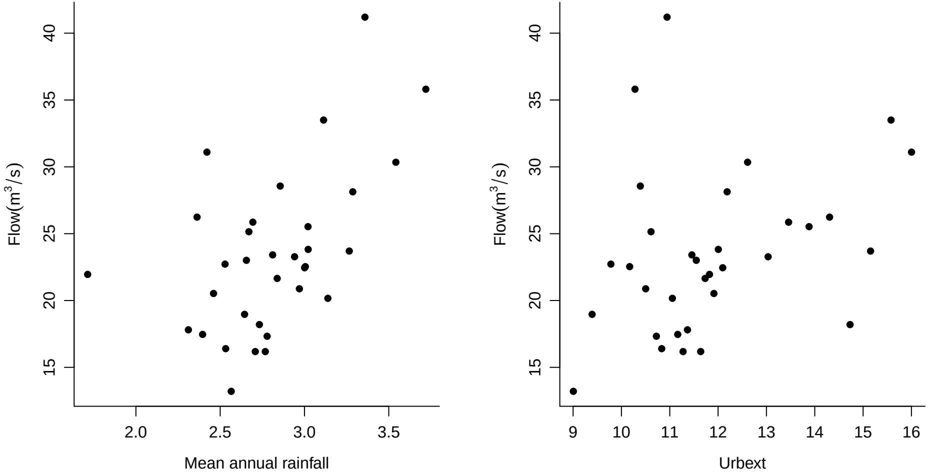
- Understand if process of interest (river flow, rainfall, etc) is evolving in time
- Understand how process of interest is affected by external drivers
- Assess risk connected to a certain hazard and its evolution
- If this is changing, how to account for this

Detection, attribution and management.

The Lostock at Littlewood Bridge



The Lostock at Littlewood Bridge



Statistical tools

We assume that $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a random sample from a population.

We are interested in discovering some property of the population.

Statistical tools

We assume that $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a random sample from a population.

We are interested in discovering some property of the population.

Inference framework:

- Parametric: assume that y_i is a realisation of some distribution described by **parameters** $\boldsymbol{\theta}$ ($f(y_i; \boldsymbol{\theta})$)
- Non-parametric: no assumption on the distribution of $f(y)$ is made (well, less assumptions...)

Parameteric framework

Advantage of parametric framework:

- Describe the whole distribution (including, for example, quantiles)
- A very general framework
- Easy to extend to very complex models (but estimation can be complicated)

Parameteric framework

Advantage of parametric framework:

- Describe the whole distribution (including, for example, quantiles)
- A very general framework
- Easy to extend to very complex models (but estimation can be complicated)

The parametric framework:

- Assume that each member of the sample y_i comes from some distribution Y_i
- Often assumed: $(Y_1, \ldots, Y_n)$ are independent and identically distributed (iid)
- Assume that Y_i follows a known distribution parametrised by $\boldsymbol{\theta}$
- (for example $Y_i \sim N(\mu, \sigma)$, with $\boldsymbol{\theta} = (\mu, \sigma)$)
- Find estimates $\hat{\boldsymbol{\theta}}$ based on the sample

Estimation methods

- Method of moments
- Maximum likelihood
- Bayesian approaches

Estimation methods

- Method of moments
- Maximum likelihood
- Bayesian approaches

Choice of framework and estimation method should depend on:

- Actual data properties
- Main inferential question (and importance of uncertainty assessment)
- Computational hurdle
- Model complexity
- Presence of prior information (which can be formalised)

Maximum likelihood estimation

The likelihood function is defined as

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}),$$

but calculations typically employ the log-likelihood

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta}).$$

$\hat{\boldsymbol{\theta}}_{ML}$ is the value that maximises $l(\boldsymbol{\theta}; \boldsymbol{y})$.

Maximum likelihood estimation

The likelihood function is defined as

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}),$$

but calculations typically employ the log-likelihood

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta}).$$

$\hat{\boldsymbol{\theta}}_{ML}$ is the value that maximises $l(\boldsymbol{\theta}; \boldsymbol{y})$.

Asymptotically ($n \to \infty$) we have that $\hat{\boldsymbol{\theta}}_{ML} \sim N(\boldsymbol{\theta}, I_E(\boldsymbol{\theta})^{-1})$ where $I_E(\boldsymbol{\theta})$ is the expected information matrix, with elements

$$e_{i,j}(\theta) = E\left[-\frac{d^2 l(\theta)}{d\theta_i d\theta_j}\right]$$

Typically $I_E(\boldsymbol{\theta})$ is unknown: use the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$.

Parametric models for change

- Assume Y_i comes from a distribution $f(\boldsymbol{\theta}_i, y_i)$
- Assume $\boldsymbol{\theta}_i = g(\boldsymbol{x}_i)$
- So $Y_i = (Y|X = x_i)$ with $f(g(\boldsymbol{x}_i), y_i)$

Example. Linear regression (with two explanatory variables):

- $Y_i \sim N(\mu_i, \sigma)$; $\boldsymbol{\theta}_i = (\mu_i, \sigma)$.
- $\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ - linear relationship.
- σ is constant.
- As a consequence: $E[Y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$, $V[Y_i] = \sigma^2$.
- We can describe Y and how it varies with X

Parametric models for change

Linear regression likelihood:

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta}) \propto -n \log \sigma - \frac{(y - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2}{2\sigma^2}$$

ML estimates can be derived analytically: $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma})$.

And we have, for example, $\hat{\beta}_i \sim N(\beta_i, \hat{\sigma}_{\beta_i})$.

[1] Prosdocimi et al, NHESS, doi:10.5194/nhess-14-1125-2014

Parametric models for change

Linear regression likelihood:

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta}) \propto -n \log \sigma - \frac{(y - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2}{2\sigma^2}$$

ML estimates can be derived analytically: $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma})$.

And we have, for example, $\hat{\beta}_i \sim N(\beta_i, \hat{\sigma}_{\beta_i})$.

From this once can construct confidence intervals for β_i or perform a test such as:

$$H_0 : \beta_0 \geq \tilde{\beta} \qquad VS \qquad H_1 : \beta_0 < \tilde{\beta}$$

By default $\tilde{\beta} = 0$, but one can test for any value $\tilde{\beta}$ and statistical test ($=$, $\leq$, $\geq$).[1]

Notice that if x_j is a factor one can account for step changes (change points).

[1] Prosdocimi et al, NHESS, doi:10.5194/nhess-14-1125-2014

Parametric models of change in extremes

Describing extremes is a different task than describing the typical behaviour.

$(y_1, \ldots, y_n)$ is a sample of extremes: what is a reasonable assumption for Y?

Extreme Value Theory gives theoretical derivation, but practice is often different.

Regardless of the choice of $f(y, \boldsymbol{\theta})$ - parametric models of change for extremes can be easily constructed assuming $Y_i = (Y|X = x_i)$ and $\boldsymbol{\theta}_i = g(\boldsymbol{x}_i)$.

Parametric models of change in extremes

Describing extremes is a different task than describing the typical behaviour.

$(y_1, \ldots, y_n)$ is a sample of extremes: what is a reasonable assumption for Y?

Extreme Value Theory gives theoretical derivation, but practice is often different.

Regardless of the choice of $f(y, \boldsymbol{\theta})$ - parametric models of change for extremes can be easily constructed assuming $Y_i = (Y|X = x_i)$ and $\boldsymbol{\theta}_i = g(\boldsymbol{x}_i)$.

What is an extreme?

- Largest event over a certain amount of time (eg water year, season)
- Events larger than a certain high threshold (independent events?)

Parametric models in extremes

Traditional (asymptotic) results based on extremes of stationary series:

- Block maxima: $Y \sim GEV(\mu, \sigma, \xi)$
- Threshold exceedance magnitude: $Y \sim GP(\sigma, \xi)$
- Threshold exceedance frequency: $N \sim Pois(\lambda)$

Parametric models in extremes

Traditional (asymptotic) results based on extremes of stationary series:

- Block maxima: $Y \sim GEV(\mu, \sigma, \xi)$
- Threshold exceedance magnitude: $Y \sim GP(\sigma, \xi)$
- Threshold exceedance frequency: $N \sim Pois(\lambda)$

Using exceedances typically results in larger samples (so less variability in estimates).

Parametric models in extremes

Traditional (asymptotic) results based on extremes of stationary series:

- Block maxima: $Y \sim GEV(\mu, \sigma, \xi)$
- Threshold exceedance magnitude: $Y \sim GP(\sigma, \xi)$
- Threshold exceedance frequency: $N \sim Pois(\lambda)$

Using exceedances typically results in larger samples (so less variability in estimates).

In practice other distributions are often assumed for Flow maxima.

Generalised Extreme Value distribution

$$\text{The GEV CDF:} \quad F(y, \boldsymbol{\theta}) = \exp\left\{ -\left(1 + \xi \frac{y - \mu}{\sigma}\right)^{-1/\xi} \right\}$$

$\boldsymbol{\theta} = (\mu, \sigma, \xi)$:

- $\mu \in \mathbb{R}$: location parameter
- $\sigma > 0$; scale parameter
- $\xi \in \mathbb{R}$: shape parameter.

$Y \sim GEV(\mu, \sigma, \xi)$ is defined on $y : 1 + \xi(y - \mu)/\sigma > 0$, this means:

- $y \in [\mu - \sigma/\xi, \infty)$, if $\xi > 0$ (Frechet)
- $y \in (-\infty, \mu - \sigma/\xi]$, if $\xi < 0$ (Weibull)
- $y \in (-\infty, \infty)$, if $\xi = 0$ (Gumbel)

BUT! In engineering/hydrology $Y \sim GEV(\xi, \alpha, \kappa)$ and $\kappa = -\xi$. Software can use different parametrisation.

Generalised Extreme Value distribution

$$\text{Quantile function (for } \xi \neq 0) : \; q(y, \theta) = \mu + \frac{\sigma}{\xi} \left[(-\log(1-p))^{-\xi} - 1 \right]$$

Generalised Extreme Value distribution

$$\text{Quantile function (for } \xi \neq 0) : \; q(y, \theta) = \mu + \frac{\sigma}{\xi} \left[(-\log(1-p))^{-\xi} - 1 \right]$$

Modelling change:

$$\mu = \mu_0 + \mu_1 x$$

Generalised Extreme Value distribution

$$\text{Quantile function (for } \xi \neq 0): \; q(y, \theta) = \mu + \frac{\sigma}{\xi}\left[(-\log(1-p))^{-\xi} - 1\right]$$

Modelling change:

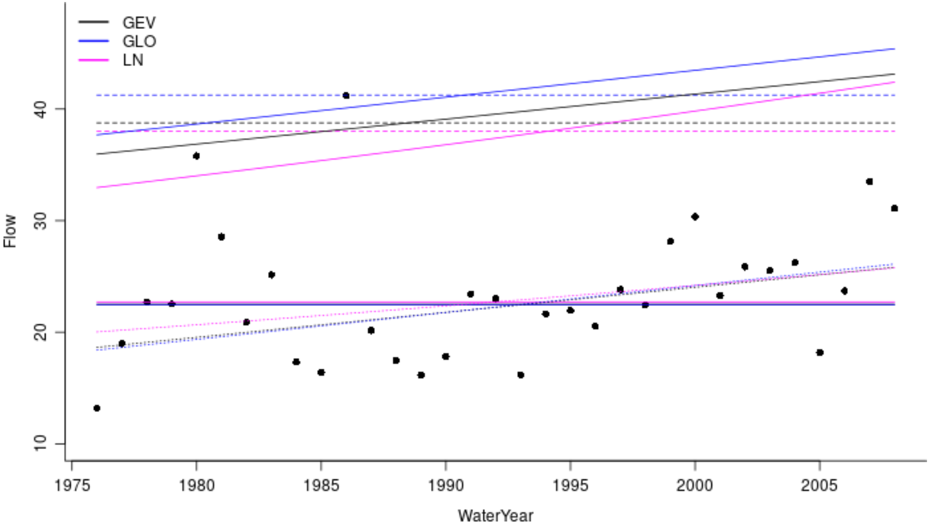
$$\mu = \mu_0 + \mu_1 x$$

Effective quantile for $x = x^*$:

$$q(y, \theta(x^*)) = \mu_0 + \mu_1 x^* + \frac{\sigma}{\xi}\left[(-\log(1-p))^{-\xi} - 1\right]$$

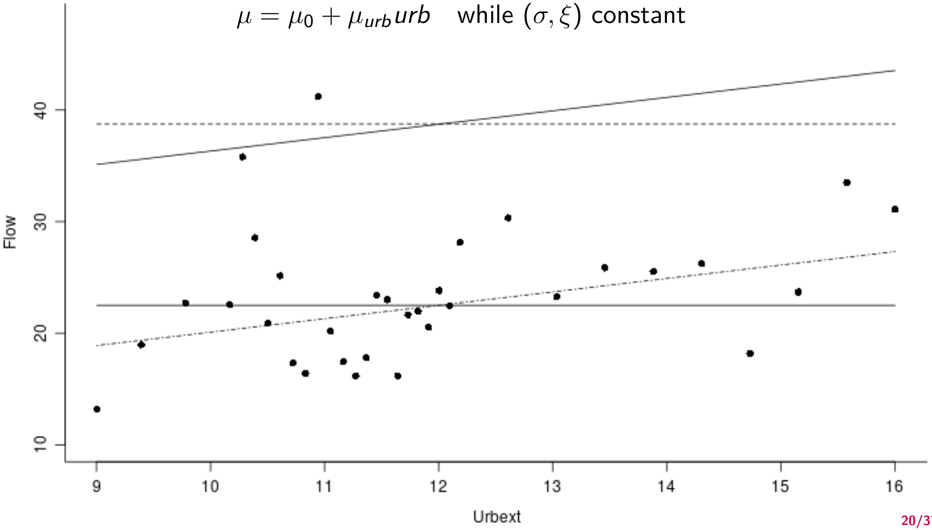
Changes in annual maxima - choice of distribution

The Lostock at Littlewood Bridge: median and effective 50-yrs event.



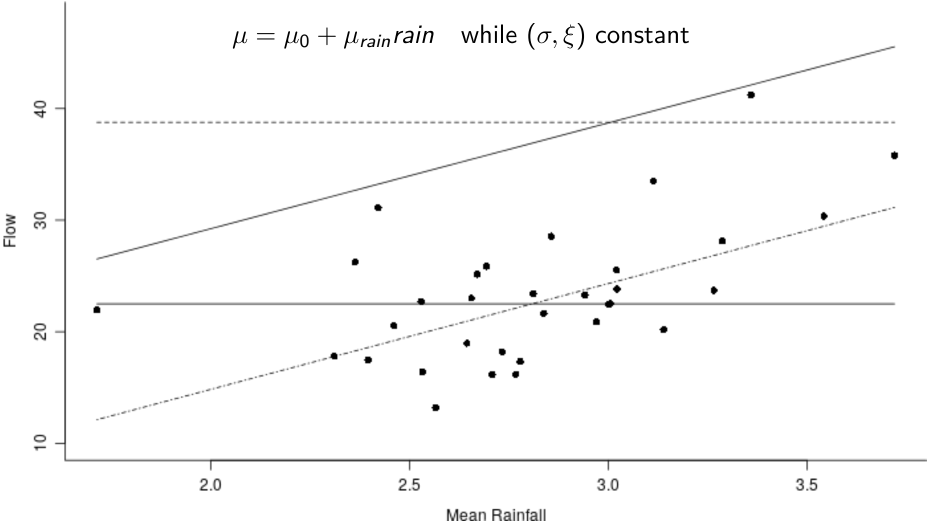
Changes in annual maxima

Time is not a cause for change, but land cover changes impact peak flow.



Changes in annual maxima

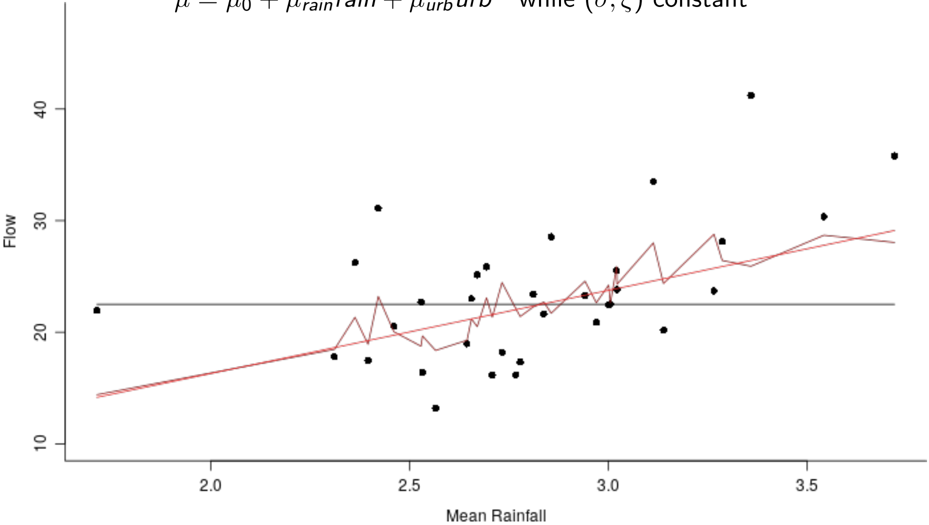
Time is not a cause for change, but soil wetness impact peak flow.



Changes in amax - effect of rain given Urbext

Separate effect of rain and urbanisation:

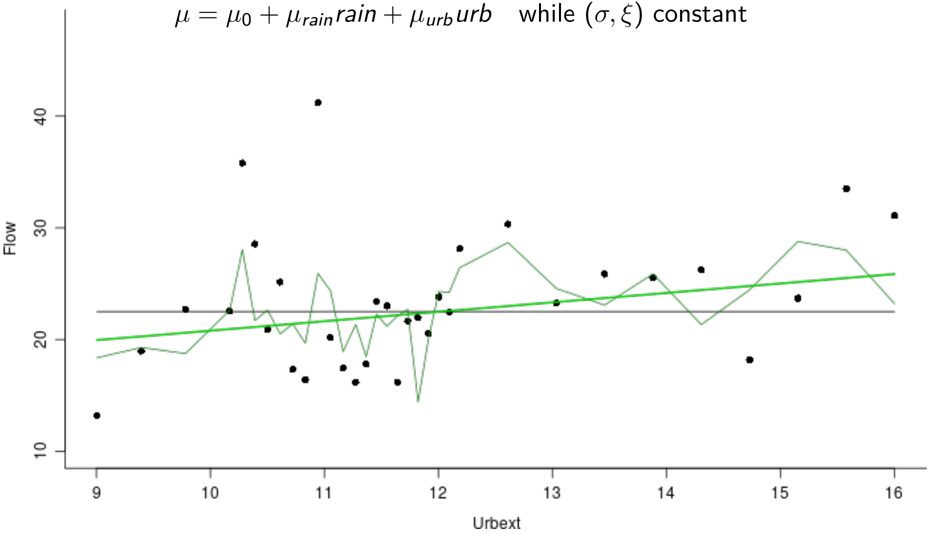
$$\mu = \mu_0 + \mu_{rain} rain + \mu_{urb} urb \quad \text{while } (\sigma, \xi) \text{ constant}$$



Changes in amax - effect of Urbext given rain

Separate effect of rain and urbanisation:

$\mu = \mu_0 + \mu_{rain} rain + \mu_{urb} urb$ while (σ, ξ) constant



Changes in annual maxima - estimated parameters

Rain as covariate (log-lik: -98.39)

	μ_0	μ_{urb}	μ_{rain}	σ	ξ
MLE	-5.604	-	9.479	4.042	0.003
se	8.158	-	2.830	0.622	0.168

Urbext as covariate (log-lik: -100.0004)

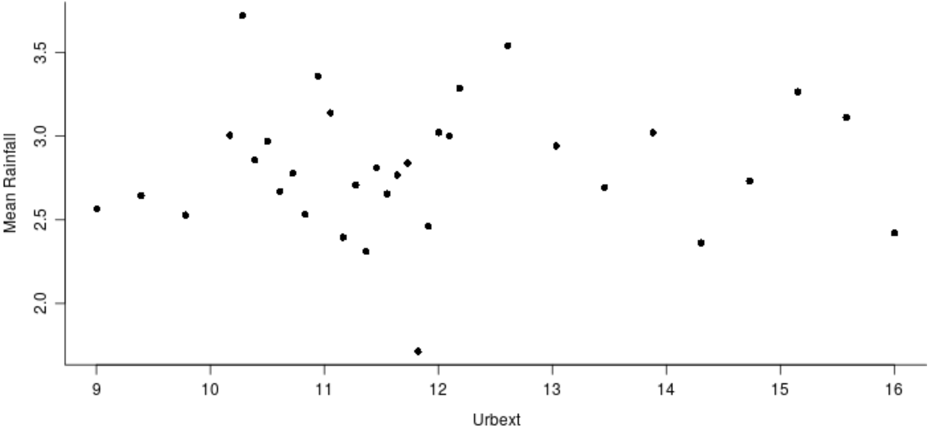
	μ_0	μ_{urb}	μ_{rain}	σ	ξ
MLE	6.53	1.20	-	4.17	0.04
se	4.79	0.40	-	0.60	0.13

Rain and urbext as covariate (log-lik: -96.47)

	μ_0	μ_{urb}	μ_{rain}	σ	ξ
MLE	-9.767	0.845	7.449	3.862	-0.016
se	7.344	0.422	2.668	0.580	0.153

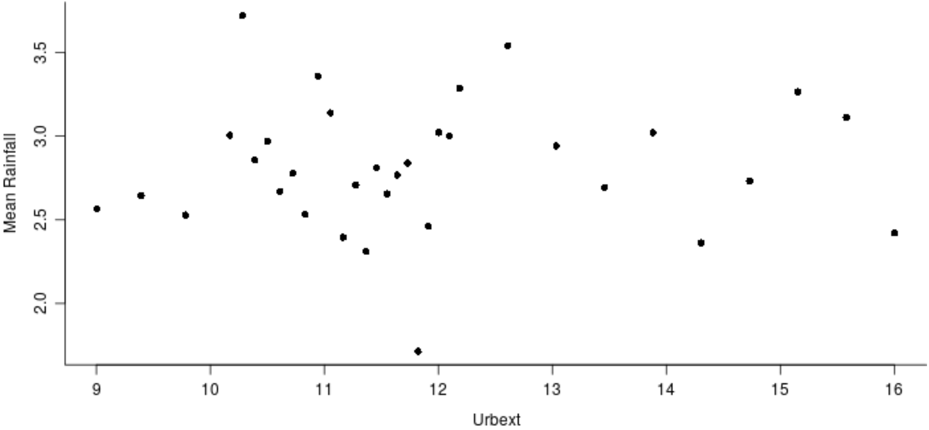
Changes in extremes - attribution

Kendall's $\hat{\tau}(\text{Urbext, Rain}) = 0.068$.



Changes in extremes - attribution

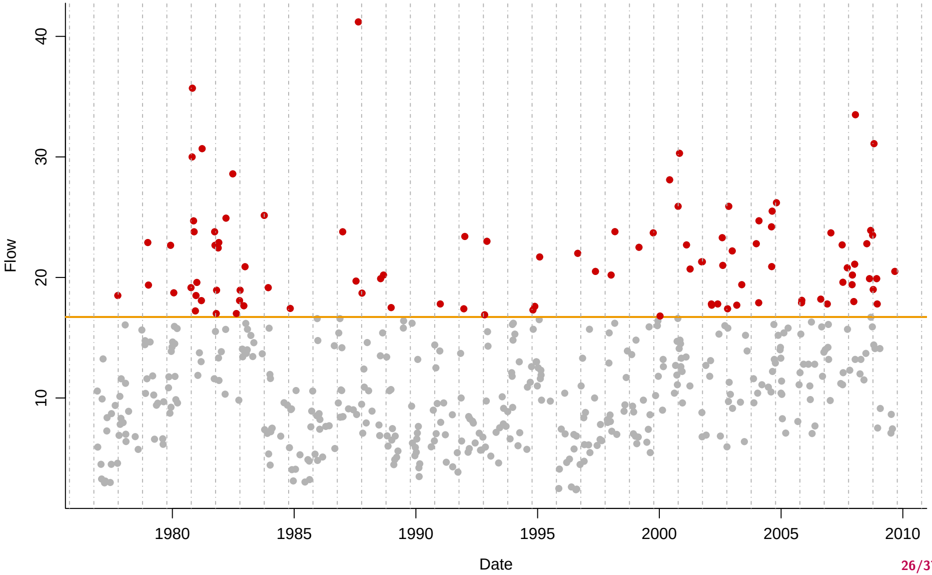
Kendall's $\hat{\tau}$(Urbext, Rain) $= 0.068$.



Reality is complex: linear models are a (over-simplified!) representation.

Changes is peaks over threshold

Extract observations above a high threshold



Generalised Pareto Distribution

Y is taken to be the observations above a high threshold u ($Y = (X|X > u)$).

GP is the limiting distribution for the magnitude of exceedances.

$$F(y, u, \boldsymbol{\theta}) = 1 - \left(1 + \xi \frac{y - u}{\tilde{\sigma}}\right)^{-1/\xi}$$

u is a constant, $\boldsymbol{\theta} = (\sigma, \xi)$:

- $\sigma > 0$; scale parameter
- $\xi \in \mathbb{R}$: shape parameter.

The domain changes depending on the sign of ξ: $y \in [u, \infty)$, if $\xi \geq 0$;
$y \in (-\infty, u - \sigma/\xi]$, if $\xi < 0$.

$$\text{Quantile function:} \quad q(p, u, \boldsymbol{\theta}) = u + \frac{\sigma}{\xi}(p^{-\xi} - 1)$$

Generalised Pareto Distribution

Y is taken to be the observations above a high threshold u ($Y = (X|X > u)$).

GP is the limiting distribution for the magnitude of exceedances.

$$F(y, u, \boldsymbol{\theta}) = 1 - \left(1 + \xi \frac{y - u}{\tilde{\sigma}}\right)^{-1/\xi}$$

u is a constant, $\boldsymbol{\theta} = (\sigma, \xi)$:

- $\sigma > 0$; scale parameter
- $\xi \in \mathbb{R}$: shape parameter.

The domain changes depending on the sign of ξ: $y \in [u, \infty)$, if $\xi \geq 0$;
$y \in (-\infty, u - \sigma/\xi]$, if $\xi < 0$.

$$\text{Quantile function:} \quad q(p, u, \boldsymbol{\theta}) = u + \frac{\sigma}{\xi}(p^{-\xi} - 1)$$

Modelling change: $\sigma_0 + \sigma_1 x$

Point Process representation of extremes

Exceedances frequency and magnitude traditionally modelled as separate processes.

They can be modelled in a unique framework using a Point Process representation of extremes [2].

[2] Smith, Statist. Sci., doi:10.1214/ss/1177012400

Point Process representation of extremes

Exceedances frequency and magnitude traditionally modelled as separate processes.

They can be modelled in a unique framework using a Point Process representation of extremes [2]. This representation is under-utilised in hydrology.

$N = \{\text{no. Exceedance in a Year}\}$. $N \sim Pois(\lambda)$

$P(\text{no. Exceedance in a Year})$ is linked to magnitudes.

[2]Smith, Statist. Sci., doi:10.1214/ss/1177012400

Point Process representation of extremes

Exceedances frequency and magnitude traditionally modelled as separate processes.

They can be modelled in a unique framework using a Point Process representation of extremes [2]. This representation is under-utilised in hydrology.

$N = \{\text{no. Exceedance in a Year}\}$. $N \sim Pois(\lambda)$

$P(\text{no. Exceedance in a Year})$ is linked to magnitudes.

Express this using GEV-parameters:

$$\log \lambda = -\frac{1}{\xi} \log \left[1 + \xi \frac{u - \mu}{\sigma} \right]$$

[2] Smith, Statist. Sci., doi:10.1214/ss/1177012400

Point Process representation of extremes

Exceedances frequency and magnitude traditionally modelled as separate processes.

They can be modelled in a unique framework using a Point Process representation of extremes [2]. This representation is under-utilised in hydrology.

$N = \{\text{no. Exceedance in a Year}\}$. $N \sim Pois(\lambda)$

$P(\text{no. Exceedance in a Year})$ is linked to magnitudes.

Express this using GEV-parameters:

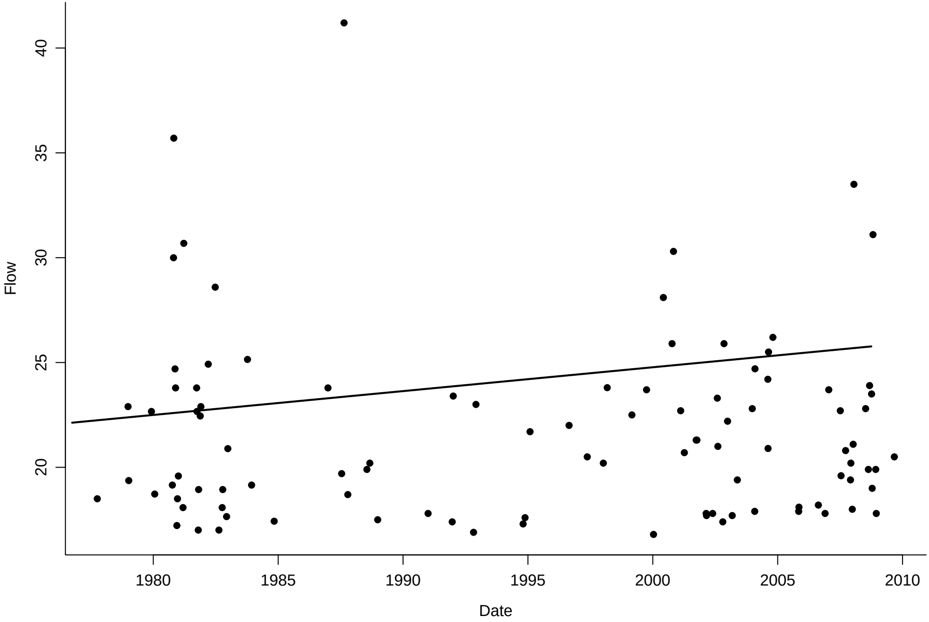
$$\log \lambda = -\frac{1}{\xi} \log \left[1 + \xi \frac{u - \mu}{\sigma} \right]$$

Express changes in magnitude and frequency in the same model

Same meaning as GEV models of change

[2] Smith, Statist. Sci., doi:10.1214/ss/1177012400

Changes in Peaks - Point Process



Changes in extremes - comparing the models

Rain and urbext as covariate - GEV:

	μ_0	μ_{urb}	μ_{rain}	σ	ξ
MLE	-9.767	0.845	7.449	3.862	-0.016
se	7.344	0.422	2.668	0.580	0.153

Rain and urbext as covariate - PP:

	μ_0	μ_{urb}	μ_{rain}	σ	ξ
MLE	-12.139	0.930	8.007	4.622	-0.184
se	6.757	0.320	1.723	0.368	0.064

Changes in extremes - comparing the models

Rain and urbext as covariate - GEV:

	μ_0	μ_{urb}	μ_{rain}	σ	ξ
MLE	-9.767	0.845	7.449	3.862	-0.016
se	7.344	0.422	2.668	0.580	0.153

Rain and urbext as covariate - PP:

	μ_0	μ_{urb}	μ_{rain}	σ	ξ
MLE	-12.139	0.930	8.007	4.622	-0.184
se	6.757	0.320	1.723	0.368	0.064

Larger sample size leads to more precise estimation (statistically)

Tail estimate is quite different

Changes in extremes

Parametric approaches: easy to include predictors and test for significance

Changes in extremes

Parametric approaches: easy to include predictors and test for significance

This might be a bug and not a feature

Changes in extremes

Parametric approaches: easy to include predictors and test for significance

This might be a bug and not a feature

The assumption is that $Y_i = (Y|X = x_i)$ follows $f(y; \boldsymbol{\theta})$ - goodness of fit should be carried out on **residuals**

Statistical EVT and practice are not aligned

Detection

Methods sometimes chosen because of data availability

Statistical models rely on assumption of iid random observations

Short records: hard to identify complex evolutions

Short records: hard to observe a good range of the explantory variable

When detecting “change”: what are we detecting?[3]

[3] Merz et al, HESS, doi:10.5194/hess-16-1379-2012

Attribution

Golden standard of causality is randomised trials: what about observational studies?

Climate sciences reproduce the treatment/placebo framework with numerical experiments (how good for extremes?).

Some numerical experiments done in hydrology - but systems are complex.

Causality: a cascade of impacts (with feedback[4])

[4] Zhang et al, Nature, doi:10.1038/s41586-018-0676-z

Changes in annual maxima - uncertainity

Structures are designed for the "T-Year" event: estimated as the 1-1/T quantile.

If the distribution is changing so is the quantile.

Changes in annual maxima - uncertainity

Structures are designed for the "T-Year" event: estimated as the 1-1/T quantile.

If the distribution is changing so is the quantile.

	Q100	95% lb	95% ub	width
no-change	30.514	41.837	53.159	11.322
Rain = max(Rain)	33.676	48.403	63.130	14.727

Adding parameters adds variation to the estimates - is it worth it?

Changes in annual maxima - uncertainity

Structures are designed for the "T-Year" event: estimated as the 1-1/T quantile.

If the distribution is changing so is the quantile.

	Q100	95% lb	95% ub	width
no-change	30.514	41.837	53.159	11.322
Rain = max(Rain)	33.676	48.403	63.130	14.727

Adding parameters adds variation to the estimates - is it worth it?

Bias-variance trade-off and parsimonious models.

Changes in extremes - consequences

How to quantify risk under change?[5]

Choice of distribution has an impact on estimates of rare events

Today I used "effective design events": $q(p; \hat{\boldsymbol{\theta}})$. So at $X = x^*$: $q(p; \hat{\boldsymbol{\theta}}(x^*))$.

Choice of distribution/model has an impact on estimates of rare events.

Choice of model has an impact of description of change[6].

$$\text{GEV quantile function (for } \xi \neq 0\text{)}: \; q(y, \theta) = \mu + \frac{\sigma}{\xi}\left[(-\log(1-p))^{-\xi} - 1\right]$$

Compare effective return levels for x^* and x_0:

$$q(p; \hat{\boldsymbol{\theta}}(x^*)) - q(p; \hat{\boldsymbol{\theta}}(x_0)) = \mu_1(x^* - x_0)$$

[5] Volpi, Wires Water, doi:10.1002/wat2.1340

[6] Vogel et al JAWRA doi:10.1111/j.1752-1688.2011.00541.x

(Statistical) recommended reading

Coles, S (2001), An introduction to statistical modeling of extreme values, Springer

Katz, R.W., Parlange, M.B. and Naveau, P., 2002. Statistics of extremes in hydrology. Advances in water resources, 25(8-12), pp.1287-1304.

Katz, Richard (2013) Statistical Methods for Nonstationary Extremes, Chapter 2 in A. AghaKouchak et al. (eds.), Extremes in a Changing Climate, Water Science and Technology Library 65, DOI 10.1007/978-94-007-4479-0 2,

Doing science the right way

Reproducibility crisis in several fields - open science movement as a result.

Replicability (i.e. being able to re-run the analysis) should be a given.

Start any project in a replicable way: literate programming and programmatic interaction with data (access, manipulation, analysis).

In R (and Python) this is increasingly feasible.

Slides code at github.com/ilapros - done in rmarkdown