BAYESIAN MODELLING OF EXTREME EVENTS

Jan Picek

Technical University of Liberec, Czech Republic

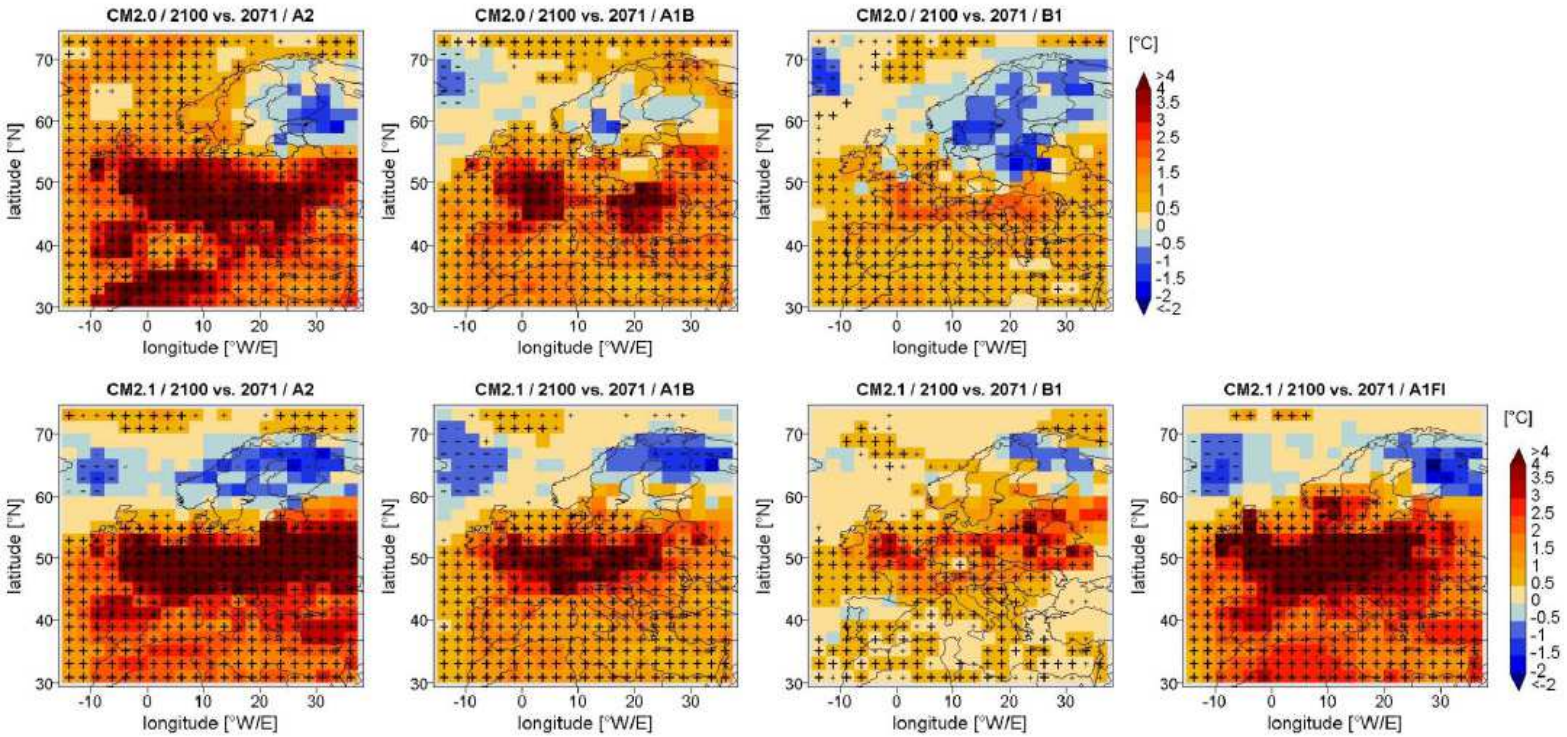
Statistical hydrology - STAHY 13, Kos, Greece

Motivation

Development of extreme value models with time-dependent parameters in order to estimate (time-dependent) high quantiles of maximum daily air temperatures over Europe in climate change simulations (1961-2100).

Kyselý, Picek and Beranová (2010): Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold, Global and Planetary Change, 72, 55-68.

Data motivation



Differences between 20-yr return values of TMAX estimated using the non-stationary POT model for years 2100 and 2071. Large (small) crosses mark gridpoints in which the estimated 90% (80%) CIs do not overlap.

Theoretical models

Fisher-Tippett Theorem: "If suitable normalized maxima converge in distribution to a non-degenerate limit, then the limit distribution must be an extreme value distribution."
$\Longrightarrow$ Method block maxima

Threshold view – it is reasonable to involve all values exceeding a given high threshold u.

The method is known as "peaks-over-threshold" (POT) and leads to the Poisson process model for threshold exceedances and the Generalized Pareto distribution (GPD) for their magnitudes.

Pickands (1975) showed that the limiting distribution of normalized excesses of a threshold u as the threshold approaches the endpoint u_{end} is the GPD.

Theoretical models

The classical block maxima and POT methods assume stationarity which is often violated in climatology and hydrology by the presence of a trend or long-term variability in the data.

If we describe a variable of primary interest by using covariate information (time index, variables based on atmospheric circulation ...)
$\Longrightarrow$ an approach based on the theory of point processes developed by Smith (1989) and Coles (2001).

The method leads to a likelihood function that can be treated in a usual way to obtain maximum likelihood estimates, standard errors and confidence intervals of the model parameters. One of its main advantages is that it enables a straightforward incorporation of time-dependency (or other variable of inerest) of parameters of the extreme value distribution. BUT also the threshold may depend on the covariates.

Bayesian analysis

- We observe the data $\boldsymbol{x}$.
- We model them as realizations of random variables $\boldsymbol{X} = X_1, ..., X_n$ – density function $f(\boldsymbol{x}|\boldsymbol{\theta})$.
- $\boldsymbol{\theta} \in \Theta \subset R^p$ is an unknown parameter.
- $g_0(\boldsymbol{\theta})$ – prior distribution .
- The posterior distribution of parameter given the observed data:

$$g(\boldsymbol{\theta}|\boldsymbol{x}) = C(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\theta}) g_0(\boldsymbol{\theta}).$$

- The posterior distribution of model parameters given the data - the MCMC methods

Let $X_1,\ X_2, \ldots$ be iid random variables with distribution function F and let $M_n = \max(X_1, \ldots, X_n)$.
Suppose that n is large, so that the distribution of M_n can be approximated by the GEV(μ, σ, ξ). Then for large threshold u exceedances is approximately a non-homogeneous Poisson process with intensity function

$$\lambda(x) = \frac{1}{\sigma}\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-(\xi+1)/\xi}, \quad u < x$$

The log-likelihood function can be derived (Coles, 2001) as

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = -\left\{1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right\}^{-1/\xi} + \sum_{i=1} n_u \log(\lambda(x_{(i)})),$$

where n_u of the n observation exceed the threshold u and $x_{(i)}$ denotes the ith exceedence, $\boldsymbol{\theta} = (\mu, \sigma, \xi)$.

Bayesian analysis

If we assume that the data $x = (x_1, \ldots, x_n)$ are independent realizations from GPD($\boldsymbol{\theta}$), the log-likelihood function is

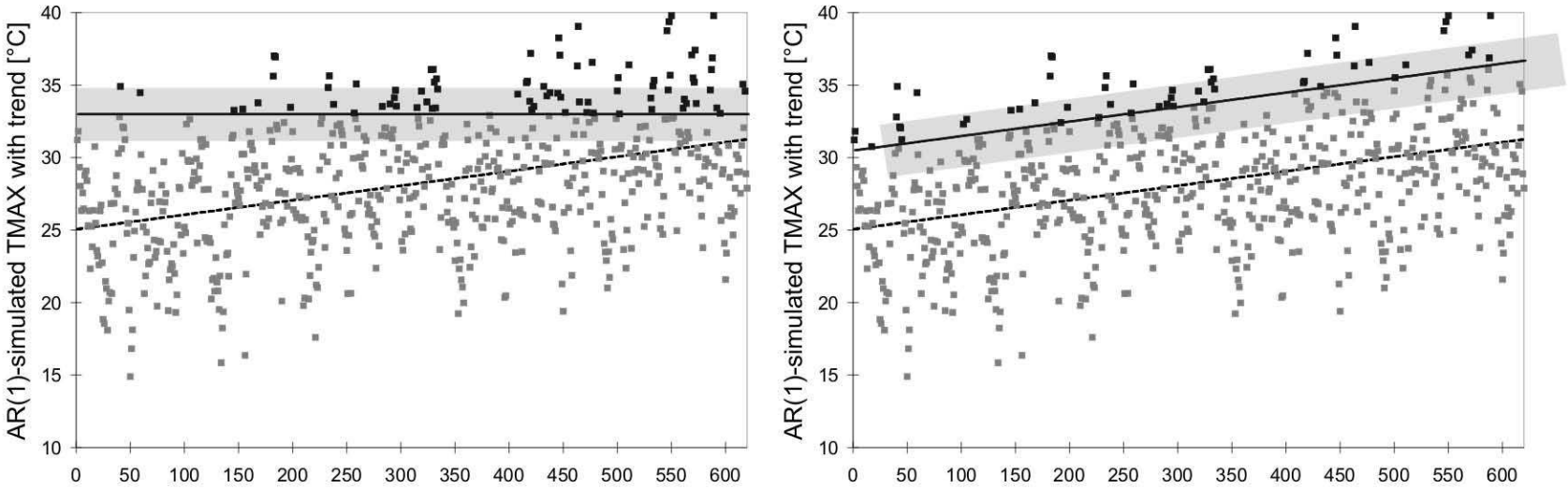
$$f(\boldsymbol{x}|\boldsymbol{\theta}) = -n \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log \left[1 + \xi(x_i - \mu)/\sigma\right],$$

Construction of prior Distribution – trivariate normal

Posterior distribution - Markov Chain Monte Carlo (MCMC) technique

BUT also the threshold may depend on the covariates.

TIME-DEPENDENT THRESHOLD



When a significant trend is present in the data, no fixed threshold in the POT models is suitable over longer periods of time: there are either too few (or no) exceedances over the threshold in an earlier part of records or too many exceedances towards the end of the examined period.

We use of a time-dependent threshold based on the quantile regression methodology.

Regression quantiles

Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}, \tag{1}$$

where $\mathbf{Y}$ is an $(n \times 1)$ vector of observations, $\mathbf{X}$ is an $(n \times (p+1))$ matrix, $\boldsymbol{\beta}$ is the $((p+1) \times 1)$ unknown parameter ($p \geq 1$)and $\mathbf{E}$ is an $(n \times 1)$ vector of i. i. d. errors.

We assume that the first column of $\mathbf{X}$ is $\mathbf{1}_n$, i.e. the first component of $\boldsymbol{\beta}$ is an intercept.

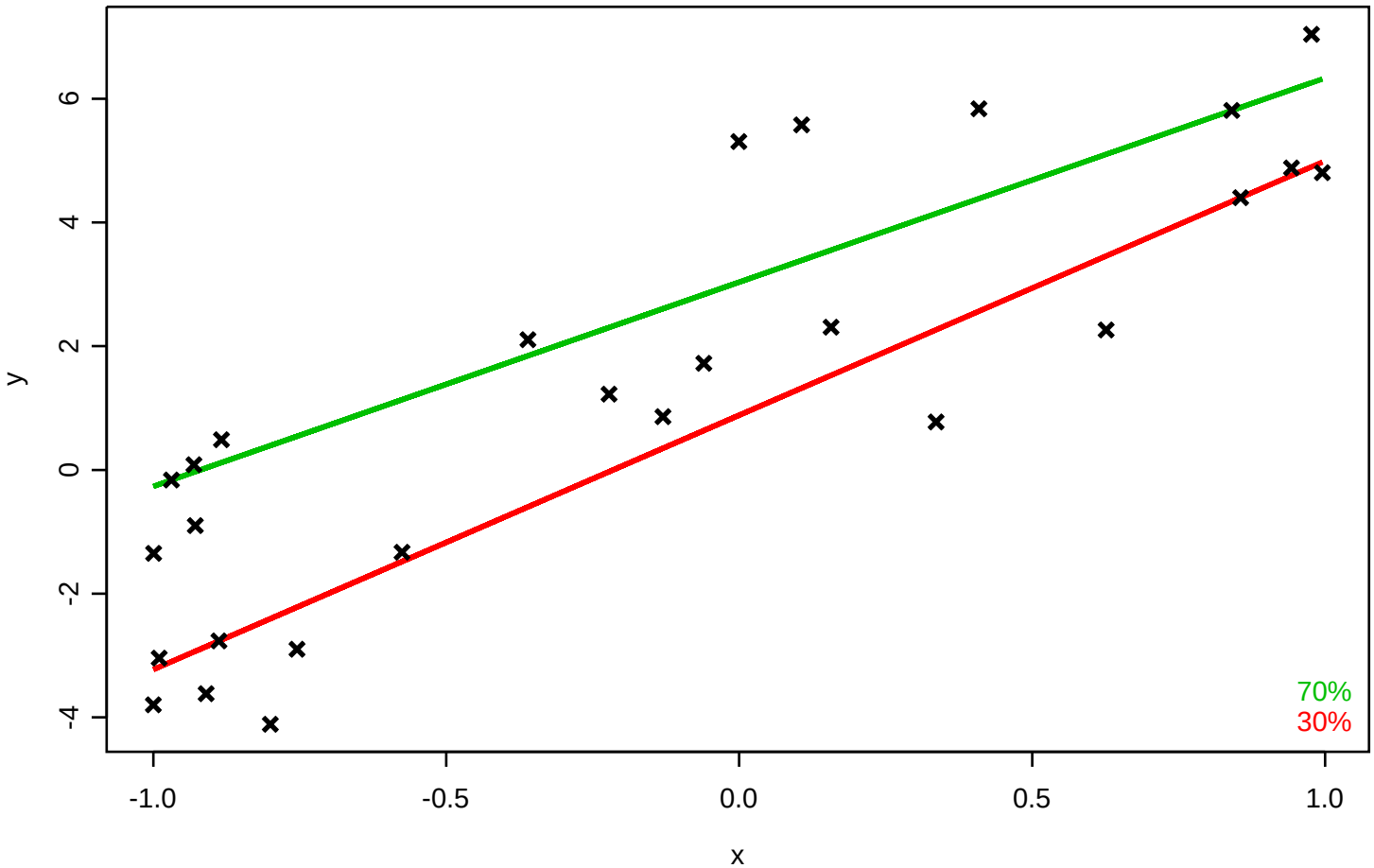
R. Koenker a G. Basset (1978) defined the α-regression quantile $\widehat{\boldsymbol{\beta}}\,(\alpha)$ $(0 < \alpha < 1)$ for the model (1) as any solution of the minimization

$$\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i'\mathbf{t}) := \text{min}, \quad \mathbf{t} \in I\!R^{p+1}, \tag{2}$$

where

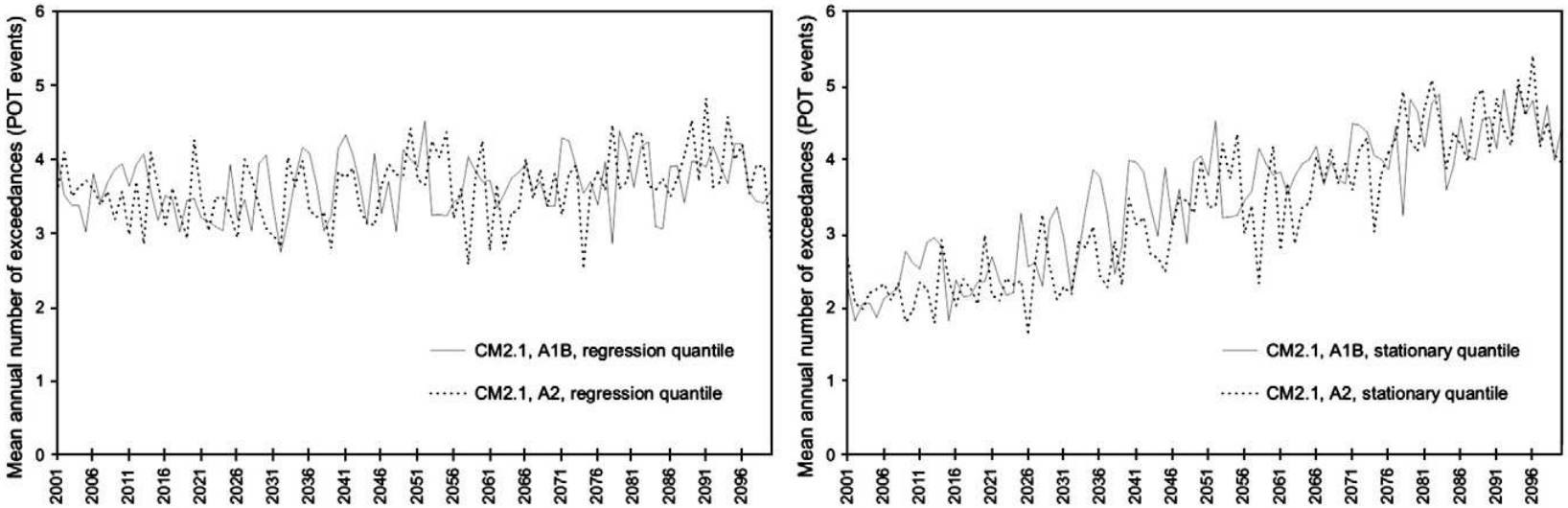
$$\rho_\alpha(x) = x\psi_\alpha(x),\ x \in I\!R^1 \text{ and } \psi_\alpha(x) = \alpha - I_{[x<0]},\ x \in I\!R^1. \tag{3}$$

Regression quantiles



The advantage is that many aspects of usual quantiles and order statistics are generalized naturally to the linear model.

Regression quantiles



Mean annual number of exceedances above the threshold (averaged over gridpoints) for the 95% regression quantile and the 95% quantile.

Bayesian analysis

If we assume that the data $x = (x_1, \ldots, x_n)$ are independent realizations from GPD($\boldsymbol{\theta}$), the log-likelihood function is

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = -n \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log \left[1 + \xi(x_i - \mu)/\sigma\right],$$

Construction of prior Distribution – trivariate normal

Posterior distribution - Markov Chain Monte Carlo (MCMC) technique

But we use of a time-dependent threshold based on the regression quantiles $\Longrightarrow$ is GPD is a suitable model ?

The linear regression model: Picek, Dienstbier (2013) - the limiting distribution of normalized excesses of the threshold based on the regression quantiles is the GPD.

$\Longrightarrow$

The log-likelihood function is

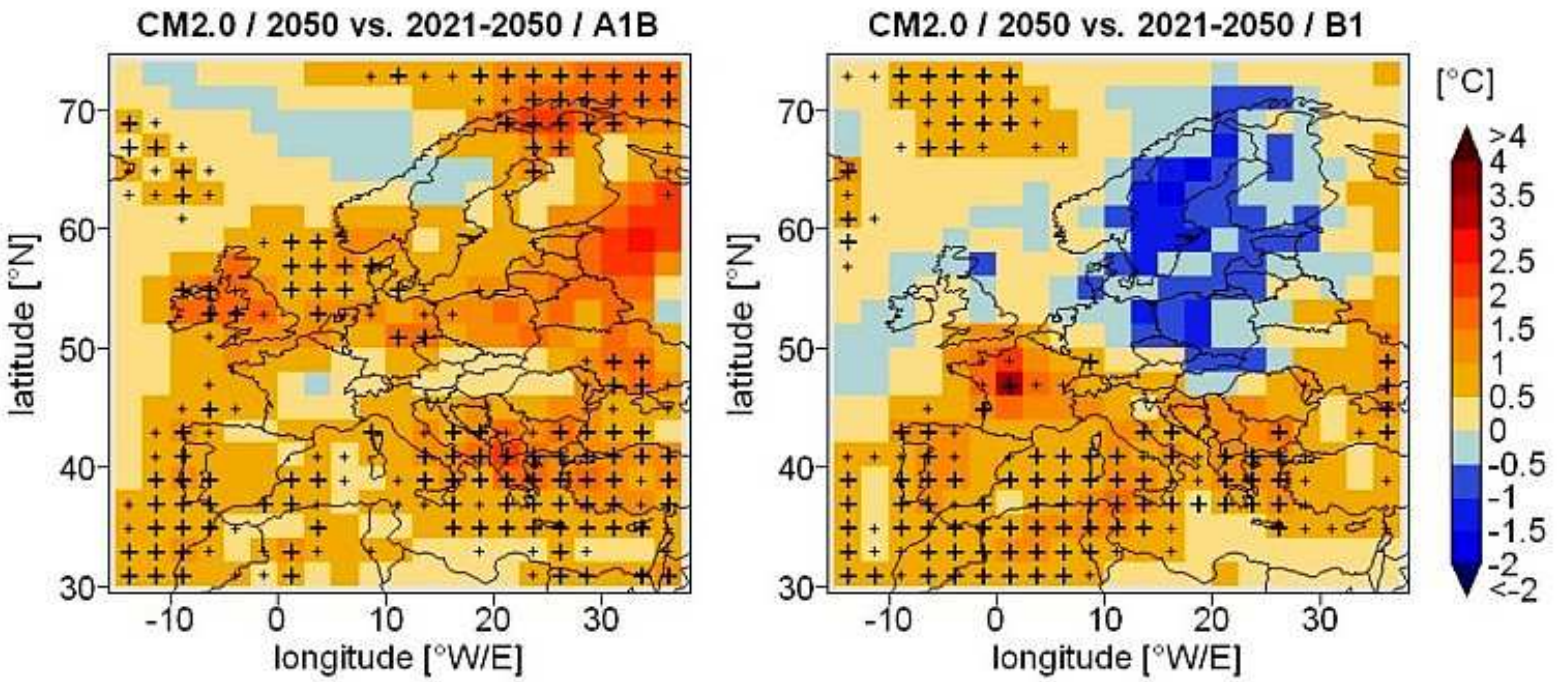
$$f(\boldsymbol{y}|\boldsymbol{\theta}) = -n\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=m_n-k+1}^{m_k}\log\left[1+\xi(\hat{\beta}_{n,0}\left(\tau_i\right)-\mu)/\sigma\right],$$

where $\hat{\beta}_{n,0}$ is intercept component of $\widehat{\boldsymbol{\beta}}_n$

Construction of prior Distribution – trivariate normal

Posterior distribution - Markov Chain Monte Carlo (MCMC) technique

Results



Differences between 20-yr return values of TMAX estimated using bayesian POT model for year 2050 and stationary POT model over 2021-2050. Large (small) crosses mark gridpoints in which the estimated 90% (80%) CIs do not overlap.

Conclusions

- The proposed method with thresholds estimated using regression quantiles is computationally straightforward
- Picek, Dienstbier (2013) showed that the limiting distribution of normalized excesses of a regression quantile threshold is the Generalized Pareto.
- We can use usual tools of Bayesian analysis in the models based on regression quantiles.
- The extension of the Poisson process log-likelihood for linear trend of parameters is similar.