On the effects of model misspecification in the study of non-stationary series of maxima: a stochastic simulation perspective

Tipaluck Krityakierne [1] & David Ginsbourger [1] & Jörg Franke [2,3] &
Christoph Welker [2,3] & Olivia Martius [2,3] & Martin Grosjean [2,3]

[1] *Department of Mathematics and Statistics, University of Bern, Switzerland*:
tk338@cornell.edu, ginsbourger@stat.unibe.ch
[2] *Institute of Geography, University of Bern, Switzerland*: franke@giub.unibe.ch,
christoph.welker@giub.unibe.ch, olivia.romppainen@giub.unibe.ch
[3] *Oeschger Centre for Climate Change Research, University of Bern, Switzerland*:
martin.grosjean@oeschger.unibe.ch

Résumé. La prise en compte de non-stationnarités dans l'étude des séries de maxima est un sujet crucial pour la quantification des risques liés à l'évolution du climat. Pour autant, lorsqu'une série de valeurs extrêmes est modélisée via la loi d'extremum généralisée (GEV), il peut arriver aux practiciens de ne pas tenir compte de possibles non-stationnarités, ou encore de tronquer les données pour réduire l'influence de tendances passées. Ici nous adoptons une démarche de simulation stochastique pour étudier les effets d'une mauvaise spécification de modèle sur l'erreur d'estimation des niveaux de retour dans le cas où les données simulées suivent indépendamment des distributions GEV, avec des paramètres de location dépendant ou non du temps. Nos résultats suggèrent qu'en présence d'une tendance linéaire en temps, l'approche par troncature permet de mieux estimer les niveaux de retour pour de petites périodes de retour, mais dégrade fortement l'estimation en ce qui concerne les plus grandes périodes de retour. Nous présenterons finalement des résultats obtenus sur des séries de maxima annuels issues de mesures climatiques et hydrologiques enregistrées sur le territoire Suisse depuis plus d'un siècle.

Mots-clés. Valeurs extrêmes, GEV, Séries temporelles, Sciences du climat

Abstract. Accounting for possible non-stationarities in series of maxima is of crucial importance for quantifying risks in a changing climate. However, when appealing to models relying on the Generalized Extreme Value distribution, it happens that practitioners do not take such non-stationarities into account, or simply truncate data sets in order to reduce the influence of past trends. Here we adopt a stochastic simulation approach for studying the effects of model misspecifications on return level estimation errors in the case of GEV-distributed simulated data, both with fixed and time-varying location parameters. Our results suggest that in the case of a location parameter with a linear trend in time, truncating the data does lead to an improved estimation of return levels with small return periods, but turns out to degrade estimation for larger return periods.

Finally, we will present results obtained on series of yearly maxima from climatological and hydrological series of measures recorded in Switzerland over more than a century.

Keywords. Extreme values, GEV, Time series, Climate sciences

1 Introduction

The study of extreme values in non-stationary contexts have recently received a lot of attention in climate research [1, 3, 4]. While a few non-stationary approaches have been proposed to study series of extreme events, stationary GEV models are still commonly used in practice. One simple rule of thumb, yet a practical approach, when working with non-stationary extremes under a stationary assumption is to not use all available data when calculating a return level but to first truncate the data so as to take only a shorter subset of observations closer to the current state. In this work, we attempt to answer several questions arising as a consequence of stationarity misspecification. The questions include, but are not limited to: What is the risk of miscalculation by assuming stationarity? Is it true that the location parameter at present or in the future are better estimated if one truncates the data set? What are the effects of parameter mis-estimation on short- and long-term estimated return levels? To this end, the analysis via stochastic simulation is carried out on non-stationary Generalized Extreme Value (GEV) data under several model assumptions. Additional analyses of climatological and hydrological series of extremes recorded in Switzerland over more than a century will also be discussed.

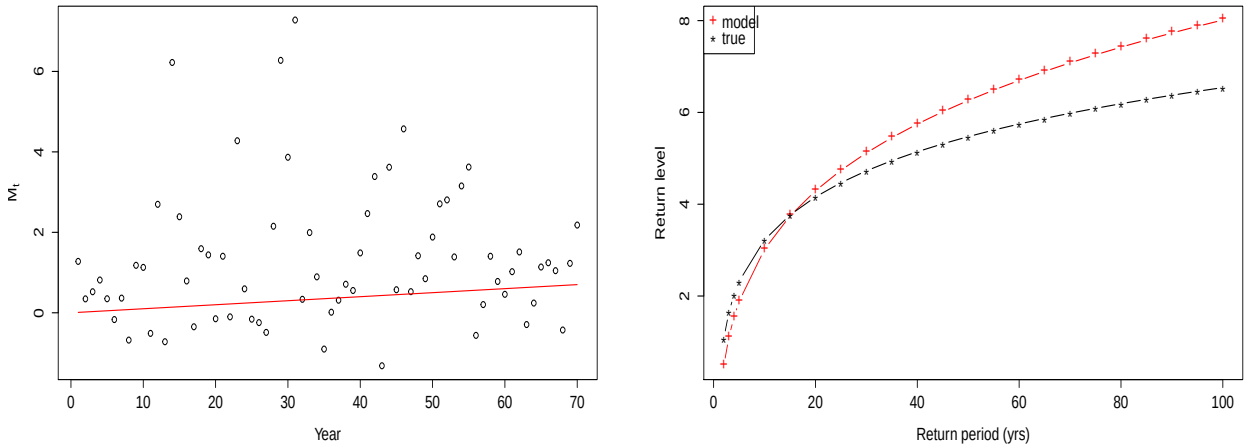
2 Methodology

2.1 Simulated non-stationary GEV series

Series of stochastically independent GEV distributed observations (M_t) are simulated. In our benchmark experiment, the shape and scale of the GEV distribution remain constant and only the location parameter varies as a function of time. That is, we generate $M_t \sim \text{GEV}(\mu(t), \sigma, \xi)$ independently, with a location parameter evolving linearly in time:

$$\mu(t) = \mu_0 + \mu_1 t. \tag{1}$$

For illustration, one realization of (M_t) simulated with $\mu_0 = 0$, $\mu_1 = 0.01$, $\sigma = 1$, and $\xi = 0.1$ together with the underlying trend are shown in Figure 1a.



(a) A realization of (M_t), with $\mu_0 = 0$ and $\mu_1 = 0.01$

(b) The true return levels and the estimated return levels with 70y-SM model at the reference time $t_0 = 70$

Figure 1: Simulation of a non-stationary GEV series and comparison between actual and estimated return levels (at $t_0 = 70$) and those estimated under stationarity assumption.

2.2 Fitted GEV Model

First we assume stationarity and fit a GEV distribution to (M_t) using the whole 70-year data set. We refer to this model as 70y-SM. Using the `'extRemes'` R package [6], the estimates of the parameters obtained from the fitted model are $\hat{\mu} = 0.0945$, $\hat{\sigma} = 1.012$, and $\hat{\xi} = 0.2136$ while the true parameters (at $t_0 = 70$) are $\mu = \mu_0 + 70\mu_1 = 0.7$, $\sigma = 1$, and $\xi = 0.1$, respectively.

We see that the estimated parameters depart from the true underlying parameters, especially for the location and shape parameters on this example. Can we find better estimates by truncating the data set? The answer is yes. If one suspects a linear trend with respect to time in the location parameter, a simple rule of thumb is to restrict attention to the most recent observations only. For example, instead of using the full data (as in the 70y-SM model), one can fit the data to a smaller window of, say, the last 30 years. We will refer to this model as 30y-SM.

We replicate the experiment 500 times. Box-plots of the distributions of estimated parameters inferred from both the 70y-SM and the 30y-SM models are shown in Figure 2 together with the true parameters (red lines). We see that, although with a higher variance, the location parameter is better estimated with the truncated data set while the scale and shape parameters are better estimated with the full data set.

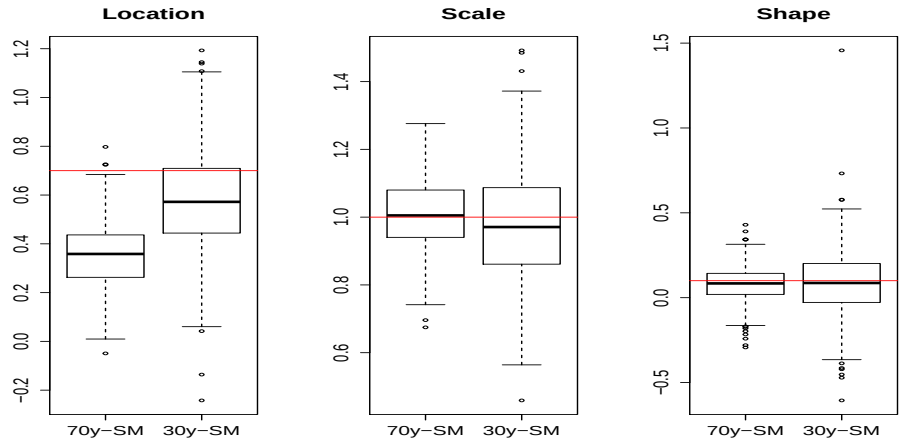


Figure 2: Distribution of estimated parameters

2.3 Different Criteria for Comparing Models

In addition to the estimates of GEV parameters, often of interest to practitioners is the T-year return level, defined as its $1/T$-level quantile z_p. Its analytical formula as a function of the GEV parameters is recalled in Property 1.

Property 1. Given a probability $p > 0$, the return level z_p associated with the return period $T = 1/p$ is defined by [3]:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], & \text{for } \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\}, & \text{for } \xi = 0, \end{cases} \tag{2}$$

where μ, σ, ξ are location, scale, and shape parameters of the GEV distribution.

A comparison between the true return levels at the reference time $t_0 = 70$ and the ones estimated based on the fitted 70y-SM model (one replicate) is shown in Figure 1b.

In addition to return levels, one may also be interested in knowing how close the distribution of the real and the fitted model are. One way to handle this is to use the Hellinger distance (Hdist) [5] to quantify the similarity between two absolutely continuous probability distributions. The definition of Hellinger distance is given in Definition 1.

Definition 1. Given f and g, two probability density functions with respect to a reference measure λ, the squared Hellinger distance between f and g is defined by

$$\text{Hdist}^2(f, g) = \frac{1}{2}\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 d\lambda(x). \tag{3}$$

2.4 A few experimental observations

We compare the mean of the absolute error of the return level as a function of the return period for several values of μ_1. The absolute error of the return level is defined by

$$\mathrm{ERL} = \left|\mathrm{RL} - \hat{\mathrm{RL}}\right|, \tag{4}$$

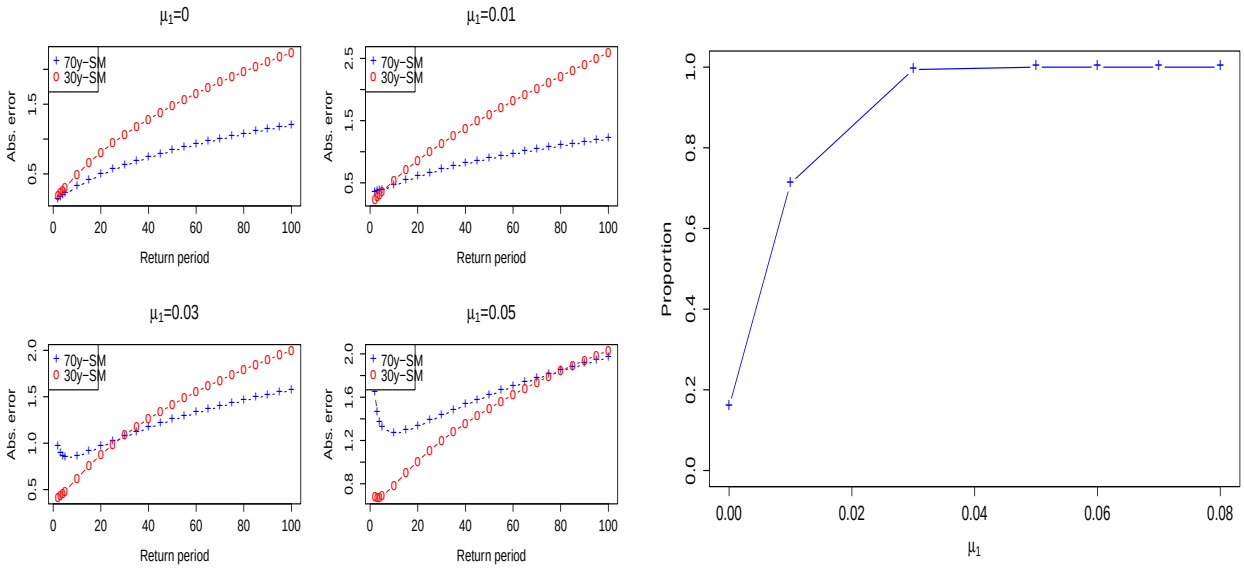
where RL and $\hat{\mathrm{RL}}$ are the true return level and the estimated return level based on a fitted model, respectively.

The results at the reference time $t_0 = 70$ years are given in Figure 3a. Under the stationary assumption, we can see that when μ_1 is small, the ERL of the 30y-SM is larger than that of the 70y-SM. As μ_1 increases, the ERL of the 70y-SM at low return periods becomes larger. In particular, when $\mu_1 = 0.03$, the ERL curve of the 70y-SM at low return periods ($T < 25$) starts to tilt upwards making the two curves cross each other at around $T = 25$. Therefore, the ERL of the 30y-SM is lower in the range of $T < 25$, but it is still higher than the 70y-SM when $T > 25$. When $\mu_1 = 0.05$, almost the whole ERL curve of the 70y-SM moves upwards, and so the ERL of the 30y-SM is smaller at almost all levels of the return period.

Next, we discuss the results based on the Hellinger distance between true and estimated GEV distributions at the reference time $t_0 = 70$. We keep track of the number of times Hdist obtained from the 30y-SM is smaller (closer to the true density) than Hdist obtained from 70y-SM and calculate the proportion (out of 500 replications). Figure 3b shows this proportion as a function of μ_1. We can see that under the stationary assumption, the proportion is increasing as μ_1 increases and reaches 1 at $\mu_1 = 0.03$, which indicates that the estimated density of the 30y-SM model is closer to the true density –in the Hellinger sense– than that of the 70y-model.

A very important conclusion one can draw from Figure 3a is that at high value of μ_1 ($\mu_1 = 0.03, 0.05$) what you gain on the location parameter by truncating the data set, you lose it on the scale and shape parameters. So, with the 30y-SM model, you get better at estimating short-term return levels but worse at long-term ones. In other words, longer-term return levels depend more on the distribution of the tail, for which having a poor estimate of the location parameter is not dramatic while poorly estimating the shape parameter has severe consequences.

Ongoing work concerns studying the case where the trend depends non-linearly on time, for instance oscillatory trends in location standing for multi-decadal variability in climate. Besides this, applications on series of yearly maxima from climatological and hydrological series of measures recorded in Switzerland over more than a century are currently in progress.



(a) Absolute error of the return level

(b) Proportion that Hdist-30y is better than Hdist-70y

Figure 3: ERL and Hdist of the 30y-SM and 70-SM at different values of μ_1

Bibliography

[1] Santiago Beguería, Marta Angulo-Martínez, Sergio M Vicente-Serrano, J Ignacio López-Moreno, and Ahmed El-Kenawy. Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis: a case study in northeast spain from 1930 to 2006. *International Journal of Climatology*, 31(14):2102–2114, 2011.

[2] George EP Box and Norman R Draper. *Empirical model-building and response surfaces*. John Wiley & Sons, 1987.

[3] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[4] Eric Gilleland. *extRemes: Extreme Value Analysis*, 2012.

[5] Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.