Erasmus University Rotterdam

MSc Thesis Econometrics

A peak-over-threshold approach for modelling covariate relations in extreme value distributions using Conditional Density Networks

Academic supervisor:
Dr. Phyllis Wan
Second assessor:
Prof. Dr. Chen Zhou

Author:
Rogier Buck, 504938

August 5, 2021

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

ERASMUS UNIVERSITY ROTTERDAM

Abstract

This paper proposes an alternative technique for non-stationary extreme value analysis using conditional density networks (CDNs) based on the peak-over-threshold (POT) approach: GP-CDN. A standard POT approach and a variant with linear thresholds determined by quantile regression (QR) are examined. A Monte Carlo simulation study is conducted to compare the RMSE of return levels estimated by the new GP-CDN technique vs. the existing GEV-CDN. A new algorithm is introduced to generate data that is simultaneously valid for GEV and GP distributions. The simulation study finds that GP-CDN mostly performs on par with GEV-CDN. For high return periods and when scale is the primary non-stationary feature of the data, GP-CDN, and particularly its QR variant, performs better. Application of the GP-CDN technique to precipitation data showcases its applicability to real-world data settings. Together, it shows that GP-CDN provides practitioners with a viable alternative for the estimation of non-stationary extremes.

Acknowledgements

I would like to thank my supervisor Dr. Phyllis Wan for introducing me to this interesting field of research, and for her valuable discussions and feedback. Furthermore, I would like to thank the second assessor Prof. Dr. Chen Zhou. Finally, my gratitude goes out to my family for their continued support.

Contents

List of abbreviations

AICc	Akaike information criterion corrected for small sample sizes
BIC	Bayesian information criterion
BM	block maxima
CaDENCE	Conditional Density Estimation Network Creation and Evaluation
cdf	cumulative distribution function
CDN	conditional density network
DGP	data-generating process
EVA	Extreme Value Analysis
GEV	Generalised Extreme Value
GEV-CDN	GEV conditional density network
GML	generalised maximum likelihood
GP	Generalised Pareto
GP-CDN	GP conditional density network
iid	independent and identically distributed
ML	maximum likelihood
MLP	multilayer perceptron
OLS	ordinary least squares regression
pdf	probability density function
PDO	Pacific Decadal Oscillation
POT	peak-over-threshold
POT-C	POT approach with constant threshold
POT-QR	POT approach with QR threshold
Q-Q	quantile-quantile
QR	quantile regression
RMSE	root mean squared error
sd	standard deviation
SOI	Southern Oscillation Index

1 Introduction

Commonly, statistical analysis concerns the average behaviour of data. However, for many real-world applications we are more interested in the likelihood of rare events that lie in the tails of probability distributions.

In recent decades, globalisation has made supply chains and our society at large more interconnected. Historically, societal systems exhibited spare capacity that could function as shock absorbers for rare (extreme) events. However, business leaders and policymakers all over the world, engaged in a continuous quest for efficiency gains, have chipped away at these shock absorbers. The house-of-cards structure of our society has increased our vulnerability to extreme shocks, such as those caused by natural phenomena. Just in the first half of 2021, with the Texas power crisis and the Canadian heatwave, we have seen two harsh reminder of the societal disruption that extreme events can cause and the responsibility of policymakers in mitigating associated risks. Crises like these underscore the need for a) accurate models for estimating the likelihood of extreme events; and b) models that can shed light on the ways in which covariates relate to extreme events. Such an understanding benefits the construction of protection systems against adverse effects and makes systems more robust.

Extreme Value Analysis (EVA) provides a useful framework for evaluating the tails of distributions. The Extreme Value Theorem was introduced in Fisher and Tippett (1928). Under reasonable assumptions, they show that the maximum of a sample of independent and identically distributed (iid) random variables asymptotically converges in distribution to one of three families of distributions, the Fréchet-, Weibull-, or Gumbel-distribution, regardless of the distribution of the data-generating process (DGP). Jenkinson (1955) showed that these three families of distributions can be combined in one distribution, named the Generalised Extreme Value (GEV) distribution. The GEV has three parameters for location, scale, and shape. Inference methods that model the GEV are usually named the block maxima (BM) approach.

An alternative method to model extreme values is the peak-over-threshold (POT) approach. The groundwork leading to this method was laid out in Pickands III (1975) and its statistical properties were analysed in more detail in Davison and Smith (1990). They show that exceedances of a sample of iid random variables above a reasonably high threshold can be modelled with the Generalised Pareto (GP) distribution. Inference for the GP distribution requires the choice of a suitable threshold, after which a scale parameter and a shape parameter can be estimated. The GEV and GP parameters are closely related; the parameters of the GP distribution are uniquely determined by the associated GEV distribution. In particular, the shape parameter is identical.

Estimation of these parameters can be done via various parametric and semi-parametric methods such as maximum likelihood (ML) (Smith, 1985), penalised maximum likelihood (Coles and Dixon, 1999), generalised maximum likelihood (GML) (El Adlouni et al., 2007), probability-weighted moments (Hosking et al., 1985), L-moments (Hosking, 1990), Hill estimator (Hill, 1975), Pickands estimator (Pickands III, 1975) as well as Bayesian methods.

Fundamentally, classical EVA assumes that the series of extremes is of sufficient length, that its elements are iid, and that it is stationary. These fundamental assumptions are not always satisfied in practice. We often encounter non-stationary extremes, e.g., more extreme floods or temperatures due to anthropogenic climate change (Jain and Lall, 2001; Kharin and Zwiers, 2005). A model with constant distribution parameters is thus not suitable; instead we require the parameters to model trends or other relations with covariates. Zhang et al. (2004) used a Monte Carlo simulation study to compare several methods for trend detection in the magnitude of extreme values and found that models that model the trend via the distributional parameters provide the highest power of detection of significant trends. ML-based methods provide an obvious estimation method that can handle relations with covariates, as they explicitly use the probability density function (pdf) of each observation in the (log-)likelihood. Davison and Smith (1990) provide an early overview of how to model GEV and GP parameters as functions of covariates in an ML framework. Successful applications of this approach are abundant in the literature in a wide range of fields such as hydrology, meteorology, oceanography, insurance and finance, see e.g., Davison and Smith (1990), Embrechts et al. (1996) and Coles (2001).

Modelling the relation between distribution parameters and covariates as a linear function is not always appropriate. Kharin and Zwiers (2005) allowed for non-linearity in the parameters by estimating linear trends in GEV parameters in overlapping rolling windows. Chavez-Demoulin and Davison (2005) introduced generalised additive modelling for sample extremes, which uses spline estimators. Both studies showed positive results when relaxing the linearity assumption, but they still require an a priori specification of the form of interaction between parameters and covariates.

Cannon (2010) introduced a more flexible approach for the non-stationary GEV model that does not require modelling an a priori specification of the interaction, by applying a probabilistic extension of the multilayer perceptron (MLP) neural network named the GEV conditional density network (GEV-CDN). Rather, the flexible nature of the MLP allows the form of the interaction to be absorbed into the estimation of the weights between the nodes in the network. He showed that the CDN exactly replicated several synthetic non-stationary GEV models introduced in El Adlouni et al. (2007) and accurately approximated others. Cannon (2010) also applied the GEV-CDN to precipitation data introduced in El Adlouni et al. (2007) and identified a non-linear relation among covariates and the GEV parameters. The GEV-CDN model has later been used in fields such as of meteorology (Vasiliades et al., 2015) and hydrology (Shrestha et al., 2017). However, even though the flexibility of the GEV-CDN model seems promising and an open-source package for GEV-CDN (Cannon, 2011) is available in the statistical programming language `R` (R Core Team, 2019), its use is not widespread in the literature.

The POT approach has slowly replaced the BM approach as the favoured method among practitioners for modelling extremes because it is believed to make more efficient use of available data, which is often scarce (Davison and Huser, 2015). This belief is confirmed for a stationary setting in a simulation study by Caires (2009), as long as there are two or more exceedances per year available. For non-stationary, serially dependent data, Caires (2009) also finds that the POT approach performs on par or better than the BM approach across a wide range of settings. Bücher and Zhou

(2018) however show that the efficiencies of the two approaches are dependent on the underlying DGP, and that no general winner is identifiable.

Even so, there is no literature available that applies the CDN framework to the GP distribution of the POT approach; a technique that could be named GP-CDN. One complicating factor could be that selecting a suitably high threshold is not straightforward in a non-stationary context. If for example an upward trend is present in the location of the extremes, using a constant threshold will produce invalid results as too few extremes are present in the early part of the sample and too many in the later part. To mitigate this issue, Kyselỳ et al. (2010) introduce and model a non-stationary threshold for the POT approach by using the quantile regression (QR) technique introduced in Koenker and Bassett Jr (1978); such an approach could be named POT-QR.

By means of a Monte Carlo simulation study, this paper aims to compare and contrast the BM and POT approaches when applying CDNs to model flexible, non-linear relations between covariates and the respective distribution parameters. Given the connection between the GEV and GP distribution parameters, and encouraged by the greater efficiency of threshold exceedances and the promising results shown by the flexibility of the GEV-CDN, this paper hypothesises that the POT approaches will yield favourable results.

The methodology will be explained in Section 2. The simulation design will be introduced in Section 3, and the simulation results can be found in Section 4. Afterwards, an application to real precipitation data is shown in Section 5. Finally, the conclusion can be found in Section 6.

2 Methodology

This section starts off with an introduction to the BM and POT approaches for EVA. Afterwards, the CDN architecture will be outlined. Lastly, the methods for parameter estimation and model selection and comparison will be described.

2.1 Block maxima and peak-over-threshold approaches for EVA

First, the non-stationary GEV and GP distributions and their likelihood functions will be reviewed. Subsequently, quantile regression for POT thresholds will be explained. Finally, an overview of the EVA frameworks employed in this research is given.

2.1.1 The non-stationary Generalised Extreme Value distribution

Let the maximum of a sequence $X_1, ..., X_n$ of n iid random variables, having a common distribution F, be denoted as $M_n = \max\{X_1, ..., X_n\}$. Then it can be shown that if there exist sequences of constant $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq y\right\} \to G(y) \quad \text{as } n \to \infty \tag{1}$$

for a non-degenerate distribution G, then G is a member of the GEV family with cumulative distribution function (cdf)

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

when $\xi \neq 0$ and $1 + \xi\frac{(y-\mu)}{\sigma} > 0$. The GEV distribution is thus characterised by three parameters, the location μ, the scale $\sigma > 0$, and the shape ξ. The shape parameter ξ is the main determinant for the tail behaviour and is sometimes also defined as $\kappa = -\xi$, as is done in El Adlouni et al. (2007) and Cannon (2010). The specific case where $\xi = 0$ requires other definitions of the distribution functions and is outside the scope of this research, but it poses no theoretical limitations to the proposed approach. This paper closely follows the notation Coles (2001) in the following theorems, while making necessary adjustments to allow for non-stationarity.

The pdf $f(y)$ for an observation from the GEV distribution with $\xi \neq 0$ is then

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

As noted in the introduction, various estimation techniques exist. ML will be used in this research. As usual, the likelihood function is defined as

$$\mathcal{L}(\mu(t), \sigma(t), \xi(t) \mid \mathbf{y}) = \prod_{t=1}^{n} f(y_t; \mu(t), \sigma(t), \xi(t))$$

where $\mathbf{y} = \{y_t, t = 1, ..., n\}$ denotes a series of n independent observations of maxima with possible non-stationary parameters $\mu(t), \sigma(t)$ and $\xi(t)$. Assuming the shape parameter $\xi(t) \neq 0$ for all t, The log-likelihood l_n of $\mathbf{y}$ is then

$$\begin{aligned} l_n\left(\mu(t), \sigma(t), \xi(t) \mid \mathbf{y}\right) = &-\sum_{t=1}^{n} \log\left(\sigma(t)\right) - \sum_{t=1}^{n}\left(1 + \frac{1}{\xi(t)}\right)\log\left[1 + \xi(t)\left(\frac{y_t - \mu(t)}{\sigma(t)}\right)\right] \\ &- \sum_{t=1}^{n}\left[1 + \xi(t)\left(\frac{y_t - \mu(t)}{\sigma(t)}\right)\right]^{-1/\xi(t)}. \end{aligned}$$

2.1.2 The non-stationary Generalised Pareto distribution

Let again the maximum of a sequence $X_1, ..., X_n$ of n iid random variables, having a common distribution F, be denoted as $M_n = \max\{X_1, ..., X_n\}$. Denote an arbitrary term in the sequence by X. Assume that F satisfies the conditions of the previous section for the GEV distribution, so that for large n,

$$\Pr\left\{M_n \leq z\right\} = F^n(z) \approx \exp\left[-\left\{1 + \xi\frac{(z-\mu)}{\sigma}\right\}^{-1/\xi}\right]$$

when $\xi \neq 0$ and $1 + \xi\frac{(z-\mu)}{\sigma} > 0$. Taking logarithms on the above gives

$$n \log F(z) \approx -\left\{1 + \xi\frac{(z-\mu)}{\sigma}\right\}^{-1/\xi}.$$

For large values of z, we can apply a first-order Taylor expansion on the logarithm, and get

$$1 - F(u) \approx \frac{1}{n}\left\{1 + \xi\frac{(u-\mu)}{\sigma}\right\}^{-1/\xi}, \tag{2}$$

where we have also rearranged the terms and renamed the argument u. Similarly, for $y > 0$, we have

$$1 - F(u+y) \approx \frac{1}{n}\left\{1 + \xi\frac{(u+y-\mu)}{\sigma}\right\}^{-1/\xi}. \tag{3}$$

We are now interested in the probability of the exceedance $y = X - u$ conditional on $X > u$

$$\Pr\left\{X > u + y \mid X > u\right\} = \frac{1 - F(u+y)}{1 - F(u)}, \quad y > 0. \tag{4}$$

Substituting Equations (2) to (3) into Equation (4) gives

$$
\begin{aligned}
\Pr\{X > u + y \mid X > u\} &\approx \frac{\frac{1}{n}\left\{1 + \xi\frac{(u+y-\mu)}{\sigma}\right\}^{-1/\xi}}{\frac{1}{n}\left\{1 + \xi\frac{(u-\mu)}{\sigma}\right\}^{-1/\xi}} \\
&\approx \left\{1 + \frac{\xi\frac{(u+y-\mu)}{\sigma}}{1 + \xi\frac{(u-\mu)}{\sigma}}\right\}^{-1/\xi} \qquad (5) \\
&\approx \left\{1 + \frac{\xi y}{\tilde{\sigma}}\right\}^{-1/\xi}
\end{aligned}
$$

where

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \qquad (6)$$

Thus, for sufficiently large threshold u, the cdf of the exceedance $y = x - u$ conditional on $x > u$ approximately follows the GP distribution

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}$$

when $\xi \neq 0$ and $1 + \xi\frac{(y-\mu)}{\sigma} > 0$. The GP distribution is characterised by two parameters, the scale $\tilde{\sigma}$ and the shape ξ, as well as the threshold u. In particular, the shape parameter ξ is identical for the GEV and GP distributions and the scale parameters are related via Equation (6).

Now for the likelihood function, suppose that $\mathbf{y} = \{y_t, t = 1, ..., k\}$ denote a series of k independent observations of exceedances of a threshold u with possible non-stationary parameters $\tilde{\sigma}(t)$ and $\xi(t)$. Then, assuming the shape parameter $\xi(t) \neq 0$ for all t, the log-likelihood l_k of $\mathbf{y}$ is

$$l_k\left(\mu(t), \tilde{\sigma}(t), \xi(t) \mid \mathbf{y}\right) = -\sum_{t=1}^{k} \log\left(\tilde{\sigma}(t)\right) - \sum_{t=1}^{k}\left(1 + \frac{1}{\xi(t)}\right)\log\left[1 + \frac{\xi(t)y_t}{\tilde{\sigma}(t)}\right].$$

2.1.3 Quantile regression for thresholds

A key aspect when using the POT approach for EVA is choosing an appropriate threshold. A common choice is to use a constant threshold, but this is not a very attractive choice when the location of extremes is non-stationary. Kyselỳ et al. (2010) however proposed to use the QR method of Koenker and Bassett Jr (1978) to determine a non-stationary threshold. Therefore, apart from constant thresholds, this research will employ linear QR fits based on low quantiles τ_{BM} of the associated BM data. The BM data is used to determine the thresholds for the POT-QR approaches because prior to choosing a threshold, POT data is obviously not yet available whereas BM data is readily available. The choice for linear QR reduces the flexibility of GP-CDN as it requires an a priori model specification for the threshold. In principle, more flexible QR methods such as nonlinear QR or smoothed additive QR could be employed. However, given that these alternative QR methods are prone to overfitting and because data is scarce, this research opts for linear QR. Hereafter, linear QR is referred to simply as QR.

QR is an intuitively easy-to-understand extension to ordinary least squares regression (OLS). Whereas OLS models the mean conditional on covariates, QR models a conditional quantile of the data. QR assumes that a quantile τ of an observation Y conditional on a covariate vector $\mathbf{X}$ can be explained by a linear function $Q_\tau(Y \mid \mathbf{X}) = \mathbf{X}^\top \beta(\tau)$. The technique followed from the intuition that a linear quantile regression on the data should produce roughly τn positive residuals and $(1-\tau)n$ negative residuals (Koenker, 2017). This leads to the minimisation problem of Equation (7) and its auxiliary function $\rho(\cdot)$ of Equation (8). A brief outline of the technique is given here, see Koenker and Bassett Jr (1978) and subsequent papers for a more elaborate overview.

Let $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top$ be an $(n \times p)$ matrix of covariates and let $Y = (y_1, ..., y_n)^\top$ be an $(n \times 1)$ vector of observations. Define quantiles τ on a range between 0 and 1, i.e., $\tau \in (0, 1)$. It can then be shown that the τ-regression quantile is the solution of the minimisation problem

$$\min_{\mathbf{b} \in R^p} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^\top \mathbf{b}) \tag{7}$$

where the function $\rho(\cdot)$ is defined as

$$\rho_\tau(x) = \begin{cases} \tau x & x \geq 0 \\ (\tau - 1)x & x < 0 \end{cases} \tag{8}$$

This minimisation problem can be defined as a linear program. The function $\rho(\cdot)$ is piecewise linear and convex, and the problem is thus solvable via several algorithms.

2.1.4 Overview of EVA frameworks

All in all, this research will compare five EVA frameworks, as outlined in Table 1 below. The threshold parameter choices for τ_{BM} and u represent high thresholds in this context. Using two options for each of the POT approaches partially addresses the well-known threshold selection problem, but other sufficiently high threshold parameter choices are obviously also possible. More light is shed on these parameter choices later on via Figures 1 to 4 and via the observations per year in Tables 4 to 7.

Table 1: Overview of EVA frameworks

	EVA framework				
Name	BM	POT-QR-1	POT-QR-2	POT-C-0	POT-C–1
Approach	BM	POT	POT	POT	POT
Threshold type	-	Quantile regression	Quantile regression	Constant	Constant
Threshold parameter	-	$\tau_{BM} = 0.1$	$\tau_{BM} = 0.2$	$u = 0$	$u = -1$

2.2 Conditional Density Networks

A MLP neural network with one hidden layer with J hidden nodes is considered. Given I covariates at time t, $\mathbf{x}_t = \{x_i(t), i = 1, ..., I\}$, define the output of the j^{th} hidden-layer node

$$h_j(t) = m\left(\sum_{i=1}^{I} x_i(t) w_{ji}^{(1)} + b_j^{(1)}\right)$$

where $w_{ji}^{(1)}$ denote the input-hidden-layer weights and $b_j^{(1)}$ the hidden-layer bias. The function $m(\cdot)$ is called the hidden-layer activation function. This function can be the identity function, which will result in a strictly linear CDN mapping. Alternatively, we could opt for a sigmoidal function such as the hyperbolic tangent function $\tanh(\cdot)$, which will result in a more flexible, non-linear mapping. We now move to the output layer. The value of the k^{th} output θ_k is defined as

$$\theta_k(t) = g_k\left(\sum_{j=1}^{J} h_j(t) w_{kj}^{(2)} + b_k^{(2)}\right)$$

where $w_{kj}^{(2)}$ denote the hidden-output-layer weights and $b_k^{(2)}$ the output-layer bias. The functions g_k denote the mapping functions for each of the outputs, which need not be identical and should be chosen to constrain the outputs to appropriate ranges.

Each of the K outputs $\theta_k(t)$ corresponds to a single distribution parameter. Define the full set of distribution parameters as $\boldsymbol{\theta}(t) = \{\theta_k(t), k = 1, ..., K\}$, then for GEV we have $\boldsymbol{\theta}(t) = \{\mu(t), \sigma(t), \xi(t)\}$ and for GP we have $\boldsymbol{\theta}(t) = \{\tilde{\sigma}(t), \xi(t)\}$. To constrain the outputs to appropriate ranges, we will choose the following mapping functions for g_k: identity function for $\mu(t)$, exponential function for $\sigma(t)$ and $\tilde{\sigma}(t)$, and the function $\xi^* \tanh(\cdot)$ for $\xi(t)$. By setting $\xi^* = 0.5$, we constrain ξ to $[-0.5, 0.5]$. Furthermore, we will force $\xi(t) = \xi$ to be constant over time. Therefore, the only function of the mapping function $\tanh(\cdot)$ for ξ is to constrain the range of appropriate values. The two assumptions on the shape parameter ξ are reasonable in this context (Cannon, 2010).

The CDN complexity can be adjusted in three ways: by choosing the activation function, by adjusting the number of hidden-layer nodes J, and by disconnecting weights. Herewith, we can create a hierarchy of increasing model complexity. See Cannon (2010) Figure 2 for an illustration of several GEV-CDN model architectures with increasing complexity. Creating GP-CDN architectures is straightforward and less complex, as it involves one less output. The hyperparameter option sets to be explored in this research are laid out below in Table 2.

Table 2: Hyperparameter option sets

	Hyperparameter option set			
Hyperparameter	HP1	HP2	HP3	HP4
Activation function	Identity	$\tanh(\cdot)$	Identity	$\tanh(\cdot)$
Number of hidden nodes	2-3	3	3	3
Fixed parameters	ξ	ξ	σ, ξ	σ, ξ

Note: Options 3 and 4 are useless for POT approaches as it leads to a stationary model. Identity function leads to a linear model, hence the number of hidden nodes equals the number of outputs.

2.3 Parameter estimation

The (log-)likelihood functions for the GEV and GP have now been defined as functions of non-stationary, covariate-dependent parameters and the CDN architecture with which to model these relations parametrically has been introduced. Simulation results in El Adlouni et al. (2007), building on the earlier work of Martins and Stedinger (2000), however show that the GML outperforms standard ML methods in a range of covariate-dependent GEV models. GML forces the shape parameter ξ to a physically valid range of values. It adds a penalty term to the likelihood in the form of a Beta distribution prior

$$\pi_\xi = Beta(\xi + 0.5, c_1, c_2).$$

The GML likelihood is then defined as the 'quasi'-posterior

$$\pi(\theta|x) \propto \mathcal{L}(x|\theta)\pi_\xi$$

where $\mathcal{L}$ denotes the likelihood of the GEV or GP distribution. The term 'quasi' is used because a prior is defined only for the shape parameter ξ. This definition of the Beta prior limits the domain to $-0.5 < \xi < 0.5$. El Adlouni et al. (2007) recommends setting c_1 and c_2 to 9 and 6, respectively; note that the definition of $\xi = -\kappa$ necessitates swapping the parameters. Cannon (2010) instead proposes to set c_1 and c_2 to 3.3 and 2, respectively, which widens the range of accepted values.

The GML will be the cost function of choice in this research, and we opt for Cannon's parameters for the Beta distribution. The cost function can be optimised to estimate the weights and biases using numerical solvers. A GP counterpart of the GEV-CDN `R`-package obviously does not yet exist. However, Cannon (2012) introduced the Conditional Density Estimation Network Creation and Evaluation (CaDENCE) package for `R`. This package extends his earlier work on the GEV-CDN and allows estimation of parameters of any user-specified, covariate-dependent distribution using the MLP architecture. This research takes the CaDENCE source code as its foundation, and makes changes as required for the specific purposes of this research.

2.4 Model selection and comparison

The most appropriate hyperparameter set for the CDN model for each of the BM and POT frameworks can be selected by comparison of the GML cost defined in the previous section. Alternatively, we could use the Akaike information criterion corrected for small sample sizes (AICc) or the Bayesian information criterion (BIC). These model selection criteria penalise the log-likelihood as a function of the number of model parameters. Given that the CDN models are not very complex, and thus the amount of parameters quite small, GML cost will likely be more informative.

It should be noted that a direct comparison of the GML costs or information criteria between the EVA frameworks is not possible. This is primarily because each of the frameworks deals with an overlapping but non-identical subset of the data, which causes their likelihoods to be in different orders of magnitude. The simulation algorithm outlined in Section 3, however, will present us with the true values of distribution parameters and hence with true quantities such as return levels. A direct comparison of these quantities will thus provide us with the best indication of performance. Therefore, the root mean squared errors (RMSEs) of return levels will be leading for both CDN model selection and EVA framework comparison. Student's t-tests are computed to test for significance differences between the RMSEs of the best BM models and the best POT models, which requires a partition of the Monte Carlo simulations.

GEV and GP return levels follow naturally from their distributions. The N-year return level for the GEV distribution, when $\xi \neq 0$, is defined as

$$z_n = \mu - \frac{\sigma}{\xi}[1 - (-\log(1 - \frac{1}{N}))^{-\xi}]. \tag{9}$$

The N-year return level for the GP distribution, when $\xi \neq 0$, is defined as

$$z_n = u + \frac{\tilde{\sigma}}{\xi}[(N n_y \zeta_u)^{\xi} - 1] \tag{10}$$

where n_y is the number of observations per year and ζ_u the rate of exceedance over threshold u.

For the GEV distribution, return levels are identical to quantiles, provided that one block corresponds to one year. True or estimated GEV return levels can then be computed directly from Equation (9) using the three true or estimated GEV parameters. For the GP distribution, return levels are very similar to quantiles, but the exceedance rate ζ also needs to be taken into account. The true exceedance rate can be computed via Equation (11), whereas the estimated exceedance rate can be approximated using the ratio of all exceedances relative to all observations. This estimation leads to a constant exceedance rate for all observations, which can lead to some bias if the true exceedance rate is not constant over time. However, the alternative of determining a unique exceedance rate for each year is deemed considerably worse; a given year only has a few exceedances and the variance of the annually determined exceedance rate over time will be enormous. After determining the exceedance rate, the true or estimated return levels can be computed directly from Equation (10) using the two GP parameters, the exceedance rate and the chosen threshold. The additional estimation of the exceedance rate should not result in an implicit advantage for the BM approach. Estimating the BM and POT return levels both require the estimation of three parameters, and the choice of a suitable threshold for the POT approach is analogous to the choice of a suitable block size for the BM approach.

3 Simulation design

This section will start with a rationale for a single DGP for BM and POT data. Subsequently, the GEV and GP will be linked in order to construct such a DGP. Next, the selection of simulation parameters for this research is covered. Finally, the validity of the proposed DGP is examined and tested.

3.1 Rationale for a single DGP for BM and POT data

For a proper comparison of the BM and POT approaches, a simulation study provides obvious benefits as the 'true' quantities of the DGP are known. Estimates of relevant quantities can then be inferred from the simulated data and can be compared to the known true quantities.

Generating simulated extreme data was feasible for El Adlouni et al. (2007) and Cannon (2010) as they focused exclusively on the BM approach, and could thus simulate extreme values directly from the GEV distribution with a predefined relation between covariates and the distribution parameters. Creating simulated data with known covariate-dependent parameters that are simultaneously suitable for BM and POT modelling, a prerequisite for this research, is complicated and has not yet been established in existing literature.

A simultaneously valid DGP is restricted by two conflicting constraints. On one hand, even though a connection between the GEV and GP distributions exists, the two sets of parameters are not completely interchangeable. The GEV distribution is characterised by three parameters that can be estimated (i.e., location, scale, and shape), whereas the GP distribution is characterised by two parameters (i.e., scale and shape) that can be estimated and one threshold that must be chosen before estimation. Thus, we cannot directly translate GP parameters into GEV parameters because it leaves the GEV location parameter undefined. Moving in the opposite direction from GEV to GP parameters is possible, with only the additional choice of a suitable threshold. On the other hand, the key motivator for wanting to use the POT approach is the ability to include more extreme observations than just the (annual) block maxima, e.g., the second or third largest values in a given block (year). Given that we require two sets of data, a smaller BM dataset and a larger POT dataset, we cannot start with creating maxima and supplement it with exceedances because the exceedances in each block (year) would be larger than the block maximum on plenty of occasions. Obviously, we also cannot model the BM and POT data separately, because a substantial number of observations should be present in both sets. That is to say, the highest exceedance in a block and the block maximum should often coincide. Therefore, we should start with the larger set of exceedances, and assign as maxima the highest exceedance within each block. To sum it up, a DGP whose data is simultaneously valid should therefore a) be defined in terms of GEV parameters, and b) take exceedances as its starting point.

In the next section, an algorithm for such a DGP is introduced whose output is simultaneously valid for covariate-dependent GEV and GP distributions and whose construction reasonably mirrors physical processes encountered in nature (e.g., in hydrology).

An alternative approach of simulating daily observations Z_i via some intermediate distribution such as the location-scale Student's t-distribution of the form $Z = \mu + \sigma * t(\nu)$ with $\nu = \frac{1}{\xi}$ degrees of freedom and where μ and σ depend on covariates, and subsequently taking BM and POT data from these daily observations does not seem to serve the purposes of this research. This is because even though the location-scale t-distribution is in the maximum domain of attraction of a GEV distribution G, the specific parameters of that distribution, and thereby the true return levels, are unknown. Accurately estimating these return levels via some simulation study will be computationally expensive, even more so since this research deals with non-stationary EVA distributions. This alternative approach however might be more realistic because now the BM and POT data do not directly follow the GEV and GP distributions, which is too good to be true in reality.

3.2 Linking the GEV and GP distributions to construct a DGP

To link the GEV and GP distributions in a way useful for the DGP, we first need to establish the distribution of the number of exceedances. Let $X_1, ..., X_n$ again be a series of iid random variables and suppose that the series is well-behaved in an extreme value sense, i.e., that Equation (1) holds. Define the probability ζ_u that a random point X_i exceeds u as

$$\zeta_u = \Pr(X_i > u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}, \tag{11}$$

where μ, σ and ξ denote annual GEV parameters; this follows from Equation (2). Due to the mutual independence of the points X_i, the number of exceedances n_u follows the binomial distribution

$$n_u \sim \mathrm{Bin}(n, \zeta_u). \tag{12}$$

This result can be exploited to simulate data that simultaneously follow the GEV- and GP distribution, and it only requires defining GEV parameters (μ, σ, ξ), a suitable threshold u and appropriate time periods. The exploitation follows from the following intuition. On one hand, given a set of daily observations $X_1, ..., X_{365}$ corresponding to one year of data, we know from Equation (12) that for a sufficient largely threshold u, the number of daily exceedances n_u can be expressed using a binomial distribution with probability ζ_u, where ζ_u is expressed in terms of GEV parameters and a threshold u. On the other hand, we know that the magnitude of the daily exceedances can be expressed using the GP distribution. Furthermore, because of their connection, the GEV parameters and the threshold uniquely define the associated GP parameters of the exceedances.

By definition, a subset of the data corresponding to one year has only one annual maximum. For suitable thresholds u, it is expected that a small, non-negative number of exceedances n_u are present in the data, whose distribution can be expressed using GP parameters. We shall distinguish three cases that differ with respect to the amount of information available to the BM and POT approaches. Firstly, if $n_u > 1$, all exceedances are extreme events applicable in the POT approach and its highest value is the annual (block) maximum. Secondly, if $n_u = 1$, that point is simultaneously applicable as the sole exceedance and as the annual maxima. Thirdly, if $n_u = 0$, no exceedances are present in the data and thus none are applicable for the POT approach; however a maximum should always be present in the data, even if it is not an extreme value. In that case, the maximum should follow a GEV distribution truncated from above at the threshold u.

This leads us to a two-step DGP for simulating data that simultaneously follows GEV and GP distributions. Given GEV parameters and a suitable threshold u for a given year, we can first sample the number of exceedances n_u from the binomial distribution of Equation (12). Then, we can sample n_u exceedances directly from the GP distribution. Conditional on a suitably high threshold u, the highest exceedance will be the BM, and it should closely follow the GEV distribution. If $n_u = 0$, the annual maximum is sampled directly from the truncated GEV distribution. This process can be replicated for many years to simulate series of BM and POT data, in which it is possible to specify non-stationarity through parametric relations in the annual GEV parameters. A step-by-step outline of the DGP that is simultaneously valid for BM and POT data can be seen in Algorithm 1 below.

The modelling of non-stationarity in discrete steps of one year, rather than modelling time continuously, is slightly restrictive. It is however a reasonable assumption, and one that is often already made by practitioners modelling natural phenomena. The single covariate of time is also a deterministic covariate. The algorithm could be adjusted to model random covariates, which is more realistic in certain settings but would also introduce an additional source of variance. Moreover, the current setup closely resembles the GEV-CDN study of Cannon (2010), which makes it attractive for comparison. For these two reasons, this simulation study only deals with a single, deterministic covariate. Including multiple, random covariates would be an interesting avenue for extending this research.

Algorithm 1: Generate BM and POT data simultaneously

```
input  : Polynomial terms for GEV parameters: $\{(m_0, m_1, m_2), (s_0, s_1), \xi)\}$
         Threshold difference: $\delta > 0$
         Number of years: $n_y$
         Number of days per year: $n \leftarrow 365$
output: Sets of indexed BM and POT data: $(Z_{BM}, Z_{POT})$
$Z_{BM}, Z_{POT} \leftarrow \emptyset$
for each year $j \in 1, \ldots, n_y$ do
    /* Set GEV and GP parameters for this year                              */
    $\mu_j \leftarrow m_0 + m_1 * j + m_2 * j^2$,
    $\sigma_j \leftarrow \exp(s_0 + s_1 * j)$
    $u_j \leftarrow \mu_j - \delta$
    $\tilde{\sigma_j} \leftarrow \sigma_j + \xi * (u_j - \mu_j)$
    /* Define daily exceedance rate and draw number of exceedances          */
    $\zeta_{u,j} \leftarrow f(\mu_j, \sigma_j, \xi_j, u_j, n)$ // Use Equation (11)
    $N_j \sim Bin(365, \zeta_{u,j})$
    /* Sample BM and POT data                                               */
    $Z^*_{BM}, Z^*_{POT} \leftarrow \emptyset$
    if $N_j > 0$ then
        $Z^*_{POT} \sim GP(N_j; \tilde{\sigma}, \xi)$ // Sample $N_j$ exceedances directly from GP
        $Z^*_{BM} \leftarrow \max(Z^*_{POT}) + u_j$ // Add highest exceedance to threshold for BM
        $Z^*_{POT} \leftarrow Z^*_{POT} + u_j$ // Add threshold in order to return data on same
            level
    else
        /* If no exceedance, draw BM directly from truncated GEV distribution
            */
        $Z^*_{BM} \leftarrow \emptyset$
        while $Z^*_{BM} = \emptyset$ do
            $\bar{Z}_{BM} \sim GEV(1; \mu, \sigma, \xi)$
            if $\bar{Z}_{BM} < u_j$ then
                $Z^*_{BM} \leftarrow \bar{Z}_{BM}$
            end
        end
    end
    /* Add data and its indices to output                                   */
    $Z_{BM} \leftarrow Z_{BM} \cup \{j, Z^*_{BM}\}$
    $Z_{POT} \leftarrow Z_{POT} \cup \{j, Z^*_{POT}\}$
end
return $(Z_{BM}, Z_{POT})$
```

3.3 Simulation parameters

Using Algorithm 1, we will simulate four types of datasets with distinct covariate relations. The first three simulations will span $n_y = 50$ years and we will replicate it 1000 times. The fourth simulation will span $n_y = 1000$ years and will be replicated 250 times. In order to compute t-tests, the simulations will be partitioned in 10 sets.

The first three simulations are the primary interest of this research. All its parameters, and in particular the shape parameter ξ and the number of years n_y, are set in such a way as to be representative for data types often encountered in environmental practice. Environmental data on extremes is often only available for several decades. The additional fourth simulation has a much longer duration of $n_y = 1000$ years. In certain other fields such as e.g., finance or insurance, data is available at a much higher frequency. This simulation type thus sheds additional light on the relative performance of the proposed POT approaches when BM data is not actually scarce. Note that we will continue defining blocks of simulation type 4 as years, but this choice is arbitrary and it can obviously be replaced by any unit specific to the field of interest.

The first simulation type will model datasets with only a linear increase in the location parameter. The second and fourth type will model datasets with an increase in the scale as well as a smaller increase in the location. The third type will model datasets with a quadratic change in the location parameter. All datasets will be modelled with a constant shape parameter $\xi = 0.1$. The threshold difference δ, which is a strictly positive, constant parameter in the DGP that defines the level of the threshold u_j relative to location μ_j. Appropriate values for δ ensure that the DGP threshold u_j is close to the location μ_j while allowing enough POT data points to be sampled. The thresholds in the fitting stage and δ are set up in such a way as to retain $2-4$ times as many data points relative to annual maxima after fitting; $\delta = 2$ suffices.

See Table 3 below for a complete overview of the simulation parameters. Examples of data for each of the three simulation types, together with their DGP threshold, $2, 5, 10$ and 100-year return levels, and fitted threshold according to POT-QR-1 can be found in Figures 1 to 4. Simulation type 4 also shows the 1000-year return level. Note that all fitted QR thresholds u above the DGP threshold are valid as long as the fitted QR threshold gradient roughly equals the DGP threshold gradient.

Table 3: Overview of simulation parameters

		Simulation type			
Parameter		1	2	3	4
Location μ_j	m_0	0	0	0	0
	m_1	0.1	0.04	0.1	0.002
	m_2	0	0	-0.003	0
Scale σ_j	s_0	0	0	0	0
	s_1	0	0.013	0	0.001
Shape	ξ	0.1	0.1	0.1	0.1
Threshold difference	δ	2	2	2	2
Number of years	n_y	50	50	50	1000

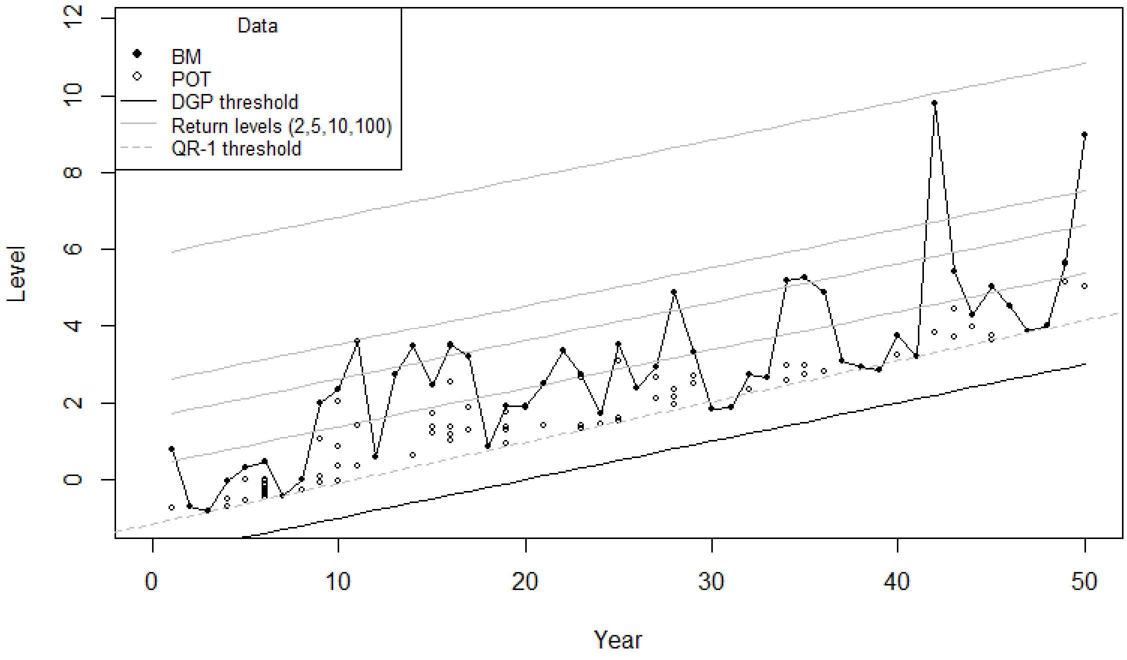


Figure 1: Example of simulation type 1

Note: Only shows POT data above the fitted QR-1 threshold. BM points above the QR-1 threshold overlap POT points.

Interpretation: Location of the points shifts steadily upwards while the scale remains constant. Gradient of QR-1 threshold roughly equal to gradient of DGP threshold.

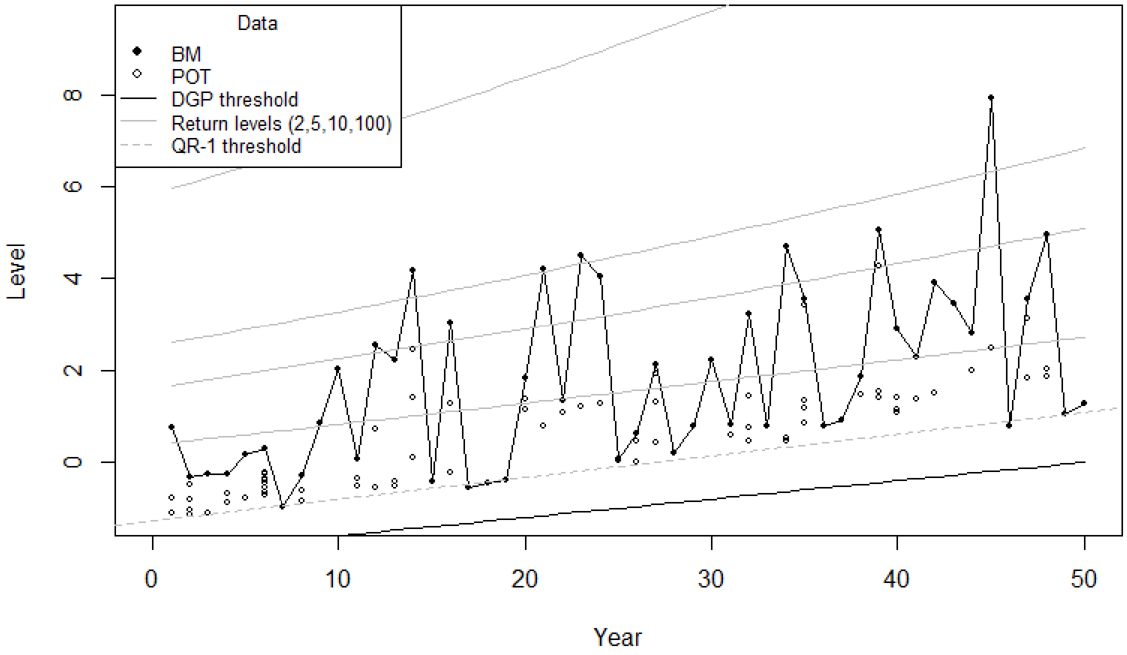


Figure 2: Example of simulation type 2

Interpretation: Location of points shifts slowly upwards. Scale grows as well, which also causes increasing return periods. Gradient of fitted QR-1 threshold again roughly equals gradient of DGP threshold.

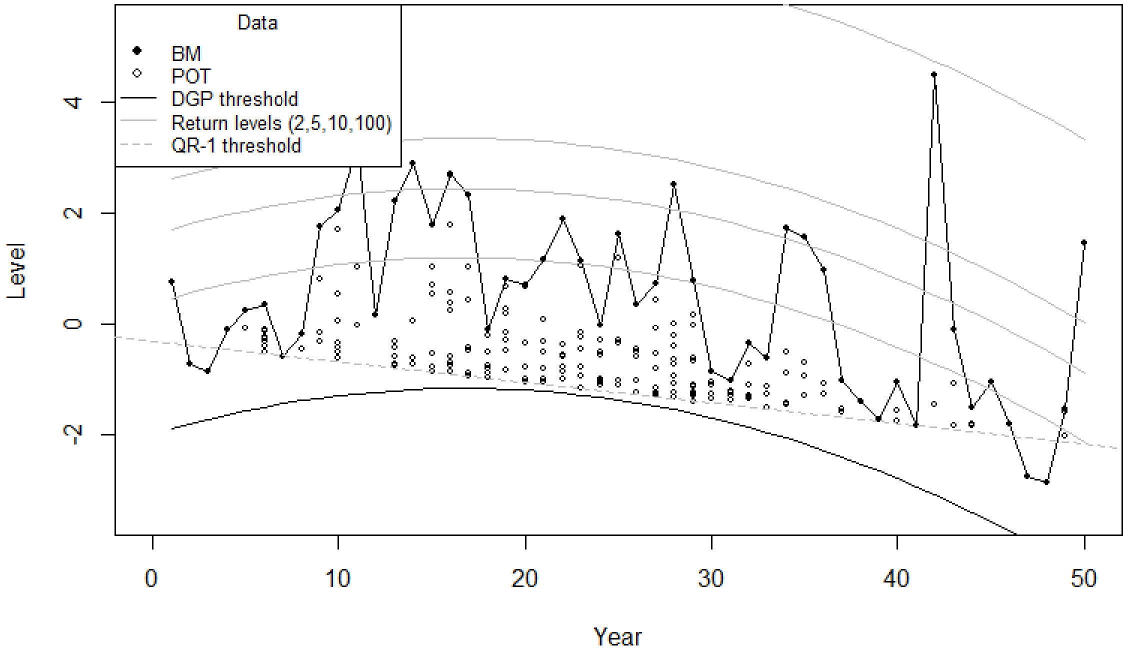


Figure 3: Example of simulation type 3

Interpretation: Location of GEV points curves downwards. Fitted QR-1 threshold is too high at the ends and too low in the centre, due to linearity of the quantile regression.

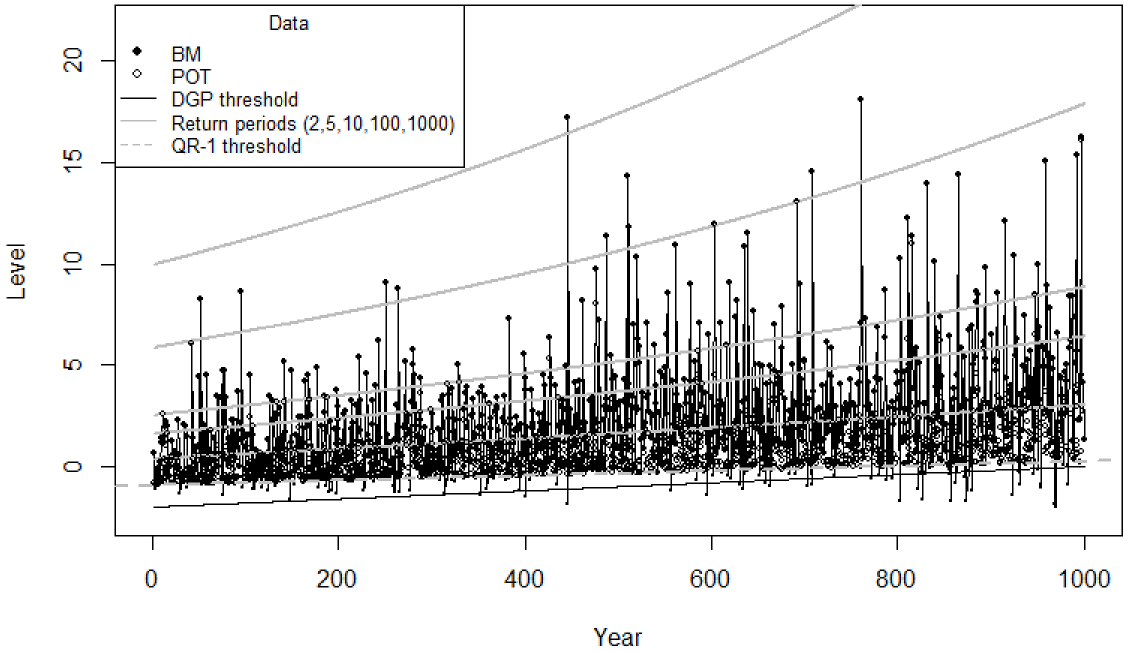


Figure 4: Example of simulation type 4

Interpretation: Similar to Figure 2, but with 1000 years

3.4 Validity of the DGP

Naturally, the validity of the DGP is conditional on the validity of the assumptions of a reasonably high threshold u and the closely related assumption of $n >> n_u$; which is manageable if we set the simulation parameter δ appropriately. As can be seen in Figure 5, true return levels for the GEV and GP distributions of simulation type 1 are very close and approach each other with increasing return periods. Similar results are seen for the other simulation types. The parameter $\delta = 2$ is thus appropriate, especially since we are mainly interested in high return periods.

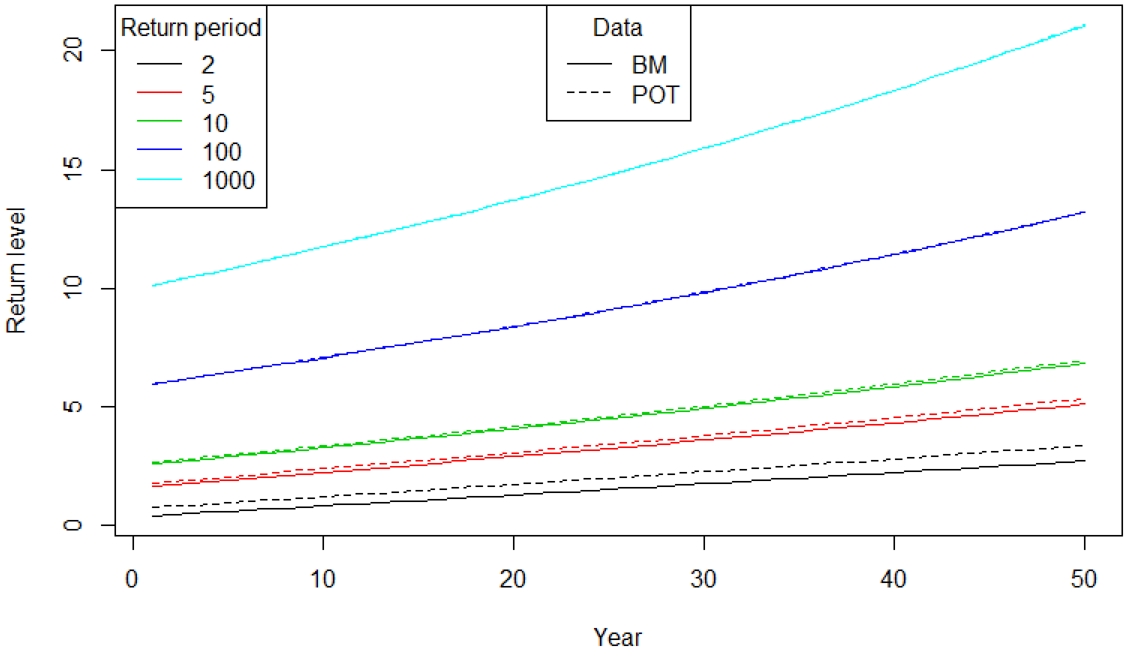


Figure 5: Comparison of true return levels for the BM and POT data for simulation type 1

Interpretation: Return levels for the BM and POT data are approximately equal and approach each other with increasing return periods.

A second concern for the validity of the DGP relates to the goodness of fit of the BM data to the GEV distribution. By construction of the DGP, the POT data will follow the GP distribution exactly, whereas the BM data will only approximately follow the GEV distribution. The approximation is assumed to be very close. However, if the approximation is not very close, it will create an unfair advantage for the POT approach in the comparison with the BM approach. We will therefore check this assumption on generated data using quantile-quantile (Q-Q) plots and the Kolmogorov-Smirnov test for goodness of fit (Massey Jr, 1951). For completeness, we also apply both checks to the GP data, even though we know that the exceedances are drawn directly from the specified GP distribution by construction.

The data is generated using Algorithm 1 with $n_y = 1000, \mu = 0, \sigma = 1, \xi = 0.1$ and $\delta = 2$. These parameters are comparable to simulation type 1 of Table 3, but without the non-stationarity and with an increased number of years n_y. These alterations are made for clarity purposes, and do not negatively impact the approximation to the GEV distribution. In fact, the increased number of years increases the power of the test.

A Q-Q plot for one randomly generated dataset is shown in Figure 6 below. Both plots closely follow the diagonal line and there is no reason to doubt that the simulated empirical GEV and GP data follow their true model distributions.

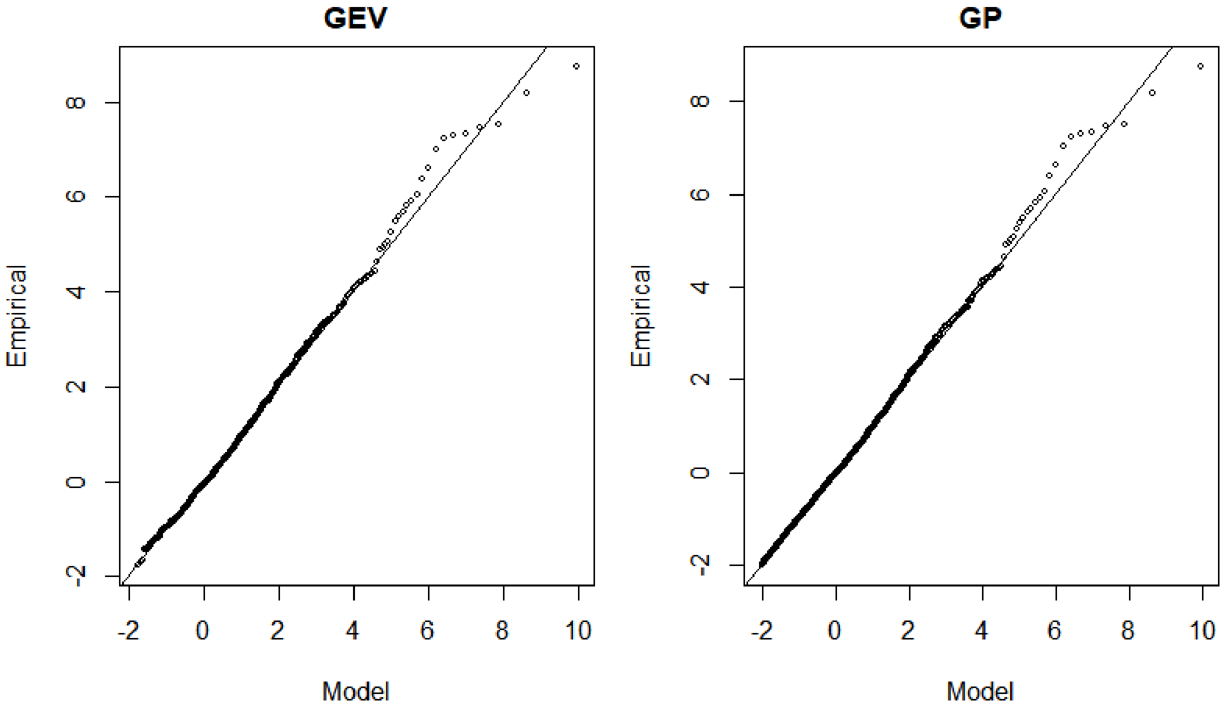


Figure 6: Q-Q plots

Note: The GEV data contains 1000 maxima, whereas the GP data contains several times more exceedances. *Interpretation:* Both sets of data generally follow the diagonal, with only some small divergence at the upper end, which is not uncommon.

We can also formally check the approximation using the Kolmogorov-Smirnov test with the null hypothesis that the distribution function which generated the data is the specified model distribution. Data generation and testing are replicated 1000 times at a significance level of 5%, with the same parameters as for the Q-Q plots above. In advance, we expect that the null hypothesis is rejected in roughly 5% of the replications. After 1000 replications, the null was rejected 4.8% and 5.6% of the times for the GEV and GP distributions, respectively. This is in line with our prior expectation.

Both the Q-Q plot and the Kolmogorov-Smirnov test show that there is no reason to doubt our assumption that the approximation of the BM data to the GEV distribution is very close. Therefore, there is no reason to assume that the DGP significantly favours the POT approach.

4 Simulation results

The RMSEs and their standard deviations (sds) of the $2, 5, 10, 100$ and 1000-year return levels for all simulation types can be found in Tables 4 to 7. These return levels match those examined in Cannon (2010). This research hypothesised that the novel POT-C and POT-QR approaches would outperform the existing BM approach due to the inclusion of more data. In general, this is not the case. However, under certain conditions and especially for high return periods, we do see that POT outperforms BM.

For simulation type 1 in Table 4, with a linear increase in location, we see in line with our expectation that the hyperparameters with an identity activation function outperform their $\tanh(\cdot)$ counterparts. For the BM model, we see that hyperparameter set 3 shows the best overall performance. This is as expected, because the location parameter μ is the only non-stationary parameter in both the DGP and the fitted model, whereas hyperparameter set 1 also assumes a non-stationary relation for the scale σ. We also see that the BM model performs best among all models for the four lowest return periods; the difference is significant in all four cases. For the highest return period of 1000 years however, the POT-QR-1 model overtakes the BM model as the best performing model, although this difference is not significant at a 5% level. In any case, the differences between the BM and POT-QR models are modest. The somewhat disappointing performance of the POT-C models is likely a result of the fact that these models do not have ability to model non-stationarity in the location, which is the only non-stationary feature in the DGP of this simulation type. To a lesser degree, the same explanation holds for the POT-QR models, whose only ability to model non-stationary in location is via the threshold. On the other hand, the BM model can directly model non-stationarity and was thus in a favourable position to begin with. The considerably better performance of POT-QR relative to POT-C show the strong positive effect of the QR threshold. Surprisingly enough, the GML cost does not provide a good metric for determining the best hyperparameter set within an EVA framework. The lowest GML values favour the models with a $\tanh(\cdot)$ activation function, whereas these models have inflated return level errors and perform worse in terms of RMSE.

Table 4: Overview of results for simulation type 1

	BM				POT-QR-1		POT-QR-2		POT-C-0		POT-C–1	
Statistic	HP1	HP2	HP3	HP4	HP1	HP2	HP1	HP2	HP1	HP2	HP1	HP2
RMSE and (sd) of return levels												
2-year	0.23	0.46	0.21**	0.46	0.28	0.61	0.28	0.64	1.10	1.39	1.74	1.55
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.25)	(0.03)	(0.08)	(0.05)	(0.03)
5-year	0.38	0.73	0.29**	0.53	0.41	0.97	0.42	1.09	1.22	1.54	1.99	1.68
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.45)	(0.05)	(0.11)	(0.09)	(0.05)
10-year	0.54	1.02	0.39**	0.60	0.55	1.27	0.56	1.47	1.28	1.63	2.05	1.69
	(0.02)	(0.05)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.60)	(0.07)	(0.13)	(0.12)	(0.06)
100-year	1.59	2.80	1.29**	1.58	1.47	2.73	1.52	3.17	2.33	2.75	2.72	2.65
	(0.08)	(0.18)	(0.09)	(0.13)	(0.08)	(0.12)	(0.09)	(1.02)	(0.07)	(0.14)	(0.14)	(0.09)
1000-year	3.96	7.02	3.55	5.29	3.44	5.13	3.58	5.87	5.00	5.52	5.07	5.53
	(0.29)	(0.56)	(0.30)	(0.51)	(0.25)	(0.32)	(0.29)	(1.35)	(0.09)	(0.11)	(0.10)	(0.14)
GML cost	-32.9	-39.49	-32.36	-36.84	113.19	108.96	81.23	76.93	569.39	563.26	765.22	757.61
Obs. per year	1				2.28		1.59		7.12		8.61	

Note: Light grey shading indicates lowest RMSE within framework, dark grey indicates lowest overall RMSE. Asterisks * and ** indicate whether the difference between the best BM model and the best POT model is significant at 5% and 1% levels, respectively.
Interpretation: RMSE of BM model is slightly but significantly lower than POT-QR models for return periods of 2-100 years. RMSE of POT-QR-1 model is lowest for 1000-year return period, but not significant. Differences between BM and POT-QR models are modest, POT-C models are considerably worse.

For simulation type 2 in Table 5, we see a more diffuse picture. For all EVA frameworks, we again find that the identity activation function outperforms the $\tanh(\cdot)$ function. Also, for the BM model, we again find that the best performing hyperparameter set is the one that matches the DGP, i.e. with non-stationarity in both location μ and scale σ. With the scale now being the main non-stationary feature of this simulation type, we see that the POT approaches have a better chance at competing with the BM approach. The BM model still performs best for return periods of $2, 5$ and 10 years, with the differences for 2 and 5 year return periods being significant at the 1% level. POT-QR-1 and POT-C–1 claim the top spot for the return periods of 100 and 1000 years, respectively, and differences are significant at the 1% and 5% level. These advantages are achieved with only $2-4$ times as much data available relative to annual maxima, as indicated by the data multiplication factor in the bottom row. In any case, differences across the frameworks again are modest.

Table 5: Overview of results for simulation type 2

	BM				POT-QR-1		POT-QR-2		POT-C-0		POT-C–1	
Statistic	HP1	HP2	HP3	HP4	HP1	HP2	HP1	HP2	HP1	HP2	HP1	HP2
RMSE and (sd) of return levels												
2-year	0.32**	0.66	0.37	0.75	0.41	0.71	0.41	0.83	0.53	1.03	0.51	0.80
	(0.02)	(0.04)	(0.01)	(0.04)	(0.01)	(0.09)	(0.02)	(0.18)	(0.03)	(0.15)	(0.03)	(0.04)
5-year	0.54**	1.07	0.70	0.98	0.59	1.12	0.61	1.39	0.69	1.59	0.78	1.16
	(0.04)	(0.06)	(0.02)	(0.03)	(0.02)	(0.16)	(0.04)	(0.32)	(0.05)	(0.24)	(0.05)	(0.05)
10-year	0.76	1.51	1.00	1.20	0.80	1.48	0.82	1.87	0.89	2.08	1.01	1.45
	(0.05)	(0.09)	(0.03)	(0.04)	(0.04)	(0.21)	(0.05)	(0.41)	(0.07)	(0.31)	(0.06)	(0.05)
100-year	2.25	4.13	2.72	3.05	2.11*	3.29	2.20	4.06	2.20	4.33	2.23	2.95
	(0.12)	(0.23)	(0.09)	(0.14)	(0.13)	(0.40)	(0.12)	(0.69)	(0.16)	(0.49)	(0.07)	(0.09)
1000-year	5.62	10.26	6.52	9.22	4.88	6.40	5.19	7.66	5.01	7.95	4.53**	5.64
	(0.35)	(0.68)	(0.26)	(0.75)	(0.31)	(0.66)	(0.34)	(0.96)	(0.40)	(0.61)	(0.22)	(0.22)
GML cost	-22.45	-29.63	-20.79	-26.02	151.55	147.54	107.96	103.42	145.72	143.03	269.58	266.38
Obs. per year	1				2.29		1.6		2.09		3.82	

Interpretation: RMSE of BM model is slightly lower than the other frameworks for return periods of 2-10 years. RMSE of POT-QR-1 is lowest for 100-year return period and RMSE of POT-C–1 is lowest for 1000-year return period; both are significant. Differences between all frameworks are modest.

For simulation type 3 in Table 6, we see that the standard BM approach bests the POT approaches across the board. The differences are significant at the 1% level for all return periods. Clearly, the curve in the location is too much for the constant or linear thresholds too handle. For the BM model, the $\tanh(\cdot)$ activation function of HP4 beats the identity function of HP3 for return periods of $2, 5$ and 10 years, whereas this is reversed for periods of 100 and 1000 years. Differences across the frameworks are moderate.

Table 6: Overview of results for simulation type 3

	BM				POT-QR-1		POT-QR-2		POT-C-0		POT-C–1	
Statistic	HP1	HP2	HP3	HP4	HP1	HP2	HP1	HP2	HP1	HP2	HP1	HP2
RMSE and (sd) of return levels												
2-year	0.69	0.46	0.61	0.45**	0.68	1.29	0.69	1.18	1.08	1.55	1.11	1.36
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.21)	(0.01)	(0.22)	(0.02)	(0.25)	(0.02)	(0.08)
5-year	0.89	0.72	0.66	0.51**	0.76	1.73	0.77	1.67	1.17	2.14	1.19	1.58
	(0.03)	(0.03)	(0.01)	(0.01)	(0.02)	(0.31)	(0.02)	(0.36)	(0.03)	(0.47)	(0.02)	(0.12)
10-year	1.06	1.01	0.73	0.58**	0.85	2.12	0.88	2.11	1.27	2.66	1.28	1.79
	(0.05)	(0.04)	(0.01)	(0.02)	(0.03)	(0.40)	(0.02)	(0.46)	(0.04)	(0.62)	(0.03)	(0.15)
100-year	2.09	2.75	1.37**	1.57	1.61	3.91	1.70	4.10	2.11	4.90	1.95	2.92
	(0.17)	(0.13)	(0.08)	(0.11)	(0.08)	(0.73)	(0.07)	(0.87)	(0.09)	(1.07)	(0.06)	(0.28)
1000-year	4.26	6.91	3.03*	5.27	3.29	6.71	3.59	7.22	4.16	8.32	3.52	4.91
	(0.42)	(0.56)	(0.21)	(0.40)	(0.20)	(1.17)	(0.22)	(1.40)	(0.24)	(1.46)	(0.15)	(0.45)
GML cost	-22.07	-35.34	-21.2	-32.63	167.39	162.66	103.08	98.83	69.92	66.94	192.86	189.76
Obs. per year	1				3.53		2.09		1.37		4.15	

Interpretation: RMSE of model 1 is slightly lower than the other frameworks for all return periods, with HP4 better for 2 to 10-year return periods and HP3 better for 100 to 1000-year return periods. The POT-QR models are slightly better than the POT-C models. Differences between all frameworks are moderate and significant.

The results for simulation type 4 can be found in Table 7. Given that 1000 years of observations are now available, the return periods of less than 100 years are not very rare or extreme. Alternative inference methods might perform better for those return periods. Rather, the 100 and 1000-year return periods are of more interest. Nevertheless, we have included the $2, 5$ and 10-year return periods for consistency with the other simulations. The results are mostly similar to the results of simulation type 2 in Table 5. BM performs best for return periods of 2-10 years, while POT-QR-1 performs best for return periods of 100 years and more. The differences are only significant for the 2 and 5-year return periods. Naturally, RMSEs for all models are much lower than simulation type 2 due to the inclusion of more data. More surprisingly, we find that the POT-QR-1 model bests the BM approach for high return periods of 100 and 1000 years, even though BM data is not scarce. Differences between the BM and POT-QR models are modest, while POT-C performs worse.

Table 7: Overview of results for simulation type 4

	BM				POT-QR-1		POT-QR-2		POT-C-0		POT-C–1	
Statistic	HP1	HP2	HP3	HP4	HP1	HP2	HP1	HP2	HP1	HP2	HP1	HP2
RMSE and (sd) of return levels												
2-year	0.09**	0.13	0.48	0.51	0.13	0.16	0.11	0.14	0.36	0.41	0.60	0.67
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.16)
5-year	0.15**	0.22	1.10	1.11	0.18	0.25	0.17	0.23	0.40	0.49	0.86	1.00
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.03)	(0.01)	(0.01)	(0.02)	(0.04)	(0.03)	(0.23)
10-year	0.22	0.30	1.55	1.57	0.24	0.35	0.24	0.32	0.45	0.57	1.02	1.24
	(0.03)	(0.03)	(0.01)	(0.01)	(0.02)	(0.04)	(0.02)	(0.02)	(0.03)	(0.05)	(0.05)	(0.29)
100-year	0.68	0.79	3.29	3.33	0.66	0.95	0.70	0.92	0.82	1.13	1.52	2.20
	(0.12)	(0.09)	(0.05)	(0.06)	(0.09)	(0.13)	(0.09)	(0.12)	(0.09)	(0.14)	(0.07)	(0.53)
1000-year	1.63	1.73	5.74	5.84	1.56	2.17	1.68	2.20	1.71	2.35	2.73	4.11
	(0.35)	(0.23)	(0.19)	(0.21)	(0.23)	(0.36)	(0.26)	(0.35)	(0.24)	(0.31)	(0.14)	(0.85)
GML cost	-1125	-1128	-1066	-1069	3460	3462	2492	2493	2819	2821	4749	4758
Obs. per year	1				2.28		1.6		1.76		3.08	

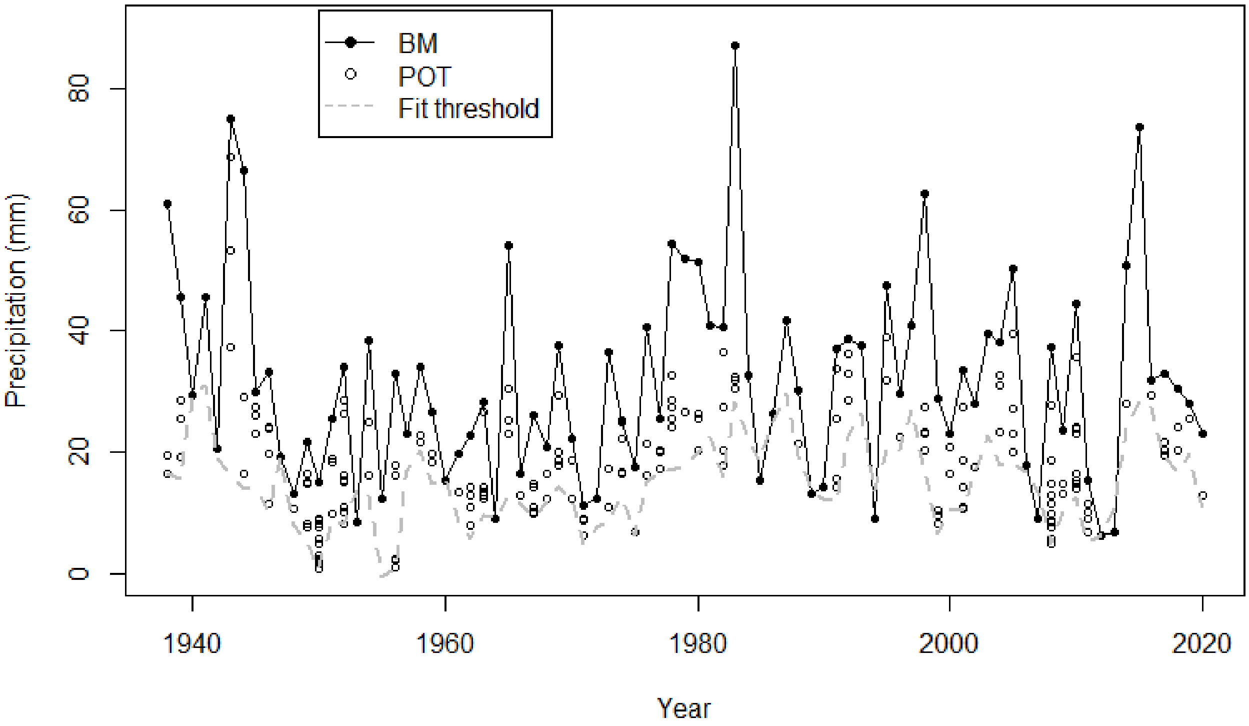
Interpretation: RMSE of BM model is slightly lower than the other frameworks for return periods of 2-10 years. RMSE of POT-QR-1 is lowest for 100 and 1000-year return periods. Differences between BM and POT-QR frameworks are modest, POT-C performs worse.

In all, we have seen that the POT approaches perform nearly on par with the BM approach for low return periods, and oftentimes best the BM approach for high return periods. This is most pronounced for datasets in which the scale parameter σ is the primary non-stationary feature. These results are persistent even when BM data is not scarce.

5 Application to precipitation data

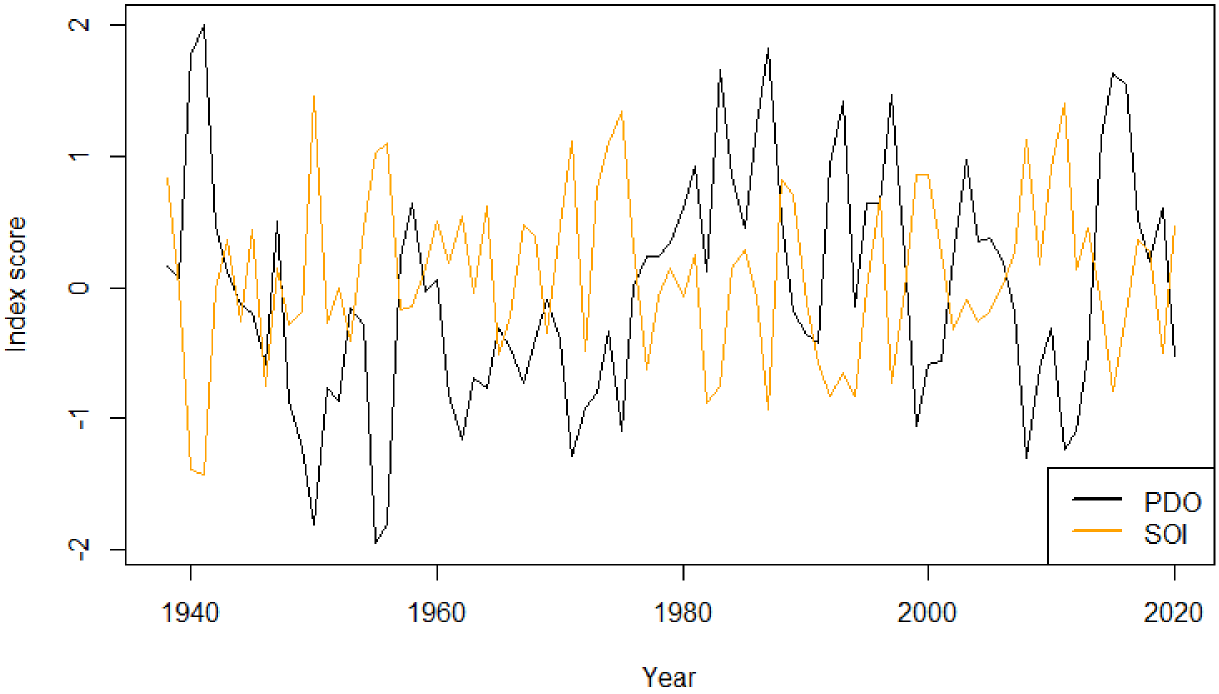
To showcase the applicability of the POT approach, and its QR variant, we will apply the same approaches of Section 2 to the Randsburg, USA, precipitation data previously analysed in El Adlouni et al. (2007) and Cannon (2010). Using his GEV-CDN approach, Cannon (2010) identified a non-linear relation between parameters of the GEV distribution and two atmospheric indicators or 'teleconnections', namely the Pacific Decadal Oscillation (PDO) and the Southern Oscillation Index (SOI).

The data of precipitation extremes and the teleconnections can be seen below in Figure 7. The correlation coefficients of the covariates with the annual extremes are $\rho_{PDO} = 0.49$ and $\rho_{SOI} = -0.29$. This data is extracted from daily data retrieved directly from the National Centers for Environmental Information of the American National Oceanic and Atmospheric Administration. Note that the data differs from the data visualised in Figure 4 of Cannon (2010). After correspondence regarding this surprising disparity, Cannon has indicated that this is due to a data handling error on his end; the data as shown here in Figure 7 can be considered correct.



(a) Annual maxima and exceedances of daily precipitation, 1938-2020

Note: Fit threshold as per POT-QR-1
Interpretation: Fitted quantile regression threshold appears to move in lockstep with the location of extremes, and thus seems to work reasonably well.



(b) PDO and SOI, 1938 - 2020

Interpretation: PDO and SOI generally move in opposite directions: $\rho = -0.63$

Figure 7: Precipitation data and teleconnections for Randsburg (USA)

The same EVA frameworks of Table 1 and hyperparameter option sets of Table 2 will be investigated. The only change being that the constant thresholds u of the POT-C models will be shifted to 12 and 20 respectively, so that the POT data of the simulation and the precipitation application are roughly equal in terms of exceedances per year and that they are thus high enough.

We now deal with two, random covariates, as opposed to the single, deterministic covariate in the simulations. In theory, this should present no particular concern for all methods applied in this research, although visualisation of results is slightly more complicated. In this bivariate setting, contour plots provide an obvious and attractive way to shed light on estimated quantities and their dependence on covariates. As practitioners in fields such as e.g. hydraulic engineering are often interested in the magnitude of natural phenomena occurring at specific, high return periods, Figures 9 to 13 show contour plots for the estimated 100-year return level for each of the aforementioned approaches. AICc, BIC and GML statistics of all approaches are shown in Table 8, together with the average number of exceedances per year.

Because of the limited amount of data available and because the scatter of the two covariates does not provide full coverage of the domain, several steps must be taken to reduce the risk of overfitting and present meaningful results. The hyperparameter options with a $\tanh(\cdot)$ activation function are particularly vulnerable to overfitting. First, optimisation for those hyperparameter sets is tried 100 times from a random starting point to avoid shallow local optima. Secondly, the contour plots are restricted to points within the 97.5% tolerance ellipsoid of the data; there is not sufficient support for points outside this boundary where behaviour will be erratic. Thirdly, the contour plots are restricted to return levels between 20 and 180mm. This range contains nearly all observations. The maximum precipitation within the 83 years of available data is less than 90mm, and there are more than 150 exceedances over 20mm. It is far more likely that 100-year return levels above 180mm and below 20mm are due to overfitting. Additionally, restricting the contour plots to this range allows us to see more contrast for all return levels within the plausible range. This restriction creates some surprising empty (white) regions in the contour plots with a $\tanh(\cdot)$ activation function of Figures 10b to 13b. A few observations exist in the empty regions for the POT-QR models Figures 10b to 11b, of which the one in the large empty region in Figure 11b is most clearly visible. These few observations all have fitted return levels below the minimum of 20mm, which are caused by very low fitted QR thresholds and exacerbated by the flexibility of the $\tanh(\cdot)$ activation function. We thus see that in these instances the POT-QR model has indeed overfit to handle these low observations, and we should not trust or try to interpret the fitted return levels in those empty regions.

As noted before in Section 2, a direct comparison of the test statistics between the EVA frameworks is not possible. For return periods of 100 years, the simulation results of Section 4 however indicated a small advantage for the POT approaches with regression quantile thresholds over the BM approach in settings where both the scale and the location of extremes change linearly over the covariates. Such a setting can be identified visually; plots of the annual maxima vs. PDO and vs. SOI can be seen in Figure 8. In relation to the PDO, a trend in both location and scale seems to be apparent, although the latter trend is somewhat obscure. For return periods of 1000 years, the simulation results of Section 4 indicated an advantage for the POT-QR approach, as long as there is no strong quadratic trend in the location. All in all, this suggest that we can have a slight preference for POT-QR estimations over BM estimations.

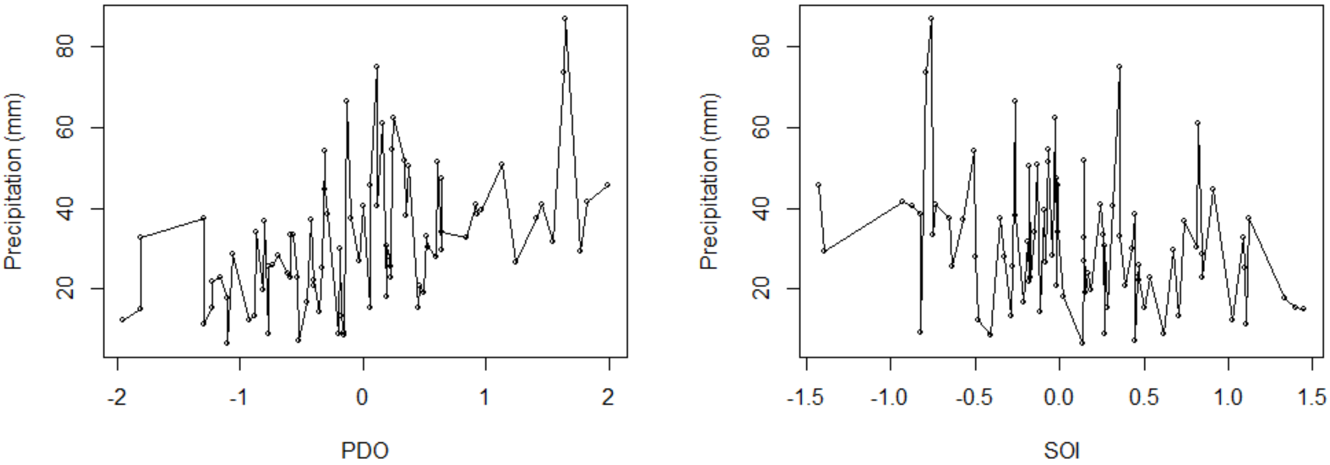


Figure 8: Annual maxima of Randsburg precipitation vs. PDO and vs. SOI

Upon analysis of the results of Table 8, we see somewhat surprisingly that the BIC identifies a different winner than the AICc and GML for three of the five frameworks. Nonetheless, the statistics offer considerable support for the non-linear relations modelled by the tanh($\cdot$) activation function.

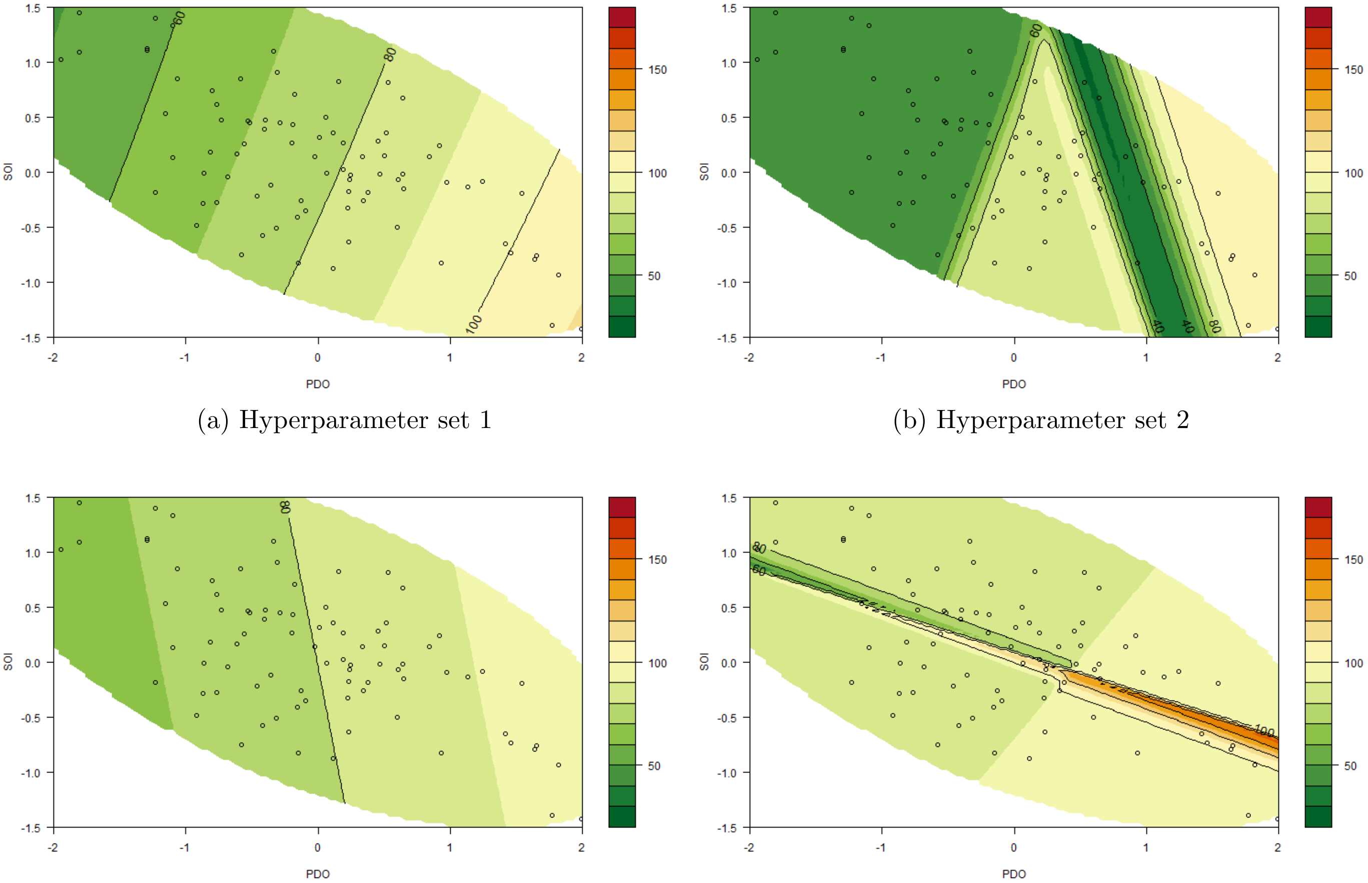
Aware of the limited amount of data, even for the POT approaches, we should not draw too strong conclusions. Three features of the contour plots, however, are noteworthy. First off, it appears that the optimisations with the tanh($\cdot$) activation function are trying to map the domain into a limited number of regions with rather straight boundaries and fairly constant return levels inside each region, rather than more curvy contours often seen in such contour plots. Secondly, the POT approaches suggest somewhat lower 100-year return levels than the BM approaches, with the bulk of the return levels between $70 - 90$mm for BM and between $40 - 70$mm for the POT approaches. Thirdly, and perhaps most importantly, we see clear support that higher values for the PDO lead to higher extremes, especially when this coincides with lower values for the SOI.

All in all, we have shown the applicability of the proposed POT approach and its QR variant. This has led to an increased understanding of the relation between precipitation extremes in Randsburg and the PDO and SOI teleconnections, and has suggested that 100-year return levels are possibly a few centimetres lower than previously estimated.

Table 8: Overview of results for Randsburg precipitation data

	BM				POT-QR-1		POT-QR-2		POT-C-0		POT-C–1	
Statistic	HP1	HP2	HP3	HP4	HP1	HP2	HP1	HP2	HP1	HP2	HP1	HP2
AICc	-50	-67	-53	-65	1775	1706	986	903	2175	2158	1040	1031
BIC	-34	-38	-42	-46	1789	1745	997	934	2190	2198	1052	1063
GML	-33	-52	-32	-43	883	841	488	440	1083	1067	516	504
Obs. per year	1				3.24		1.76		3.87		1.82	

Note: Light grey shading indicates lowest statistic for each framework



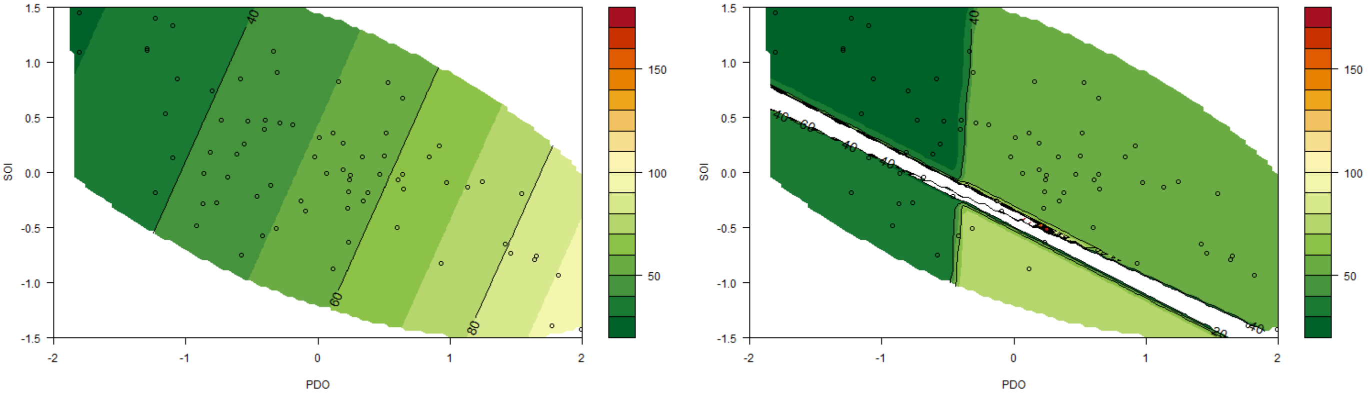
(a) Hyperparameter set 1

(b) Hyperparameter set 2

(c) Hyperparameter set 3

(d) Hyperparameter set 4

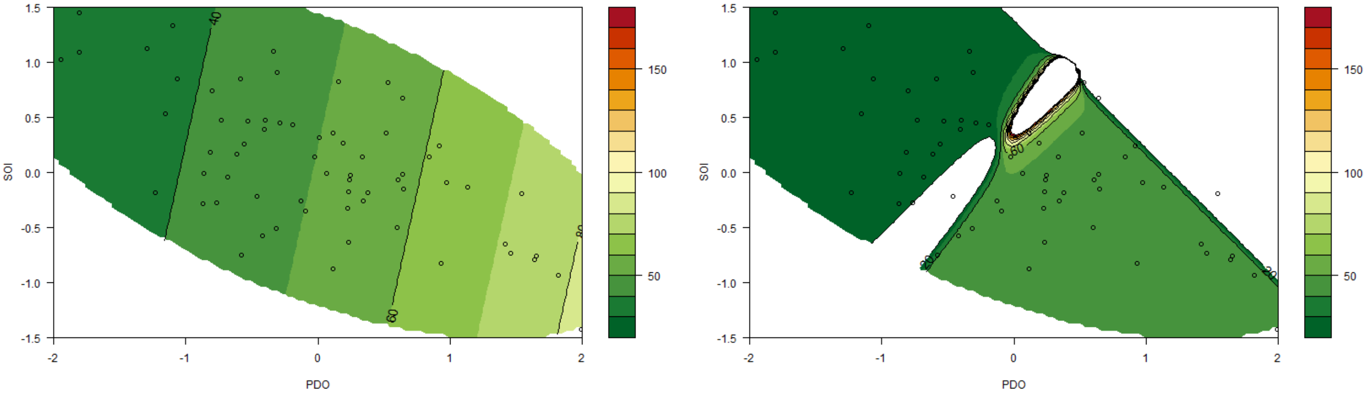
Figure 9: Contour plots for 100-year return levels for the BM model



(a) Hyperparameter set 1

(b) Hyperparameter set 2

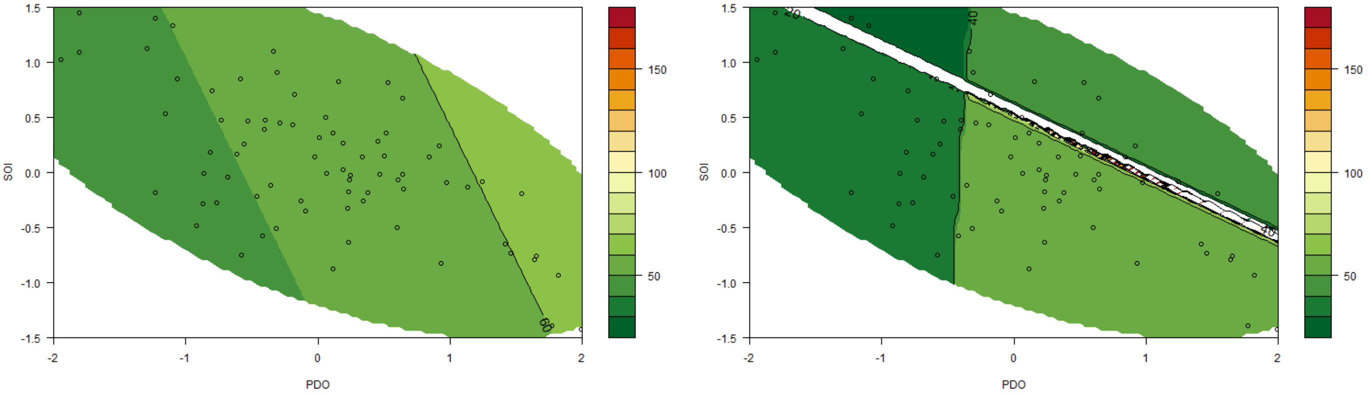
Figure 10: Contour plots for the POT-QR-1 model



(a) Hyperparameter set 1

(b) Hyperparameter set 2

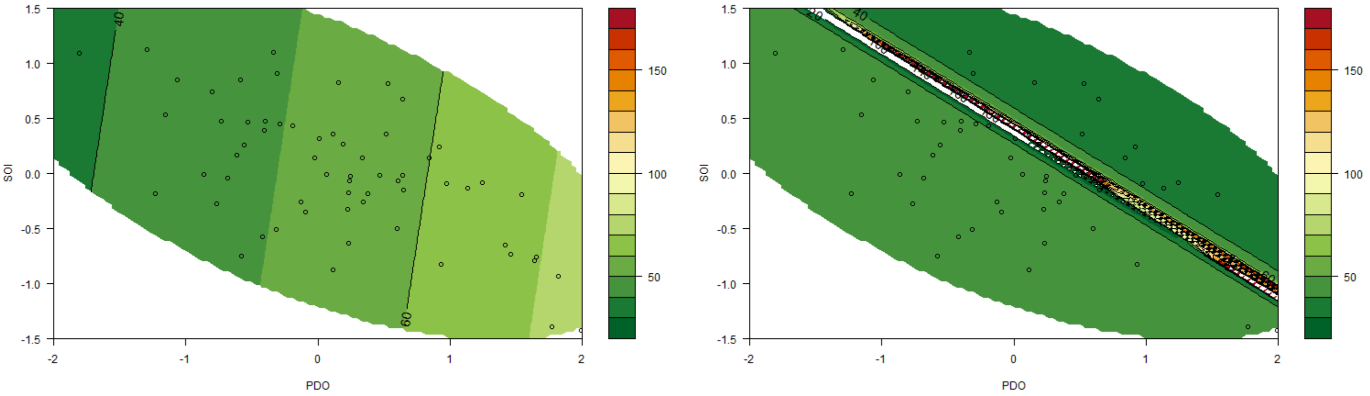
Figure 11: Contour plots for the POT-QR-2 model



(a) Hyperparameter set 1

(b) Hyperparameter set 2

Figure 12: Contour plots for the POT-C-0 model



(a) Hyperparameter set 1

(b) Hyperparameter set 2

Figure 13: Contour plots for the POT-C–1 model

6 Conclusion

This paper introduced an alternative technique for non-stationary EVA using CDNs based on the POT model: GP-CDN. Encouraged by the greater efficiency of threshold exceedances, this paper hypothesised that the POT-C and POT-QR approaches modelled using the GP-CDN yield favourable results.

A Monte Carlo simulation study was conducted to directly compare the performance of the new GP-CDN technique and the existing GEV-CDN in terms of the RMSE of estimated return levels. A new algorithm was introduced to generate data that is simultaneously valid for GEV and GP distributions, an important prerequisite for a fair comparison of the techniques. The Monte Carlo simulation study found that GP-CDN, and particularly its variant with linear thresholds based on QR, nears the performance of GEV-CDN for lower return periods of several decades across a range of data types. For return periods of a century and more, GP-CDN oftentimes modestly outperforms GEV-CDN. The benefits of GP-CDN are more pronounced when the scale of the data is the primary non-stationary feature. The improved performance of the POT-QR approach relative to the POT-C approach is likely due to the fact that the linear QR threshold gives the POT-QR approach an avenue through which non-stationary in the location can be handled. These results are persistent even when BM data is not scarce.

Application of the GP-CDN technique to precipitation data showcased its applicability to real-world data settings. It led to an increased understanding of the non-linear relation between covariates and extremes, and to an adjustment of the estimated 100-year return levels.

Policymakers requiring information on extreme events are primarily concerned with high return levels, as these events are hardest to estimate and have an outsized effect on society. This study has found that GP-CDN provides them with a viable alternative for the estimation of non-stationary extremes.

References

Bücher, A. and Zhou, C. (2018). A horse racing between the block maxima method and the peak-over-threshold approach. ArXiv e-prints, arXiv:1807.00282.

Caires, S. (2009). A comparative simulation study of the annual maxima and the peaks-over-threshold methods. *Deltares report 1200264-002 for Rijkswaterstaat, Waterdienst.*

Cannon, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes: An International Journal*, 24(6):673–685.

Cannon, A. J. (2011). GEVcdn: an R package for nonstationary extreme value analysis by generalized extreme value conditional density estimation network. *Computers & Geosciences*, 37(9):1532–1533.

Cannon, A. J. (2012). Neural networks for probabilistic environmental prediction: Conditional density estimation network creation and evaluation (CaDENCE) in R. *Computers & Geosciences*, 41:126–135.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):207–222.

Coles, S. (2001). *An Introduction to statistical modeling of extreme values*, volume 208. Springer.

Coles, S. G. and Dixon, M. J. (1999). Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23.

Davison, A. C. and Huser, R. (2015). Statistics of extremes. *Annual Review of Statistics and its Application*, 2:203–235.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):393–425.

El Adlouni, S., Ouarda, T. B., Zhang, X., Roy, R., and Bobée, B. (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research*, 43(3).

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1996). *Modelling extremal events: for insurance and finance.* Springer-Verlag.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, pages 1163–1174.

Hosking, J. R. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):105–124.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261.

Jain, S. and Lall, U. (2001). Floods in a changing climate: Does the past represent the future? *Water Resources Research*, 37(12):3193–3205.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171.

Kharin, V. V. and Zwiers, F. W. (2005). Estimating extremes in transient climate change simulations. *Journal of Climate*, 18(8):1156–1173.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics*, 9:155–176.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, pages 33–50.

Kyselỳ, J., Picek, J., and Beranová, R. (2010). Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold. *Global and Planetary Change*, 72(1-2):55–68.

Martins, E. S. and Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3):737–744.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Shrestha, R. R., Cannon, A. J., Schnorbus, M. A., and Zwiers, F. W. (2017). Projecting future nonstationary extreme streamflow for the Fraser River, Canada. *Climatic Change*, 145(3):289–303.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

Vasiliades, L., Galiatsatou, P., and Loukas, A. (2015). Nonstationary frequency analysis of annual maximum rainfall using climate covariates. *Water Resources Management*, 29(2):339–358.

Zhang, X., Zwiers, F. W., and Li, G. (2004). Monte Carlo experiments on the detection of trends in extreme values. *Journal of Climate*, 17(10):1945–1952.