GEV-Extremal random forest

Lucien M. Vidagbandji[1] & Alexandre Berred [2] & Cyrille Bertelle [3] & Laurent Amanton [4]

[1] *Univ le Havre Normandie, LMAH, France, mahutin-lucien.vidagbandji@univ-lehavre.fr*
[2] *Univ le Havre Normandie, LMAH, France, alexandre.berred@univ-lehavre.fr*
[3] *Univ le Havre Normandie, LITIS, France, cyrille.bertelle@univ-lehavre.fr*
[4] *Univ le Havre Normandie, LITIS, France, laurent.amanton@univ-lehavre.fr*

Résumé. La régression quantile est une méthode statistique couramment utilisée dans l'analyse de régression. Contrairement à la régression classique, qui se concentre sur la prédiction de la moyenne conditionnelle d'une variable dépendante en fonction des variables indépendantes, la régression quantile vise plutôt à prédire les quantiles conditionnels. Les méthodes classiques de régression quantile rencontrent des défis, particulièrement lorsque le quantile d'intérêt est extrême, en raison du nombre limité de données disponibles dans la queue de la distribution, ou lorsque la fonction quantile est complexe. Dans le cadre de cette étude, nous proposons une méthode de régression quantile extrême basée sur la théorie des valeurs extrêmes et l'apprentissage statistique pour surmonter ces défis. Conformément à l'approche de maxima de bloc (BM) de la théorie des valeurs extrêmes, nous approchons la distribution conditionnelle des BM par la distribution des valeurs extrêmes généralisée, dont les paramètres dépendent des covariables. Pour estimer ces paramètres, nous utilisons une méthode basée sur les forêts aléatoires généralisées. Les résultats obtenus à partir d'applications sur des données simulées mettent en évidence que notre méthode est compétitive avec d'autres approches de régression quantile.

Mots-clés. Régression quantile, Distribution des valeurs extrême généralisée, Forêt aléatoire généralisée, Maximum de vraisemblance, Bloc maxima, quantile extrême.

Abstract. Quantile regression is a commonly used statistical method in regression analysis. Unlike classical regression, which focuses on predicting the conditional mean of a dependent variable based on independent variables, quantile regression aims to predict conditional quantiles. Classical quantile regression methods face challenges, especially when the quantile of interest is extreme, due to the limited number of data available in the tail of the distribution or when the quantile function is complex. In this study, we propose an extreme quantile regression method based on extreme value theory and statistical learning to overcome these challenges. Following the Block Maxima (BM) approach of extreme value theory, we approximate the conditional distribution of BM by the generalized extreme value distribution, with parameters depending on covariates. To estimate these parameters, we employ a method based on generalized random forests. Results obtained from applications on simulated data highlight that our method is competitive with other quantile regression approaches.

Keywords. Quantile regression, Generalized Extreme Value Distribution, Generalized Random Forest, Maximum Likelihood, Block Maxima, Extreme Quantile.

1 Introduction

The modeling of extreme phenomena is crucial in various fields such as finance, meteorology, public health, and many others. Understanding the tails of the distribution of random variables is essential for assessing risks associated with rare but potentially devastating events. In this context, Extreme Value Theory (EVT) has emerged as a powerful tool for characterizing the behavior of extremes in a distribution. The main objective of this work is to explore extreme quantile regression by combining the fundamental principles of EVT with statistical learning techniques. Extreme quantile regression stands out for its ability to specifically model the quantiles of the distribution, providing detailed information about extremes. This approach is particularly valuable in contexts where a focus on extreme values is crucial for making informed decisions. Specifically, if $Y \in \mathcal{Y} \subset \mathbb{R}$ represents a random variable describing a risk factor dependent on a set of covariates represented by the random vector $X \in \mathcal{X} \subset \mathbb{R}^p$, the goal is to estimate the conditional extreme quantile given by :

$$\mathcal{Q}_\tau(x) = \inf\{y : F^{-1}_{Y|X=x}(y) \geq \tau\} \tag{1}$$

with τ close to 1 and $F^{-1}_{Y|X=x}$ the generalized inverse of the conditional distribution $Y|X = x$ (Koenker *et al.* (1978)).

Letting n be the sample size available for analysis and $\tau = \tau_n$ (depending on the sample size) be the order of the quantile we seek to estimate, classical methods for quantile estimation work well when $n(1 - \tau_n) \to \infty$ as $\tau_n \to 1$ (as $n \to +\infty$). In this case, the quantile to be estimated is within the sample, and there is also a large amount of data in the quantile range to be estimated. However, the situation is different when $n(1-\tau_n) \in [0, +\infty[$ and $\tau_n \to 1$ (as $n \to +\infty$). In the latter case, estimation requires extrapolation beyond the data range or into the tail of the distribution. In other words, the sought-after quantile is outside the range of the available sample, making estimation more complex and requiring specific approaches to handle these boundary situations. Thanks to the asymptotic results of extreme value theory (de Haan *et al.* (2006)), extrapolation beyond this data range is possible. Quantile regression methods face other challenges, including the complexity of the quantile function or the high dimensionality of the feature vector. To address the latter, we will use statistical learning methods, primarily the generalized random forests method by Athey *et al.* (2019), which is an extension of Breiman's classical random forests (2001).

Several works are proposed to address these different challenges of quantile regression. To address the first challenge, methods based on extreme value theory have been developed (see Chernozhukov *et al.* (2017) for an overview), while for the second, approaches based on statistical learning have been developed (*Meinshausen et al.* (2006), *Athey et al.* (2019), etc.). Recently, methods have been proposed that combine the Peaks Over Threshold (POT) approach of extreme value theory with statistical learning methods (Youngman (2018), *Farkas et al.* (2021), *Velthoen et al.* (2023), *Pasche et al.* (2023), *Gnecco et al.* (2024)). Our work aims to adapt the method of *Gnecco et al.* (2024) to the block maxima approach of extreme value theory. Explicitly, we model the conditional distribution of the relationship (1) using the generalized extreme value distribution, with parameters varying depending on the feature vector. These parameters are estimated by minimizing a local likelihood, weighted by weights obtained using generalized random forests.

2 Background

2.1 Block Maxima Approach

In this section, we will review some concepts regarding the approach of Extreme Value Theory (EVT), which we will use to address the first challenge of quantile regression stated in the introduction. The Block Maxima (BM) method is based on the limiting distribution of the maximum of a sequence of random variables $X_1, \cdots, X_m$ drawn independently and identically from a variable X with probability distribution F, as shown by Fisher *et al.* (1928), Gnedenko *et al.* (1943). These authors demonstrated that there exist normalization constants $a_m > 0$ and $b_m \in \mathbb{R}$ such that

$$\lim_{m \to +\infty} F^m(a_m x + b_m) = G_\xi(x), \ x \in \mathbb{R} \tag{2}$$

where G_ξ is a non-degenerate probability distribution defined by :

$$G_\xi(x) = \exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right),$$

with $1 + \xi x > 0$. Any function F satisfying equation (2) belongs to the max-domain of attraction of the distribution of extreme values G_ξ and is denoted as $F \in \mathcal{D}(G_\xi)$ (De Haan *et al.* (2006)).

If we consider $Y_1, \cdots, Y_N$ as a sequence of independent and identically distributed random variables according to the random variable Y with cumulative distribution function $F \in \mathcal{D}(G_{\xi_0})$ and corresponding normalization constants a_n and b_n, the Block Maxima (BM) method involves dividing the data into n blocks of the same size $m > 1$ (or nearly the same), denoted as $B_{k,m} = \{Y_{(k-1)m+1}, \cdots, Y_{km}\}$, where $k = 1, \cdots, n$. For any $m > 1$, the distribution of $Z_k = \max_{B_{k,m}}(Y_i)$ is F^m and satisfies (2), thus its distribution is approximated by the generalized extreme value (GEV) distribution with parameters (a_m, b_m, ξ_0). The BM method assumes that the distribution of these block maxima, denoted as Z_k, exactly follows the GEV distribution, and the resulting sequence of random variables $Z_1, \cdots, Z_n$ is also independent and identically distributed. Note that the choice of block size is crucial : increasing the block size results in large estimation variance and decreasing it will introduce bias in estimation. This boils down to a trade-off between bias and variance. For accurate estimation, a trade-off between bias and variance must be found when defining the blocks. The BM method is presented and discussed in the literature, and notable references include the books by Coles (2001) and De Haan *et al.* (2006). The GEV distribution is given by :

$$G_{\mu,\sigma,\xi}(z) = \begin{cases} \exp\left(-(1+\xi\frac{z-\mu}{\sigma})_+^{-\frac{1}{\xi}}\right) & \xi \neq 0 \\ \exp\left(-\exp(-\frac{z-\mu}{\sigma})\right) & \xi = 0 \end{cases}, \ \forall z \in \mathbb{R} \tag{3}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$ are the parameters of location, scale, and shape, respectively, and $a_+ = \max\{0, a\}$. The quantile of order τ of GEV is obtained through equation (1) and is given by :

$$Q_\tau = \begin{cases} \mu + \frac{\sigma}{\xi}\left((-\ln(\tau))^{-\xi} - 1\right), & \text{if } \xi \neq 0 \\ \mu + \xi \ln\left(-\ln\left(\tau\right)\right), & \text{if } \xi = 0 \end{cases} \tag{4}$$

Estimating the quantile of GEV amounts to estimating the parameters μ, σ, and ξ. There are several methods for estimating these parameters, with the most common being the maximum likelihood method. Our proposed method is based on a variation of this estimator.

2.2 Generalized random forests

The Generalized Random Forest (GRF) is an extension of the classical Random Forest method proposed by Breiman (2001). The Random Forest is an ensemble method for regression and classification that aggregates B trees fitted in parallel on bootstrap samples from the training dataset. Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be the real response variable and $X \in \mathcal{X} \subset \mathbb{R}^p$ be the vector of covariates. Classical regression analysis aims to obtain an estimate of the conditional mean $\eta(x) = \mathbb{E}(Y|X = x)$ of the response variable Y, given $X = x$. This is achieved by minimizing the expected quadratic loss :

$$\eta(x) = \arg\min_{z \in \mathcal{Y}} \mathbb{E}\left(l(Y - z)|X = x\right) \text{ with } l(y_1, y_2) = (y_1 - y_2)^2 \tag{5}$$

If $\eta_b(x)$ is the value predicted by the b-th tree for the test data $x \in \mathbb{R}^p$, in the case of regression, it is given by :

$$\hat{\eta}_b(x) = \sum_{i=1}^{n} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}}}{|\{i : X_i \in R_b(x)\}|} Y_i, \ b = 1, \cdots, B$$

where $R_b(x) \in \mathbb{R}^p$ denotes the region containing x in tree b and $|E|$ is the cardinality of set E. The prediction made by the Random Forest is given by :

$$\begin{aligned} \hat{\eta}(x) &= \frac{1}{B} \sum_{b=1}^{B} \hat{\eta}_b(x) \\ &= \sum_{i=1}^{n} w_n(x, X_i) Y_i \end{aligned} \tag{6}$$

with

$$w_n(x, X_i) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}}}{|\{i : X_i \in R_b(x)\}|}. \tag{7}$$

Each $w_n(x, X_i)$ in the Random Forest represents a similarity weight.

The Generalized Random Forest (GRF), introduced by Athey et al. in 2019, represents an extension of the Random Forest method. This approach preserves the attractive features of classical Random Forests while providing the flexibility to use custom loss functions for tree construction in the forest, i.e., the function l used in (5). An additional advantage of the Generalized Random Forest is that the similarity weights it generates capture the heterogeneity of the quantile function, unlike classical Random Forests. Indeed, the similarity weight $w_n(x, X_i)$ estimated by the classical forest is high for an observation X_i when $\mathbb{E}(Y|X = X_i) \approx \mathbb{E}(Y|X = x)$, but there are situations where this weight is high and $\mathcal{Q}(Y|X = X_i) \not\approx \mathcal{Q}(Y|X = x)$ (Athey *et al.* (2019), Gnecco*et al.* (2024)). We use the

Generalized Random Forest in our method to obtain the necessary similarity weights for estimating the parameters of the conditional GEV distribution. These parameters are then used to estimate the conditional quantile, as explained in Section (3).

2.3 Quantile Regression

When the conditional distribution $F(.|X = x)$ is continuous, the conditional quantile function given in equation (1) simplifies to :

$$\mathcal{Q}_\tau(x) = F^{-1}(\tau|X = x). \tag{8}$$

To our knowledge, the first appearance of random forest methods in the context of quantile regression dates back to Meinshausen (2006). He estimates the conditional distribution as follows :

$$\hat{F}\left(y|X = x\right) = \sum_{i=1}^{n} w_n(x, X_i)\mathbb{1}_{\{Y_i \leq y\}},$$

where the weights $w_n(x, X_i)$ are obtained from the classic random forest as defined in (7), and this estimated distribution is then used in (1) to obtain the conditional quantile. Another method is that of Athey *et al.* (2019), which is an application of generalized random forests with the quantile loss function given by $\rho_\tau(c) = c(\tau - \mathbb{1}_{\{c<0\}})$. In this work, we will use the term generalized random forest (grf) to denote the forest built using the quantile loss. Other methods are proposed to address the challenges outlined in the introduction by combining the peaks-over-threshold (POT) approach from extreme value theory with machine learning. Notable works include *Farkas et al.* (2021) using regression trees, *Velthoen et al.* (2023) with gradient boosting, *Pasche et al.* (2023) using neural networks, and *Gnecco et al.* (2024) employing generalized random forests. Our method is an extension of the BM approach proposed by *Gnecco et al.* (2024) and is explained in the following section.

3 GEV Extremal Random Forest

We propose a method for extreme conditional quantile regression addressing the challenges outlined in the introduction. To address the first issue, we model the tail of the conditional distribution $F(.|X = x)$ in equation (8) with the generalized extreme value distribution. To address the second issue, we use generalized random forests with quantile loss as the loss function to obtain the weights $w_n(x, X_i)$, which we use to estimate the parameters of the GEV distribution as weighted likelihood estimators, as proposed by Gnecco *et al.* (2024) to solve similar problems in the peaks-over-threshold (POT) approach.

In the conditional case, the parameters of the distribution GEV are functions of the covariate vector x. The conditional GEV distribution is obtained by replacing $\theta = (\mu, \sigma, \xi)$ in (3) with $\theta(x) = (\mu(x), \sigma(x), \xi(x))$ (where $\mu, \sigma, \xi : \mathcal{X} \to \mathbb{R}$). The quantile of order τ (with

τ close to 1) of conditional GEV is obtained through equation (8) and is given by :

$$Q_\tau(x) = \begin{cases} \mu(x) + \dfrac{\sigma(x)}{\xi(x)} \left((-\ln(\tau))^{-\xi(x)} - 1 \right), & \text{if } \xi(x) \neq 0 \\ \mu(x) + \xi(x) \ln\left(-\ln\left(\tau\right)\right), & \text{if } \xi(x) = 0 \end{cases} \tag{9}$$

Estimating the conditional GEV quantile amounts to estimating the parameters $\mu(x)$, $\sigma(x)$, and $\xi(x)$. Our proposed method involves estimating these parameters and then substituting them into equation (9) to obtain an estimation of $Q_\tau(x)$.

Here, we propose an alternative form of the classical maximum likelihood estimator of the GEV distribution. As suggested by Gnecco *et al.* (2024) for the POT approach, we estimate $\theta(x)$ by $\hat{\theta}(x) = (\hat{\mu}(x), \hat{\sigma}(x), \hat{\xi}(x))$, which is the weighted maximum likelihood estimator, i.e., minimizing

$$L_n(\theta, x) = \sum_{i=1}^{n} w_n(x, X_i) l_{\theta(x)}(z_i) \tag{10}$$

with $l_\theta(z_i)$ such that :

— when $\xi(x) \neq 0$,

$$l_{\theta(x)}(z_i) = \log(\sigma(x)) + (1 + \frac{1}{\xi(x)}) \log\left(1 + \xi(x)\frac{z_i - \mu(x)}{\sigma(x)}\right) + \left(1 + \xi(x)\frac{z_i - \mu(x)}{\sigma(x)}\right)^{\frac{-1}{\xi(x)}}$$

if $1 + \xi(x)\frac{z_i - \mu(x)}{\sigma(x)} > 0$ and $+\infty$ otherwise.

— when $\xi(x) = 0$,

$$l_{\theta(x)}(z_i) = \log(\sigma(x)) + \left(\frac{z_i - \mu(x)}{\sigma(x)}\right) + \exp\left(\frac{z_i - \mu(x)}{\sigma(x)}\right)$$

the $w_n(x, X_i,)$ $i = 1, \cdots, n$ are obtained using the generalized random forests method. The weighted maximum likelihood estimator is defined as :

$$\hat{\theta}(x) = \arg\min_{\theta(x) \in \Theta} L_n(\theta, x).$$

The parameter ξ is important as it determines the shape of the distribution's tail. Therefore, significant attention is paid to it in the estimation of the GEV distribution in the literature. For instance, Bücher *et al.* (2020) propose a maximum likelihood estimator for the GEV distribution by penalizing only the parameter ξ, in the non-conditional case. Gnecco *et al.* (2024) also suggested a maximum likelihood estimator for the generalized Pareto distribution by penalizing the parameter ξ in the conditional case. Building upon these works, primarily on the study by Gnecco *et al.* (2024), we also penalize ξ, considering the penalized maximum likelihood estimator defined as :

$$\hat{\theta}_{\text{pena}}(x) = \arg\min_{\theta(x) \in \Theta} L_n(\theta, x) + \lambda(\xi - \hat{\xi})^2, \tag{11}$$

where $0 \leq \lambda$ is the penalty parameter, and $\hat{\xi}$ is considered as the shape parameter of the maximum likelihood estimator obtained according to equation (10) with weights $w_n(x, X_i)$ all equal to 1, as used in Gnecco *et al.* (2024) for the POT approach. Cross-validation method is employed for the selection of λ.

4 Simulation results

This section presents some application results of our method on simulated data, based on the simulation part of the works by Gnecco *et al.* (2024) and Velthoen *et al.* (2023). We generate $N = 90000$ (denoted *ntrain* in the figures) i.i.d. samples of $X \sim \mathcal{U}_{[-1,1]^p}$ and the conditional distribution $Y|X = x \sim \gamma(x)T_{\nu(x)}$, where T_k is the Student's distribution with k degrees of freedom. For the figures below, we use $\gamma(x) = 1 + \mathbb{1}_{\{x_1>0\}}$ (denoted *model1* in the figures) and $\nu(x) = 4 - (x_1^2 - 2x_2^2 + x_3^2)$ (denoted *quadratic* in the figures) for any test data $x \in \mathbb{R}^p$, where x_i denotes the i^{th} component of x, and $p = 40$. The conditional quantile function depends only on the first three components of x, and the remaining components are noise. Our method involves dividing the N data into n blocks of size 30 and considering the maxima of the formed blocks, so only $n = 3000$ data are considered for the training process to obtain the weights $w_n(x, X_i)$ and the adjustment of the likelihood given in (11). We evaluate our method on test data $\{x^i\}_{i=1}^{N'}$ with $N' = 3000$ (denoted n_{test} in the figures), independent of the training data and generated by the Halton sequence on the cube $[-1, 1]^p$ (with $x^i \in \mathbb{R}^p$ representing the i^{th} data of the test sample), where we divide into blocks of size 30 and consider only the maxima per block as test data. Thus, only $n' = 100$ data are considered for the model evaluation.

To highlight the performance of our method, designated as GEV, on the graphs below, we compare its Mean Integrated Squared Error (MISE) with other approaches using statistical learning. This includes the **quantile regression forests** by Meinshausen (2006), denoted as QRF, the method of **generalized random forests** by Athey *et al.* (2019), denoted as GRF, and the unconditional method denoted as Uncond. Note that all the models considered are trained on the same dataset, which consists of the formed maxima, hence a set of $n = 3000$ training data. The test is also performed on the same dataset, namely a set of $n' = 100$ test data.

The results show that our method performs better for conditional quantile estimation, mainly for quantiles close to 1, more clearly for the quantile order from $\tau = 0.99$ to $\tau = 0.9999$, as shown in Figure (1). Figure (2) shows the variation of MISE as a function of the predictor size p for a quantile order fixed at $\tau = 0.999$, where we see that our method is competitive, mainly with the GRF method. We also see that it performs better than other models for $p = 40$.

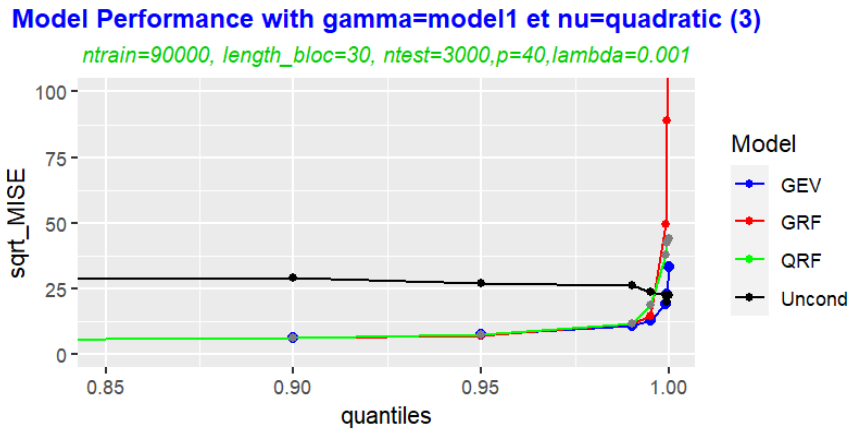

FIGURE 1 – Experiment 1

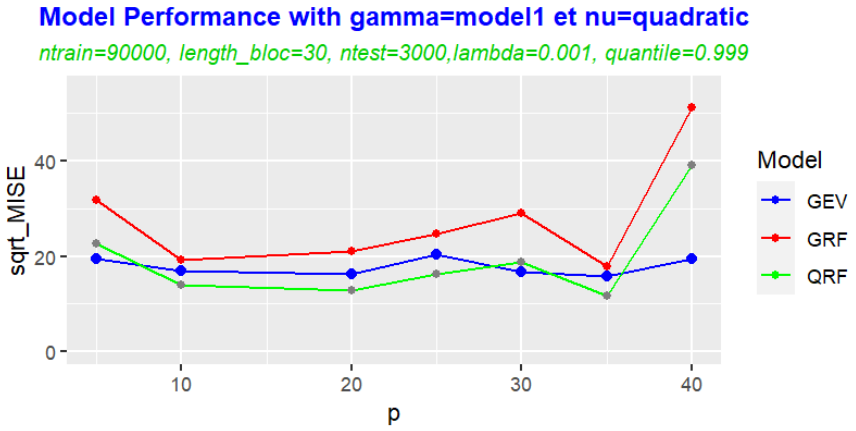

FIGURE 2 – Experiment 1

Bibliography

Coles, S., Bawa, J.,Trenner, L. et al. (2001), An introduction to statistical modeling of extreme values, *Springer*, 208.

Bücher, A.,Lilienthal,J. et al., Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis, *Extremes*, 24, pp 325-348.

Meinshausen, N. (2006), Quantile Regression Forests, *Journal of Machine Learning Research*, 7, pp. 983-999.

Breiman, L. (2001), Random forests, *Machine Learning*, 45, pp 5–32.

Athey, S. and Tibshirani, J. and Wager, S. (2019), Generalized random forests, *The Annals of Statistics*, 2, pp 1148-1178.

Gnecco, N., Terefe, E. M. and Engelke, S. (2024), Extremal Random Forests, *Journal of the American Statistical Association*, 0, pp 1-24.

Velthoen, J., Dombry, C., Cai, JJ. et al. (2023), Gradient boosting for extreme quantile regression, Extremes, 26, pp 639–667.

Pasche, O. C. and Engelke, S.(2023), Neural Networks for Extreme Quantile Regression with an Application to Forecasting of Flood Risk, arXiv.

Dombry, C. (2015), Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework, *Bernoulli*, 1, pp 420–436.

Koenker, R., Bassett G. (1978), Regression Quantiles, *Econometrica*, 46(1) :33.

Chernozhukov, V., Fernández-Val, I., Kaji, T. (2016), Extremal quantile regression : An overview, *arXiv preprint arXiv :1612.06850*

Youngman, B. (2019), Generalized Additive Models for Exceedances of High Thresholds With an Application to Return Level Estimation for U.S. Wind Gusts, *Journal of the American Statistical Association*, 114, pp 1865–1879.