Supplementary material for: "Extreme Storm Surge estimation and projection through the Metastatistical Extreme Value Distribution"

Maria Francesca Caruso[1] and Marco Marani[1]

[1]Department of Civil, Architectural, and Environmental Engineering, University of Padova, 35131, Padova, Italy

Correspondence: Caruso, M.F. (mariafrancesca.caruso@phd.unipd.it)

Contents of this file

1. Text from T1 to T3
2. Figures from S1 to S6

Introduction

This supplementary material contains two sections: 1) a general overview of the extreme value theory represented both in their conventional form (Block Maxima) and in their "threshold" form based on the generalized Pareto distribution (Peak-Over-Threshold); 2) complementary figures to the main text.

T1. Extreme Value Theory

The Extreme Value Theory (EVT) is a statistical technique that provides a theoretical framework to quantify the occurrence probability of random variables at unusually high (or low) extreme events. The cornerstone of EVT is the three type theorem introduced by Fisher and Tippett (1928) and later proved by Gnedenko (1943). This results led Gumbel (1958) to introduce a statistical methodology for extreme values. The basic idea is to study the statistical behaviour of:

$$M_n = max(X_1, \ldots, X_n) \tag{1}$$

where X_i (for $i = 1 \ldots, n$) are a sequence of independent and identically distributed (i.i.d.) random variables having a common distribution function (F). The distribution function of M_n is given by the n^{th} power of F:

$$\begin{aligned} Pr\{M_n \leq x\} &= Pr\{X_1 \leq x, X_2 \leq x, \cdots, X_n \leq x\} \\ &= Pr\{X_1 \leq x\} Pr\{X_2 \leq x\} \cdots Pr\{X_n \leq x\} \\ &= F^n(x) \end{aligned}$$

The classical EVT focuses on the asymptotic behaviour of this distribution. Under certain circumstances, it can be shown that exist scaling constants $a_n > 0$ and b_n, that allow to obtain a nondegenerate distribution ($H(x)$, i.e. it is not always either 0 or 1), such that:

$$P(\frac{M_n - b_n}{a_n} \leq x) = F^n(a_n x + b_n) \rightarrow H(x), \; as \; n \rightarrow \infty$$

The corresponding normalized variable $M_n^* = \frac{M_n - b_n}{a_n}$ has a limiting distribution that must be one of the three types of extreme value distributions (Gumbel, Fréchet and Weibull) characterized by different behaviours and shapes of the tail. Von Mises (1936) proposed a single distribution which combines all three types of asymptotic extreme value distributions into a single family known as Generalized Extreme Value (GEV) distribution:

$$H(x; \mu, \psi, \xi) = exp\{-[1 + \frac{\xi}{\psi} \cdot (x - \mu)]\}^{-1/\xi} \tag{2}$$

where:

1. μ is a location parameter;
2. ψ is a scale parameter;
3. ξ is a shape parameter which controls the type of the tail distribution: 1) $\xi \rightarrow 0$ defines the light tailed case (Gumbel type or EV1) characterized by an exponential tail; 2) $\xi > 0$ identifies the heavy tailed case (Fréchet type or EV2) described by a power law; 3) $\xi < 0$ gives the short tailed case (negative Weibull case or EV3) which has a bounded upper tail.

The EVT identifies the GEV distribution as a general model to describe the distribution of extreme events. The three GEV parameters can be estimated by using well-known statistical methods: maximum-likelihood, probability weighted moments or Bayesian methods.

In many applications to environmental and hydrology processes, two fundamental approaches are widely used to extreme value statistics, based on: 1) maxima over some fixed time period (Block Maxima), and 2) exceedances over high threshold (Peaks-Over-Threshold). The following subsections outline the concepts underlying these methods.

T2. Modeling block maxima

The Block Maxima (BM) approach consists of dividing the observation sample into a sequence of maximum values extracted from blocks of fixed time intervals and fitting the GEV distribution (Eq. (2)) to the set of block maxima obtained.
The choice of the suitable block size is a preliminary step, amounting to a trade-off between bias and error variance. In most environmental processes is commonly used a block size of one year leading to study the annual maxima time series.
Generally, for practical applications we are interested in estimating the T-years return levels associated with the extreme values. If the GEV distribution is an appropriate model for block maxima, it is possible to estimate the quantile x_T which is the level expected to be exceeded on average once every $1/T$ years. The cumulative probability is given by $H(x_T) = 1 - 1/T$ and the estimates of extreme quantiles of the annual maxima distribution are then obtained by inverting the Eq. 2:

$$x_T = \begin{cases} \mu - \frac{\sigma}{\xi} \cdot \{1 - [-ln(1 - \frac{1}{T})]^{-\xi}\} & \xi \neq 0 \\ \mu - \frac{\sigma}{\xi} \cdot \{1 - [-ln(1 - \frac{1}{T})]\} & \xi = 0 \end{cases}$$

The BM method is commonly used both for its simplicity and also because the annual maxima are undoubtedly independent variables. Despite its simplicity, the BM method is a wasteful approach because only one value from each block is used with loss of some important available information.

T3. Peak-over-threshold

To overcome the limitations of the previous method, an alternative approach is widely used in the study of the extreme events known as the Peak-Over-Threshold (POT, introduced by Balkema and de Haan (1974) and Pickands (1975)). The POT method allows to analyze all data exceeding a specific threshold value. The idea under this approach is to set an high threshold u, and to study all the exceedances of u.
Suppose X_i (for $i = 1, \ldots, n$) is a sequence of i.i.d. random variables whose distribution function is F and let define the excesses over u as $Y_i = X_i - u$ conditioned on $X_i > u$, the cumulative distribution of exceedances is defined by:

$$Pr\{Y_i \leq y\} = Pr\{X_i \leq u + y | X_i > u\} = F_u(u) = \frac{F(u+y) - F(u)}{1 - F(u)}$$

Pickands (1975) established the connection between EVT and the Generalized Pareto Distribution (GPD). He showed that a GPD approximation is possible if the distribution of the X_i satisfies $Pr\{M_n \leq z\} \approx G(z)$, where M_n and $G(z)$ are given respectively by Eq. (1) and Eq. (2). Moreover, for very large threshold u, the distribution function of the exceedances ($Y_i =$

$X_i - u$) can be approximated by the generalized Pareto family:

$$H(y;\sigma_u,\xi) = 1-(1+\frac{\xi}{\sigma_u}\cdot y)^{(-1/\xi)}$$

a) defined on $\{y : y > 0$ and $(1+\xi/\sigma_u \cdot y) > 0\}$, where $\sigma_u = \sigma + \xi(u-\mu)$;

b) with two parameters: shape (σ) and scale (ξ).

This result implies that, if block maxima have approximate distribution G, then threshold excesses have a corresponding approximate distribution within the generalized Pareto family. In this case, the parameters of the GPD of threshold exceedances are determined by those of the associated GEV distribution of block maxima. In particular, the shape parameter (ξ) is equal to that of the corresponding GEV distribution and is invariant to block size (n) while σ_u is unaffected by changes in u and σ. As it happens for the GEV distribution, the shape parameter is dominant in determining the behavior of the GPD tail:

1. $\xi > 0$ the distribution has no upper limit (equivalent to Pareto distribution) and the tail distribution function satisfies $1-H(y) \sim cy^{(-1/\xi)}$ with $c > 0$, i.e. the polynomial distribution;
2. $\xi < 0$ the distribution of excesses has an upper endpoint at $\omega_F = \sigma_u/(|\xi|)$;
3. $\xi = 0$ the distribution is unbounded. This case is interpreted as $\xi \rightarrow 0$ i.e. the exponential distribution with mean σ.

Fixed a threshold u, the number of exceedances is assumed to be a random variable itself and it is modeled with Poisson distribution leading to the so called Poisson-GPD model. According to this model, if we assume the number of yearly exceedances to have a Poisson distribution (with mean λ) and all the exceedances to be independent realizations and GPD distributed, the probability that the annual maximum of the process is less than a certain value x is:

$$\begin{aligned}
Pr\{max_{1\le i\le N} Y_i \le x\} = H(x-u;\lambda,\xi,\sigma) = exp\{-\lambda[1+\frac{\xi}{\sigma}\cdot(x-u)]\}^{-1/\xi} = \\
= Pr\{N=0\} + \sum_{n=1}^{+\infty} Pr\{N=n, Y_1 \le x, \cdot, Y_n \le x\} = \\
= e^{-\lambda} + \sum_{n=1}^{+\infty} \frac{\lambda^n \cdot e^{-\lambda}}{n} \cdot \{1-(1+\frac{\xi}{\sigma_u}\cdot(x-u))^{(-1/\xi)}\}^n = \\
= exp\{-\lambda[1+\frac{\xi}{\sigma}\cdot(x-u)]\}^{-1/\xi}
\end{aligned}$$

This property suggests that the probability distribution of the annual maxima of a GPD-Poisson model is the same as the GEV distribution (see Eq. (2)). The GEV and GPD models are consistent with one another if $\xi^{GEV} = \xi^{GPD}$, $\sigma = \psi + \xi(u-\mu)$ and $\lambda = [1+\frac{\xi}{\psi}\cdot(u-\mu)]^{-1/\xi}$.

The POT method allows to estimate the GEV parameters based on a greater number of events, whereas the traditional fitting methods considering only the annual maxima with consequent distortion in the shape of the tail. However, the optimal threshold selection requires particular attention in order to satisfy the two hypothesis underlying the method: 1) the number of events/year is Poisson-distributed; 2) exceedances over the threshold come from a Generalized Pareto Distribution (GPD). Threshold choice involves a trade-off between variance, which increases with higher thresholds due to the smaller number of excesses, and bias, which arise when the threshold is too low.

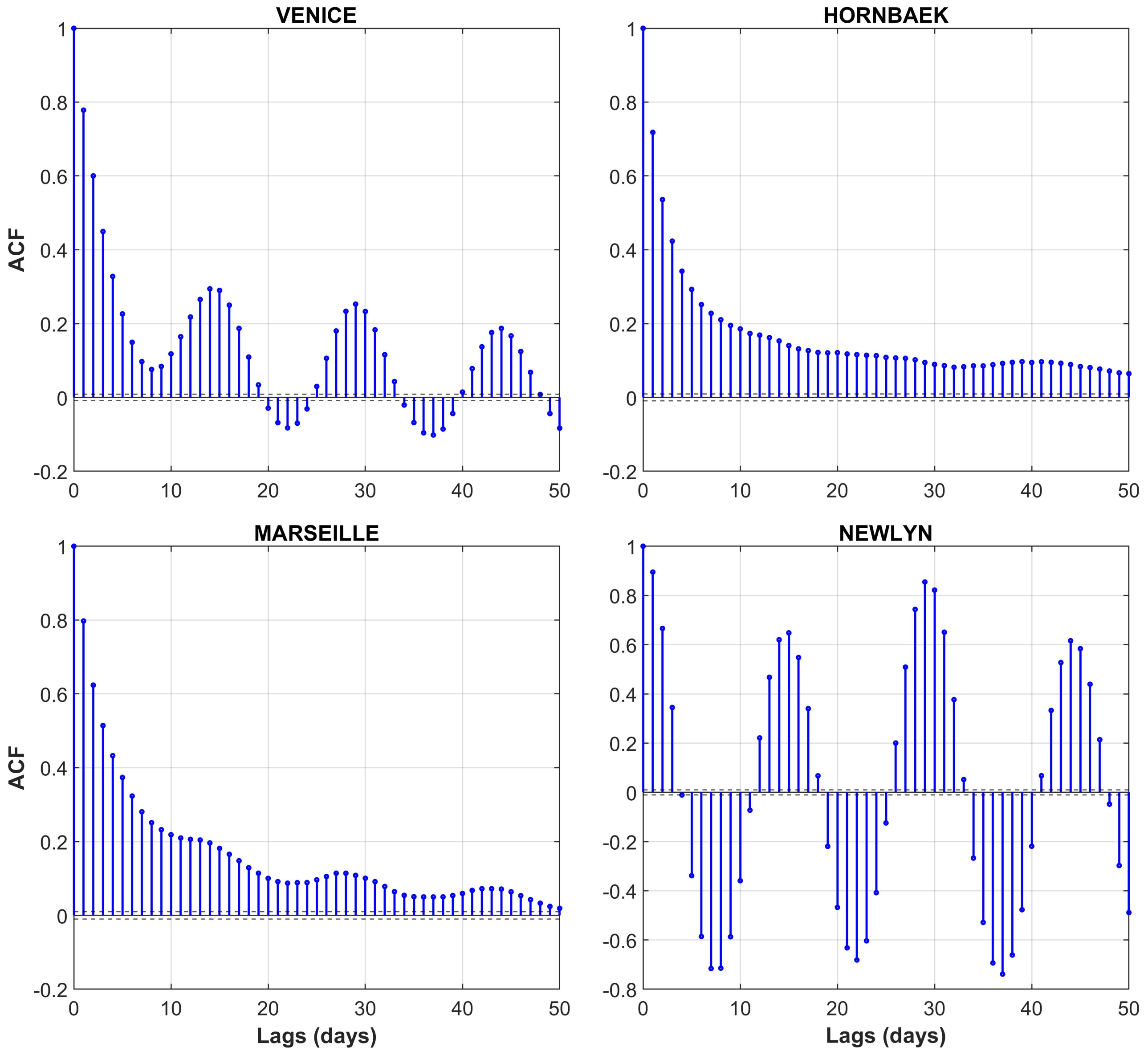

Figure S1. Correlogram plots for daily maxima levels and for all observed sites.

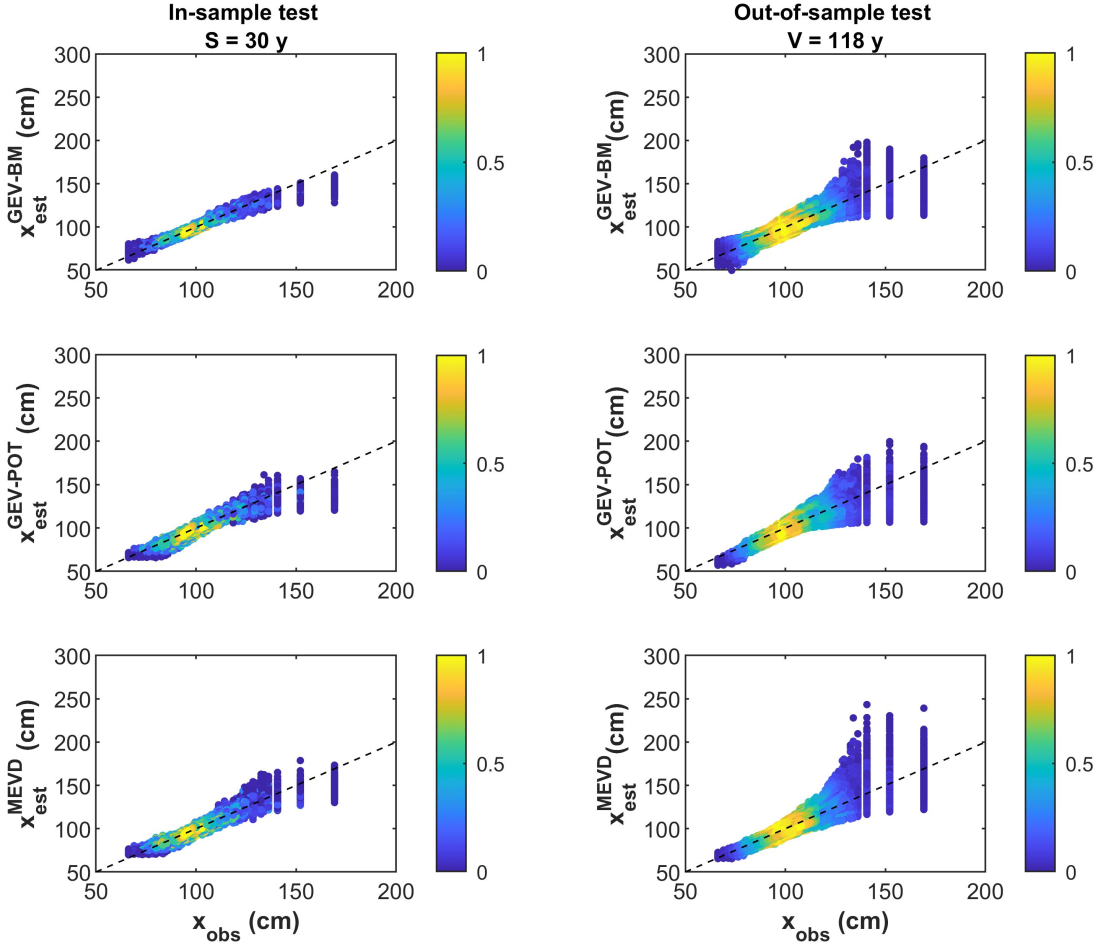

Figure S2. VENICE (IT) - QQ plots of extreme storm surge quantiles computed for the GEV-based approaches (BM and POT) and MEVD for the Venice station. The plots are obtained as a result of the cross validation method used to test the global performance of the models and are estimated for 1,000 random realizations and for sample size: a) S = 30 years (in-sample-test on the left column); b) $V = L - S$ years (out-of-sample test on the right column). The colours represent the points density around the 45° line (black dashed line) corresponding to the best fit.

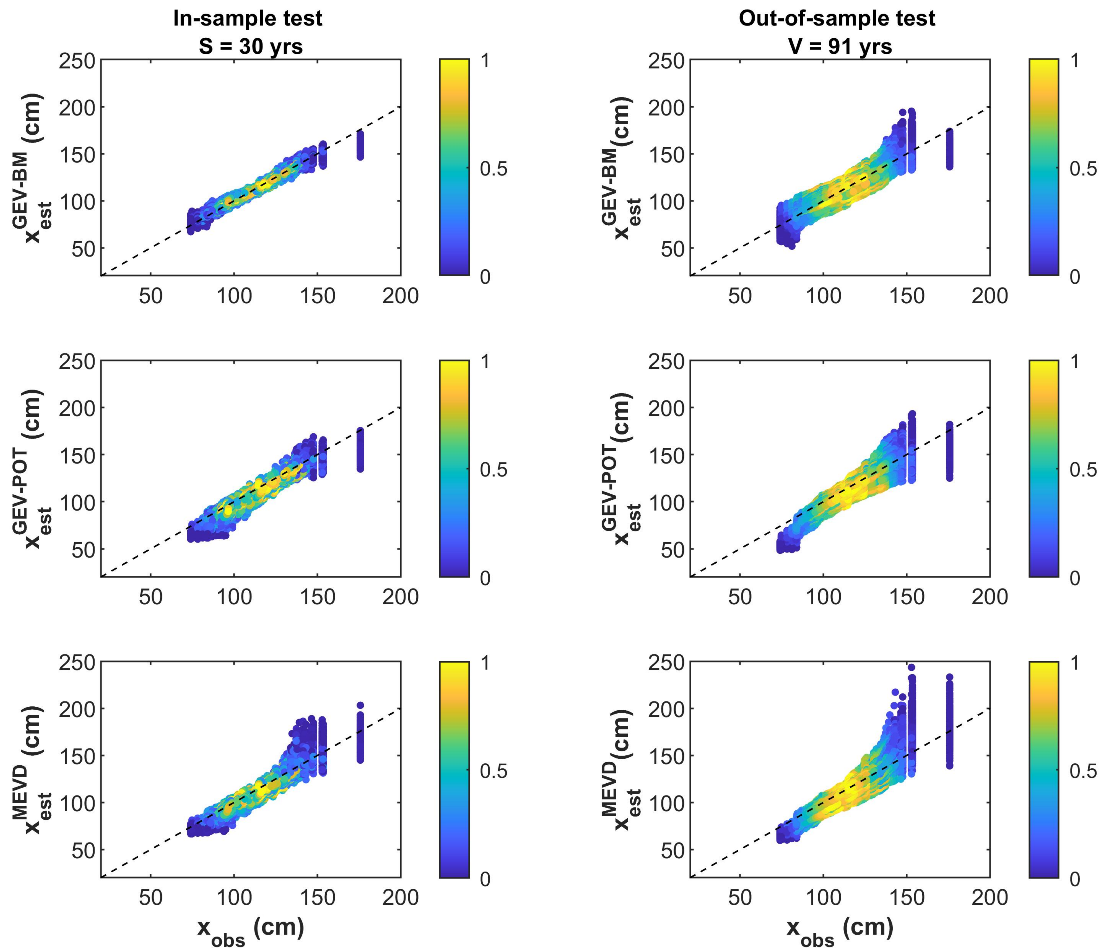

Figure S3. HORNBÆK (DK) - QQ plots of extreme storm surge quantiles computed for the GEV-based approaches (BM and POT) and MEVD for the Hornbæk station. The plots are obtained as a result of the cross validation method used to test the global performance of the models and are estimated for 1,000 random realizations and for sample size: a) S = 30 years (in-sample-test on the left column); b) $V = L - S$ years (out-of-sample test on the right column). The colours represent the points density around the 45° line (black dashed line) corresponding to the best fit.

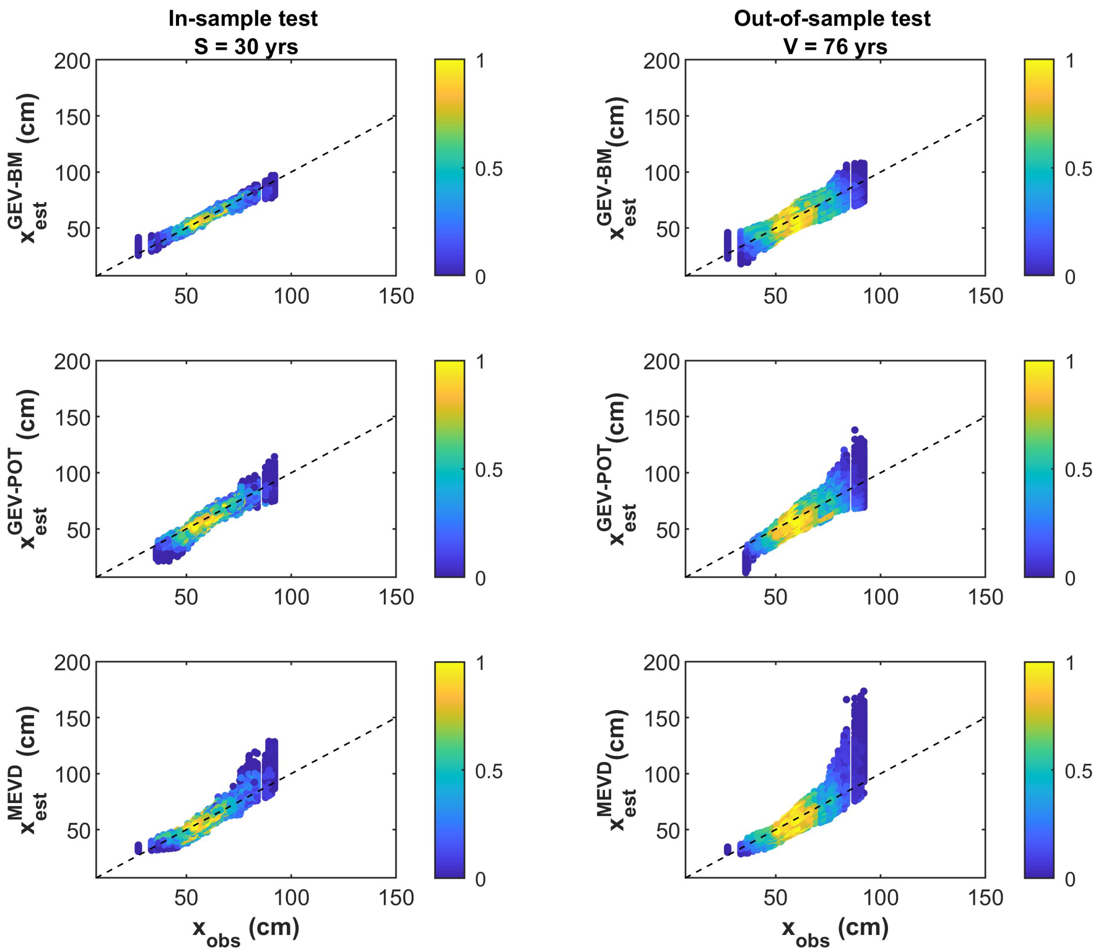

Figure S4. MARSEILLE (FR) - QQ plots of extreme storm surge quantiles computed for the GEV-based approaches (BM and POT) and MEVD (with parameter estimation *on non-overlapping sub-samples of fixed size (5 years)*) for the Marseille station. The plots are obtained as a result of the cross validation method used to test the global performance of the models and are estimated for 1,000 random realizations and for sample size: a) S = 30 years (in-sample-test on the left column); b) $V = L - S$ years (out-of-sample test on the right column). The colours represent the points density around the 45° line (black dashed line) corresponding to the best fit.

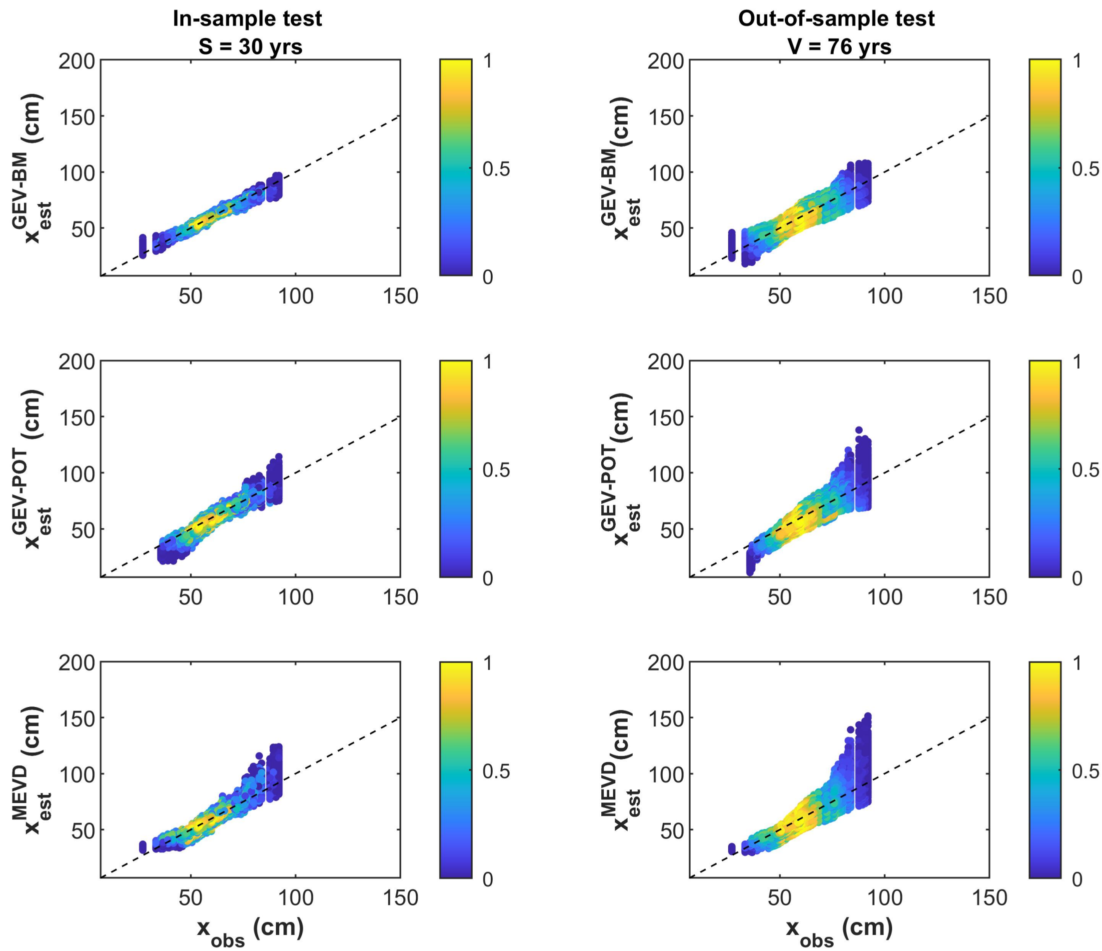

Figure S5. MARSEILLE (FR) - QQ plots of extreme storm surge quantiles computed for the GEV-based approaches (BM and POT) and MEVD (with parameter estimation *on data from the whole calibration sample*) for the Marseille station. The plots are obtained as a result of the cross validation method used to test the global performance of the models and are estimated for 1,000 random realizations and for sample size: a) S = 30 years (in-sample-test on the left column); b) $V = L - S$ years (out-of-sample test on the right column). The colours represent the points density around the 45° line (black dashed line) corresponding to the best fit.

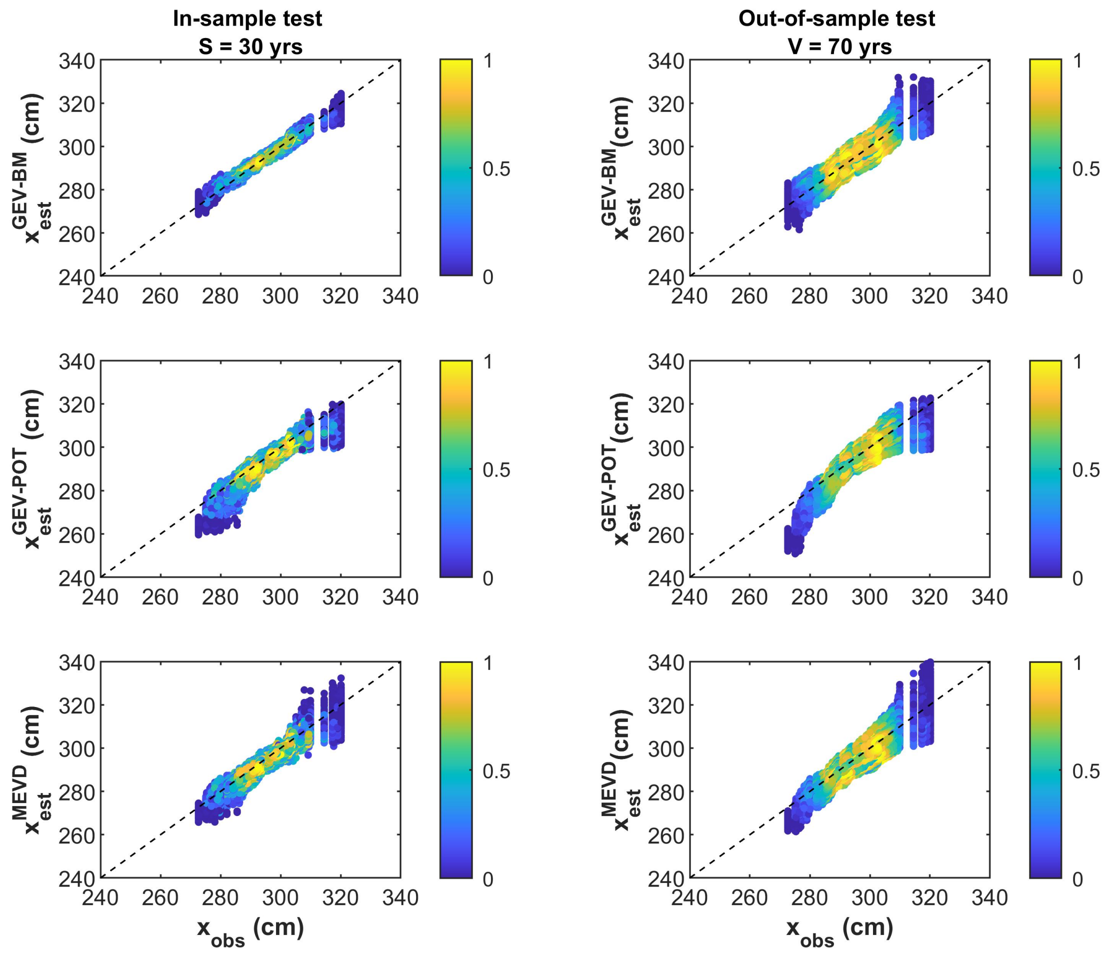

Figure S6. NEWLYN (UK) - QQ plots of extreme storm surge quantiles computed for the GEV-based approaches (BM and POT) and MEVD for the Newlyn station. The plots are obtained as a result of the cross validation method used to test the global performance of the models and are estimated for 1,000 random realizations and for sample size: a) S = 30 years (in-sample-test on the left column); b) $V = L - S$ years (out-of-sample test on the right column). The colours represent the points density around the 45° line (black dashed line) corresponding to the best fit.

References

Balkema, A. A. and de Haan, L.: Residual life time at great age, Ann. Probab., 2, 792–804, https://doi.org/10.1214/aop/1176996548, 1974.

Fisher, R. A. and Tippett, L. H. C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample, Math. Proc. Cambridge, 24, 180–190, https://doi.org/10.1017/S0305004100015681, 1928.

Gnedenko, B. V.: Sur la distribution limite du terme maximum d'une serie aleatoire, Ann. Math., 44, 423–453, 1943.

Gumbel, E. J.: Statistics of Extremes, Columbia University Press, New York, 1958.

Pickands, III, J.: Statistical inference using extreme order statistics, Ann. Stat., 3, 119–131, https://doi.org/10.1214/aos/1176343003, 1975.

Von Mises, R.: La distribution de la plus grande de n valuers, Rev. math. Union interbalcanique, 1, 141–160, 1936.