Hydrol. Earth Syst. Sci., 14, 1909–1917, 2010
www.hydrol-earth-syst-sci.net/14/1909/2010/
doi:10.5194/hess-14-1909-2010

Hydrology and Earth System Sciences

Exploiting the information content of hydrological "outliers" for goodness-of-fit testing

F. Laio, P. Allamano, and P. Claps

DITIC, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Received: 6 July 2010 – Published in Hydrol. Earth Syst. Sci. Discuss.: 22 July 2010
Revised: 27 September 2010 – Accepted: 5 October 2010 – Published: 12 October 2010

Abstract. Validation of probabilistic models based on goodness-of-fit tests is an essential step for the frequency analysis of extreme events. The outcome of standard testing techniques, however, is mainly determined by the behavior of the hypothetical model, $F_X(x)$, in the central part of the distribution, while the behavior in the tails of the distribution, which is indeed very relevant in hydrological applications, is relatively unimportant for the results of the tests. The maximum-value test, originally proposed as a technique for outlier detection, is a suitable, but seldom applied, technique that addresses this problem. The test is specifically targeted to verify if the maximum (or minimum) values in the sample are consistent with the hypothesis that the distribution $F_X(x)$ is the real parent distribution. The application of this test is hindered by the fact that the critical values for the test should be numerically obtained when the parameters of $F_X(x)$ are estimated on the same sample used for verification, which is the standard situation in hydrological applications. We propose here a simple, analytically explicit, technique to suitably account for this effect, based on the application of censored L-moments estimators of the parameters. We demonstrate, with an application that uses artificially generated samples, the superiority of this modified maximum-value test with respect to the standard version of the test. We also show that the test has comparable or larger power with respect to other goodness-of-fit tests (e.g., chi-squared test, Anderson-Darling test, Fung and Paul test), in particular when dealing with small samples (sample size lower than 20–25) and when the parent distribution is similar to the distribution being tested.

1 Introduction

An outlying observation, or outlier, is a record that appears to deviate markedly from other members of the sample to which it belongs (Grubbs, 1969). It is clear from this definition that an outlier can occur either because data values are incorrect (for example due to inaccurate recording or transcription), or because the population has an heavy-tailed distribution, which increases the probability of having single observations which stand way apart from the others (e.g., Barnett and Lewis, 1994). Still, the practitioners are often tempted to omit the outliers from the available data samples, because this choice allows one to proceed with the statistical analysis using simpler and well-behaved distributions. While application of outlier detection methods may be extremely important for screening the data and recognizing gross errors, unsupervised outlier rejection may result in a remarkable loss of information, in particular when the behavior of the tails of the distribution is fundamental to the performed statistical analyses (which of course is exactly the case in the frequency analysis of hydrological extremes). To quote Gumbel (1960): "The rejection of outliers on a purely statistical basis is and remains a dangerous procedure. Its very existence may be a proof that the underlying population is, in reality, not what it was assumed to be". In this paper we accept this viewpoint and show how extreme observations, possibly marked as outliers, can be used to select the probabilistic model for the frequency analysis of extreme events.

This objective changes the statement of the problem, from one where the extreme observations are screened as potential outliers to be rejected, to one where they are used for validation of a probabilistic model, i.e. for goodness-of-fit purposes.

Correspondence to: F. Laio
(francesco.laio@polito.it)

Published by Copernicus Publications on behalf of the European Geosciences Union.

A simple approach to the validation of the probabilistic model in hydrology is based on plotting the data on probability charts (Stedinger et al., 1992) and verifying if the observations fall approximately on a straight line. Limitations of this approach derive from (i) the subjectivity inherent in the visual verification of the alignment of the empirical points, and (ii) the fact that the method is available only for two-parameters distributions. Several different testing techniques have been developed for application to the frequency analysis of hydrological extremes as, for example Vogel (1986), Ahmad et al. (1988), Chowdhury et al. (1991), Vogel and McMartin (1991), Fill and Stedinger (1995), Wang (1998) and Laio (2004). Alternatively, the model validation issue has been recast as a model selection problem, where several candidate models are compared, and the best model to represent the available data is selected (e.g., Strupczewski et al., 2002; Mitosek et al., 2006; Di Baldassarre et al., 2008; Laio et al., 2009). However, a common drawback of goodness-of-fit and model selection techniques is that their outcome is mainly determined by the behavior of the hypothetical model in the central part of the distribution, while the behavior in the tails of the distribution, is relatively unimportant for the outcome of the test: standard goodness-of-fit tests seldom reveal an ill-fitting tail without a very large amount of data (Bryson, 1974).

This problem can be overcome by using the maximum-value test, which was originally proposed by Grubbs (1969) as a technique for outliers detection in a Gaussian setting, and subsequently extended to Gumbel-distributed parents by Rossi et al. (1984). The test is specifically targeted to verify if the maximum (or minimum) values in the sample are consistent with the hypothesis that the distribution $F_X(x)$ corresponds to the real parent distribution. However, usual applications of this test to non-Gaussian distributions are complicated by the fact that the parameters of the hypothetical distribution, $F_X(x)$, are unknown and need to be estimated using the same sample used for the test, which in turn implies that the acceptance region for the test needs to be calculated through numerical simulation (see e.g. Rossi et al., 1984).

2 Methods

This section is devoted to describing the basic features of the standard maximum-value test (Sect. 2.1), and of the necessary modifications to the testing procedure due to parameter estimation on the same data sample used for testing (Sect. 2.2). In the following of the paper, the standard version of the test will also be referred to as SMV test (Standard Maximum Value), and the modified version as MMV (Modified Maximum Value) test.

2.1 Basic definitions

Suppose that $x_{(1)}\leq\ldots\leq x_{(n)}$ is an ordered sample of n independent observations from an unknown parent distribution with cumulative distribution function $G_X(x)$; also suppose that one wishes to test the null hypothesis that the data were sampled from a distribution $F_X(x|\Theta)$, where Θ is a vector of parameters that need to be estimated. In symbols, the null hypothesis to be tested is H_0:$G_X(x)=F_X(x|\Theta)$. In this paper we will consider two- and three-parameter distributions as candidate operational models, i.e. as hypothetical parent distributions $F_X(x)$. More in detail, we will direct our attention to: (i) two-parameter distributions belonging to the location-scale family, i.e. distributions that can be written in the form

$$F(x|\theta_1,\theta_2)=\Phi\left(\frac{x-\theta_1}{\theta_2}\right) \tag{1}$$

where $\Phi(\cdot)$ is a generic function, θ_1 is a position (or location) parameter and θ_2 is a scale parameter; (ii) three-parameter distributions belonging to the location-scale-shape family, characterized by the property

$$F(x|\theta_1,\theta_2,\theta_3)=\Phi\left(\frac{x-\theta_1}{\theta_2};\theta_3\right) \tag{2}$$

where $\Phi(\cdot;\theta_3)$ is a generic function with two arguments, the second argument being the shape parameter θ_3. Most distribution commonly used in the hydrologic practice belong to one of the two families indicated above, including the Gumbel distribution, the normal distribution, the two-parameter exponential distribution, the GEV distribution, the Pearson type III (or three-parameter gamma) distribution, etc. (see Table 1 for details on the parametrization adopted in this paper). Other commonly used distributions, as for example the lognormal, log-Pearson type III, and Frechet distributions, can be traced back to these families with a preliminary log-transformation of the data.

Parameter estimation is here carried out with the L-moments method as defined, for example, in Hosking and Wallis (1997), because this method is especially suitable to be used in combination with the maximum value test, as will be clarified in the following. The method of L-moments provides parameter estimators based on the matching of distribution and sample L-moments. The former are defined as

$$\begin{aligned}
L_1 &= \int_0^1 x(u)du \\
L_2 &= \int_0^1 (2u-1)x(u)du \qquad (3)\\
L_3 &= \int_0^1 (6u^2-6u+1)x(u)du,
\end{aligned}$$

Table 1. Probability models considered in this paper. $\Gamma[\cdot]$ is the gamma function.

Distribution	Acronym	CDF or PDF	Range
Exponential	EXP	$F(x,\theta)=1-\exp\left[-(x-\theta_1)/\theta_2\right]$	$\theta_1<x<\infty$
Gumbel or Extreme Value Type I	EV1	$F(x,\theta)=\exp\left[-\exp\left[-\frac{x-\theta_1}{\theta_2}\right]\right]$	$-\infty<x<\infty$
Normal or Gaussian	NORM	$f(x,\theta)=\frac{1}{\sqrt{2\pi}\theta_2}\exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right]$	$-\infty<x<\infty$
Generalized Extreme Value	GEV	$F(x,\theta)=\exp\left[-\left(1+\frac{\theta_3(x-\theta_1)}{\theta_2}\right)^{-1/\theta_3}\right]$	$\frac{\theta_3(\theta_1-x)}{\theta_2}<1$
Gamma or Pearson Type 3	GAM	$f(x,\theta)=\frac{1}{\lvert\theta_2\rvert\Gamma[\theta_3]}\left(\frac{x-\theta_1}{\theta_2}\right)^{\theta_3-1}\exp\left[-\frac{x-\theta_1}{\theta_2}\right]$	$\frac{x-\theta_1}{\theta_2}>0$

where $x(u)$ is the quantile function of x, i.e., $F(x(u))=u$, $0<u<1$. Unbiased estimators of sample L-moments are commonly written as

$$
\begin{aligned}
l_1 &= \frac{1}{n}\sum_{i=1}^{n} x_{(i)} \\
l_2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2(i-1)}{n-1}-1\right)x_{(i)} \qquad (4)\\
l_3 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{6(i-1)(i-2)}{(n-1)(n-2)}-\frac{6(i-1)}{(n-1)}+1\right)x_{(i)}.
\end{aligned}
$$

Explicit relations for the estimation of distribution parameters are obtained for position-scale(-shape) distributions. In fact, for these distributions the quantile function can be written as

$$x(u)=\theta_1+\theta_2\cdot z(u,\theta_3), \qquad (5)$$

where $z(u,\theta_3)$ is the quantile function of the standardized variable $z=(x-\theta_1)/\theta_2$, which only depends on the probability level u and on the shape parameter θ_3 (for two-parameter distributions of course the dependency on θ_3 is lost). Using Eq. (5) in Eq. (3), distribution L-moments are re-written as

$$
\begin{aligned}
L_1 &= \theta_1+\theta_2\int_0^1 z(u,\theta_3)du=\theta_1+\theta_2 A(\theta_3) \\
L_2 &= \theta_2\int_0^1 z(u,\theta_3)(2u-1)du=\theta_2 B(\theta_3) \qquad (6)\\
L_3 &= \theta_2\int_0^1 z(u,\theta_3)(6u^2-6u+1)du=\theta_2 C(\theta_3),
\end{aligned}
$$

where $A(\theta_3)$, $B(\theta_3)$ and $C(\theta_3)$ are distribution dependent functions (or constant values in case of two-parameter distributions). For example, for the Gumbel distribution one easily obtains $A=\gamma_E$ (where γ_E=0.5772... is the Euler constant) and B=ln(2) (the C value is not required for two-parameter distributions). The values of $A(\theta_3)$, $B(\theta_3)$ and $C(\theta_3)$ for the distributions considered in this paper are reported in Table 2.

Estimators for location, scale, and shape parameters are now obtained by equating Eqs. (4 and 6). Consequently, one obtains the following system of equations:

$$
\begin{cases}
\hat{\theta}_1^{\ell}=l_1-\hat{\theta}_2^{\ell}A\left(\hat{\theta}_3^{\ell}\right) \\
\hat{\theta}_2^{\ell}=l_2/B\left(\hat{\theta}_3^{\ell}\right) \\
\dfrac{C\left(\hat{\theta}_3^{\ell}\right)}{B\left(\hat{\theta}_3^{\ell}\right)}=\dfrac{l_3}{l_2}.
\end{cases} \qquad (7)
$$

These estimators are represented by using the superscript ℓ to denote that they are the classical L-moments estimators, in order to avoid confusion with the modified estimators introduced in the following subsection. The equations in the system can be separately solved by starting from the bottom one, which allows one to find $\hat{\theta}_3^{\ell}$; then this result is used to find $\hat{\theta}_2^{\ell}$ from the central equation; finally $\hat{\theta}_2^{\ell}$ and $\hat{\theta}_3^{\ell}$ are used in the top equation to find $\hat{\theta}_1^{\ell}$. In case of two-parameter distributions, only the first two equations are needed, and the solution is analytically explicit.

Once the parameters have been estimated, the maximum-value test can be applied. This test is specifically targeted to verify if the maximum (or minimum) values in the sample are consistent with the hypothesis that the distribution $F_X(x|\Theta)$ is the real parent distribution. In detail, the testing procedure requires that

$$\left[F_X(x_{(n)}|\hat{\Theta})\right]^n<1-\alpha \qquad (8)$$

where $[F_X(x|\hat{\Theta})]^n$ is the maximum value distribution (as defined, for example, by Kendall and Stuart, 1979 and Kottegoda and Rosso, 1998), $x_{(n)}$ is the n-th order statistic (i.e. the maximum value in the sample) and α is the significance

Table 2. Distribution dependent functions to be used for parameter estimation, as defined in Eqs. (6 and 10). $\Gamma[\cdot]$ is the gamma function, $\Gamma[\cdot;\cdot]$ is the incomplete gamma function, and Norminv is the inverse of the standardized Gaussian cumulative distribution function.

	$A(\theta_3)$	$B(\theta_3)$	$C(\theta_3)$	$D(n,\theta_3)$
EXP	1	0.5	–	$-\ln[1-0.5^{1/n}]$
EV1	γ_e	$\ln[2]$	–	$-\ln[\frac{\ln[2]}{n}]$
NORM	0	$1/\sqrt{\pi}$	–	$\mathrm{Norminv}[0.5^{1/n}]$
GEV	$\frac{\Gamma(1-\theta_3)-1}{\theta_3}$	$(2^{\theta_3}-1)\frac{\Gamma(1-\theta_3)}{\theta_3}$	$(1+3^{\theta_3}-2^{\theta_3})\frac{\Gamma(1-\theta_3)}{\theta_3}$	$\frac{\left(\frac{\ln[2]}{n}\right)^{-\theta_3}-1}{\theta_3}$
GAM	θ_3	$\frac{\Gamma[\theta_3+1/2]}{\Gamma[\theta_3]\sqrt{\pi}}$	$\frac{\Gamma[\theta_3+1/2]}{\Gamma[\theta_3]\sqrt{\pi}}\cdot\tau_3$ *	$0.5^{1/n}=\frac{\Gamma[\theta_3,D(n,\theta_3)]}{\Gamma[\theta_3]}$ **

* τ_3 is the *L*-coefficient of skewness of the distribution. It can be computed using Eqs. (A.86 and A.88) in Hosking and Wallis (1997).
** Equation to be solved numerically.

level of the test. A similar test statistic was originally proposed by Grubbs (1969) as a technique for outliers detection in a Gaussian setting, and subsequently extended to Gumbel-distributed parents by Rossi et al. (1984). However, applications of this test in a non-Gaussian setting either require extensive numerical simulations, or fail to consider the effects of parameter-estimation, i.e. of the substitution of Θ with $\hat{\Theta}$ in Eq. (8). In fact, when parameters are estimated from the same sample used for goodness-of-fit purposes, the limiting values for the goodness-of-fit test (in this case, $1-\alpha$) should be suitably recalculated; this is a general requirement for all goodness-of-fit tests, see D'Agostino and Stephens (1986) and Laio (2004), among others, for details. Here we follow a different approach, as described in the following subsection, which allows one to suitably account for this effect, still maintaining the simplicity and full analytical tractability that are the distinctive features of the maximum value test.

2.2 Modified maximum-value test

In this section a simple and analytically explicit technique is proposed to account for the effects of parameter estimation. Since the maximum-value test is based only on the n-th order statistic $x_{(n)}$, the basic idea is to avoid using $x_{(n)}$ in the parameter estimation, so that the test statistic will turn out to be only slightly dependent on $x_{(n)}$ (as discussed in Laio et al., 2010). It is not possible to simply eliminate $x_{(n)}$ from the sample and carry out parameter estimation on the remaining $n-1$ values, because the resulting parameter estimators would be significantly negatively biased (due to the fact that the larger value in the sample – and not a value taken at random – is excluded from the sample). Therefore we explore the possibility to substitute $x_{(n)}$ with an estimator which is provided by the median of the hypothetical maximum value distribution,

$$\left[F_X(\tilde{x}_{(n)}|\hat{\Theta})\right]^n = 0.5. \qquad (9)$$

The decision of using the median of the hypothetical maximum value distribution as an estimator of $x_{(n)}$ comes from the possibility to obtain analytically explicit results with this estimator. The same was not guaranteed by using the mean of the maximum value distribution as the estimator of $x_{(n)}$. Consequently, by considering Eqs. (9 and 5) one obtains

$$\tilde{x}_{(n)} = \theta_1 + \theta_2 \cdot z\left(0.5^{1/n}, \theta_3\right) = \theta_1 + \theta_2 \cdot D(n,\theta_3), \qquad (10)$$

where $D(n,\theta_3)$ is a distribution dependent function of the sample size n and shape parameter θ_3 (only of n in case of two-parameter distributions). For example, for the Gumbel distribution $D(n)=-\ln[\ln[2]/n]$. The values of $D(n,\theta_3)$ for the other distributions considered in this paper are reported in Table 2. New parameter estimators, weakly dependent on $x_{(n)}$ and therefore amenable for use in Eq. (8), can now be obtained by resorting to the substitution $\tilde{x}_{(n)} \to x_{(n)}$ in Eq. (4), and resolving the L-moments equations to find out the new parameter estimators. More in detail, one can note that $x_{(n)}$ always appears with a weight $1/n$ in the summations of Eq. (4), for L-moments of any order. As a consequence, the substitution $\tilde{x}_{(n)} \to x_{(n)}$ trivially leads to the following modified form of sample L-moments estimators:

$$\begin{aligned}
l_1^* &= \frac{1}{n}\sum_{i=1}^{n-1} x_i + \frac{1}{n}\tilde{x}_{(n)} = l_1' + \frac{1}{n}\tilde{x}_{(n)} \\
l_2^* &= \frac{1}{n}\sum_{i=1}^{n-1}\left(\frac{2(i-1)}{n-1} - 1\right)x_i + \frac{1}{n}\tilde{x}_{(n)} = l_2' + \frac{1}{n}\tilde{x}_{(n)} \\
l_3^* &= \frac{1}{n}\sum_{i=1}^{n-1}\left(\frac{6(i-1)(i-2)}{(n-1)(n-2)} - \frac{6(i-1)}{(n-1)} + 1\right)x_i + \frac{1}{n}\tilde{x}_{(n)} \\
&= l_3' + \frac{1}{n}\tilde{x}_{(n)},
\end{aligned} \qquad (11)$$

where l'_1, l'_2, and l'_3 are the first three sample L-moments calculated by excluding the largest value in the sample, or $l'_1{=}l_1-x_{(n)}/n$, $l'_2{=}l_2-x_{(n)}/n$, and $l'_3{=}l_3-x_{(n)}/n$. By equating the modified sample L-moments in Eq. (11) to the distribution L-moments in Eq. (6) one obtains

$$\begin{cases} \hat{\theta}_1\left(1-\dfrac{1}{n}\right)+\hat{\theta}_2\left(A(\hat{\theta}_3)-\dfrac{D(n,\hat{\theta}_3)}{n}\right)=l'_1 \\[2ex] -\dfrac{1}{n}\hat{\theta}_1+\hat{\theta}_2\left(B(\hat{\theta}_3)-\dfrac{D(n,\hat{\theta}_3)}{n}\right)=l'_2 \\[2ex] -\dfrac{1}{n}\hat{\theta}_1+\hat{\theta}_2\left(C(\hat{\theta}_3)-\dfrac{D(n,\hat{\theta}_3)}{n}\right)=l'_3, \end{cases} \tag{12}$$

where Eq. (10) has also been used. The solution of the system of Eqs. (12) allows one to find out the modified estimators of the position, scale, and shape parameters, $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$. We will denote these estimators as *censored* L-moments estimators of the distribution parameters, because the procedure of substitution of the maximum sample value resembles a Type 2 censoring.

By rearranging the term in Eq. (12) one obtains

$$\frac{B(\hat{\theta}_3)-C(\hat{\theta}_3)}{B(\hat{\theta}_3)\left(1-\frac{1}{n}\right)+\frac{A(\hat{\theta}_3)}{n}-\frac{D(n,\hat{\theta}_3)}{n}}=\frac{l'_2-l'_3}{l'_2\left(1-\frac{1}{n}\right)+\frac{l'_1}{n}}, \tag{13}$$

which can be used to find out the estimator of the shape parameter (by using the distribution-dependent functions defined in Table 2). This $\hat{\theta}_3$ value can then be used in

$$\hat{\theta}_2=\frac{l'_2\left(1-\frac{1}{n}\right)+\frac{1}{n}l'_1}{B(\hat{\theta}_3)\left(1-\frac{1}{n}\right)+\frac{1}{n}A(\hat{\theta}_3)-\frac{D(n,\hat{\theta}_3)}{n}}, \tag{14}$$

to find out the scale parameter estimator, $\hat{\theta}_2$. As usual, for two-parameter distributions Eq. (14) can be directly used, without preliminary application of Eq. (13), because of course no shape parameter exists in this case. For example, for the Gumbel distribution, by using the functions in Table 2 in Eq. (14), one easily finds

$$\hat{\theta}_2=\frac{l'_2\left(1-\frac{1}{n}\right)+\frac{1}{n}l'_1}{\ln[2]\left(1-\frac{1}{n}\right)+\frac{1}{n}\gamma_E-\frac{\ln[\ln[2]/n]}{n}}. \tag{15}$$

Finally, the two estimators $\hat{\theta}_2$ and $\hat{\theta}_3$ can be used to estimate the location parameter through the relation

$$\hat{\theta}_1=l'_1-l'_2-\hat{\theta}_2[A(\hat{\theta}_3)-B(\hat{\theta}_3)], \tag{16}$$

which specifies to

$$\hat{\theta}_1=l'_1-l'_2-\hat{\theta}_2[\gamma_E-\ln[2]] \tag{17}$$

for the Gumbel distribution.

Some comments can be useful at this point to better contextualize the obtained results:

- The censored L-moments estimators (Eqs. 13, 14 and 16) are only weakly dependent on the sample maximum value (see Laio et al., 2010), and for this reason are amenable for use in the modified maximum-value test. In particular, these estimators are asymptotically independent of $x_{(n)}$, as can be inferred by the theory reported in Falk and Reiss (1988).

- For two-parameter distributions the result is analytically explicit, while for three-parameter distributions it requires to numerically solve Eq. (13), which, however, is a rather trivial task with computers. We also note that the level of complexity of censored L-moments estimators is exactly the same as that of the standard L-moments estimators, that again require the numerical solution of the third of Eq. (7) to perform parameter estimation for three-parameter distributions.

- For $n\to\infty$ the censored L-moments estimators converge to the standard L-moments estimators, $\hat{\theta}_1\to\hat{\theta}_1^{\ell}$, $\hat{\theta}_2\to\hat{\theta}_2^{\ell}$, and $\hat{\theta}_3\to\hat{\theta}_3^{\ell}$. This can be easily verified by considering that for $n\to\infty$ one recovers the third of Eq. (7) from Eq. (13), the second of Eq. (7) from Eq. (14), and the first of Eq. (7) from Eq. (16). Therefore, asymptotically one finds $l'_1\to l_1$, $l'_2\to l_2$, and $l'_3\to l_3$.

- In a rather different context (trying to compensate for rainfall outliers in short time series) Hershfield (1961, 1965) developed a partially similar procedure to account for the effect of maximum value elimination on parameter estimation. However, his results are limited to the first two moments (the mean and the variance) and they are not valid for any distribution as the ones we present (because they were obtained from real rainfall data). Moreover, these results were provided only in a graphical form, even if Koutsoyiannis (2000, p. 22) (in greek) provides them in closed analytical form.

- Sometimes systematic records of data can be integrated with additional data, derived from measurements of significant occasional events (e.g., Frances, 1998). When a number of occasional additional measurements is available, one can merge them with the systematic ones (e.g., Bayliss and Reed, 2001), producing a new time series of “equivalent” length m, where m is the number of years between the first and the last measurement of both the systematic and the occasional record, merged together. The MV test can be easily applied also to these merged samples of systematic and non-systematic data, by simply substituting m for n in Eq. (8), and by using Wang (1990) estimators of sample L-moments instead of the standard estimators in Eq. (4). The possibility to be applied when non-systematic data are present is a unique feature of the MV test, not shared by other commonly used testing techniques.

3 Assessment of the power of the test through synthetic data

In the previous section we have outlined the structure of the modified maximum-value test, showing how the testing procedure should be applied, i.e. by introducing Eqs. (13, 14 and 16) into Eq. (8), given a candidate operational model $F_X(x)$ and a significance level of the test α. In this section we compare the power of the MMV test to the performances of other goodness-of-fit tests and outlier detection procedures. To this aim the parent distribution, $G_X(x)$, from which synthetic samples of independent observations will be generated, is supposed to be known. The null hypothesis to be tested is $H_0 : G_X(x){=}F_X(x)$ with the Gumbel distribution as hypothetical distribution, i.e. $F_X(x){=}$ EV1 (see notation in Table 1). The power of the tests (i.e. the ability of the test to recognize that H_0 is false) is analyzed under different parametrization of $G_X(x)$. Particular attention is payed to the behavior of the tests when dealing with small samples. The cases $G_X(x)$=GEV and of $G_X(x)$=TCEV (Rossi et al., 1984) are treated in Sects. 3.1 and 3.2, respectively. The GEV and TCEV distributions are used here as possible parents due to their widespread use in the frequency analysis of the hydrological extremes (see e.g. papers citing Hosking et al., 1985 and Rossi et al., 1984). The benchmark tests, in both cases, are the classical Pearson test and the Anderson-Darling test (referred to, respectively, as CHI and AD test in the following of the paper), plus a specific test for outliers detection (Fung and Paul, 1985), called Fung-Paul (FP) test in the following of the paper.

The classical Pearson test falls in the category of the chi-squared type tests. The testing procedure requires that the range of x is partitioned in classes; a convenient procedure to avoid arbitrariness and maximize the power of the test entails the choice of k equiprobable classes under the hypothesized distribution, with $k{=}2n^{0.4}$ (Moore, 1986). The test statistic for the case 0 (i.e., when the parameters of $F_X(x)$ are fully specified a priori) is the chi-squared distribution with $k{-}1$ degrees of freedom. Conversely, when the distribution $F_X(x)$ is not completely known there is a partial recovery of degrees of freedom of the chi-squared distribution with respect to the commonly recommended value of $k-1$, with the consequence that the asymptotic distribution will lie somewhere between a chi-square distribution with with $k{-}p{-}1$ and $k{-}1$ degrees of freedom (e.g., Kendall and Stuart, 1979).

The Anderson-Darling test is based on the comparison between the hypothetical, $F_X(x)$, and empirical distribution function, $F_n(x)$, i.e. a cumulative probability distribution function that concentrates a probability $1/n$ at each of the n values in the sample. The discrepancy between the two distributions can be measured with quadratic statistics of the form $\int_x [F_n(x){-}F_X(x)]^2\Psi(x)dx$, where $\Psi(x)$ is a weighting function. When $\Psi(x){=}[F_X(x)(1{-}F_X(x))]^{-1}$ one obtains the Anderson-Darling statistic, called A^2, which has the property of assigning more weight to the tails of the distribution than to the central part. Critical values and percentage points for the AD test for EV1, NORM, GAM and GEV distributions can be calculated following the procedure described by Laio (2004).

The outlier detection procedure proposed by Fung and Paul (1985) is intended for testing the discordancy of one or more outliers in a Gumbel sample. The test statistic is expressed as $T{=}(x_{(n)}{-}x_{(n-k)})/(x_{(n)}{-}x_{(1)})$, where $k{=}1,2,3$ is the number of outliers. The tabulated significance levels can be found also in Barnett and Lewis (1994), abridged from Fung and Paul (1985). The test is recommended for sample sizes in the range $n=[5{\div}20]$.

3.1 Gumbel vs. GEV

The power of the MMV test to recognize as non-Gumbel a GEV-distributed sample is evaluated hereinafter. In rigorous terms this corresponds to assuming as parent a GEV distribution, while the distribution to be tested is an EV1. In detail, 10 000 samples formed by n elements (with $n{=}[20\div 50]$) are generated from a GEV distribution with fixed parameters $\theta_1{=}1$, $\theta_2{=}1$, and variable θ_3 values. Note that positive θ_3 values correspond to positive skewness of the distribution; when $\theta_3{=}0$ the GEV reduces to an EV1 distribution. For each sample, the parameters of an EV1 distribution are estimated according to Eqs. (17 and 15), which introduced in Eq. (10) give $\tilde{x}_{(n)}$. The MV test statistic, as expressed in Eq. (8), can then be resolved by resorting to the substitution $\tilde{x}_{(n)} \rightarrow x_{(n)}$.

The results are reported in Fig. 1, by comparison with the FP, CHI and AD test performances, at the 10% significance level. Also the case of the SMV test (i.e., applied with the classical L-moments estimators, as in Eq. 7) is shown. Observe that the power of the different tests converges to the significance level for null values of the shape parameter, i.e. when the parent distribution collapses on a Gumbel. The left and right-hand side graphs are referred to the case of n=20 and n=50, respectively. One can observe that, for rather small samples (e.g., n=20), which is a very common situation in hydrological applications, the MMV test performs better than the SMV test (which even fails to converge to the significance level when θ_3=0). Also, the CHI test results to be by far less effective. By comparison to the FP test, the MMV test turns out to be slightly more powerful. As for the AD test, the performances are comparable with a slight prevalence of the MMV test when the parent distribution is similar to the distribution considered in the null hypothesis (i.e., when $\theta_3{\rightarrow}0$). The behavior is again similar, but more favorable to the AD test, for larger samples (as shown in the right panel of Fig. 1). The intersection between the MMV and AD curves, in fact, is shifted to the right, with smaller differences between the test performances. This is due to the lesser influence of the maximum value in presence of larger samples, and is indicative of the MMV test being more suited for application to small samples. Note that the comparison

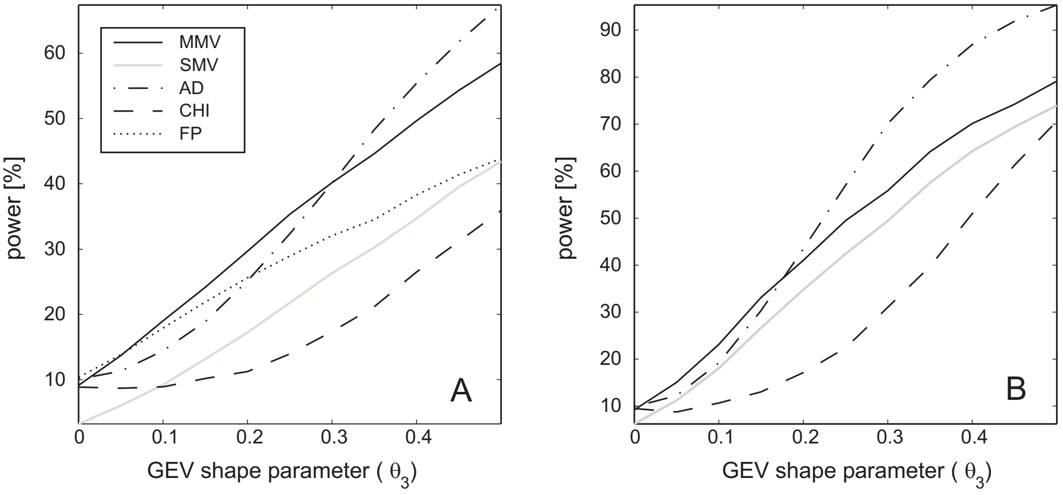

Fig. 1. Power of the tests considered in this paper (MMV = modified maximum value; SMV = standard maximum value; AD = Anderson-Darling; CHI = chi-squared; FP = Fung-Paul) as a function of the shape parameter θ_3 of the parent GEV distribution. The hypothetical distribution is Gumbel, the sample size is n=20 in panel **(A)** and n=50 in panel **(B)**.

with the FP test is not present on the right-hand panel because Fung and Paul (1985) did not give the critical values for n>20.

A possible explanation for the larger power of the MMV test when θ_3 is close to zero is provided in Fig. 2. The design event x_T is plotted as a function of return period T (i.e., Eq. (5) is used with u=$1-1/T$) for a GEV and a Gumbel distribution sharing the first two L-moments. When θ_3=0.15 (Fig. 2a) the two distributions are substantially overlapped up to a 20-year return period, and diverge only in the upper tail. In contrast, when θ_3=0.45 the two distributions are rather different also for low return periods (Fig. 2b). It is clear that the AD test, which is based on the comparison of the distributions in the whole probability range, will be favored in the latter situation, while the MMV test will perform better in situations like the one depicted in Fig. 2a, where the differences concentrate in the upper tail of the distributions. We note in passing that the ability of the MMV test to falsify the null hypothesis in these cases may be very important in practical applications: for example, in the case of Fig. 2a, the wrong assumption of a Gumbel distribution would lead to a 30% underestimation of the 100-year design value.

3.2 Gumbel vs. TCEV

The MMV test and AD test are compared also for the case when the parent distribution is a TCEV distribution (Rossi et al., 1984; Fiorentino et al., 1985) that is usually expressed in the form

$$G_X(x) = \exp\left[-\exp\left[-\frac{x-\theta_1}{\theta_2}\right] - \theta_3 \cdot \exp\left[-\frac{1}{\theta_4}\frac{x-\theta_1}{\theta_2}\right]\right], \qquad (18)$$

where θ_1 is the position parameter, θ_2 is the scale parameter, θ_3 and θ_4 are two shape parameters. When θ_3=0 the TCEV distribution reduces to a Gumbel distribution. A significance level α of 10% is again assumed for the tests. The CHI, FP and SMV tests are not considered in this example to

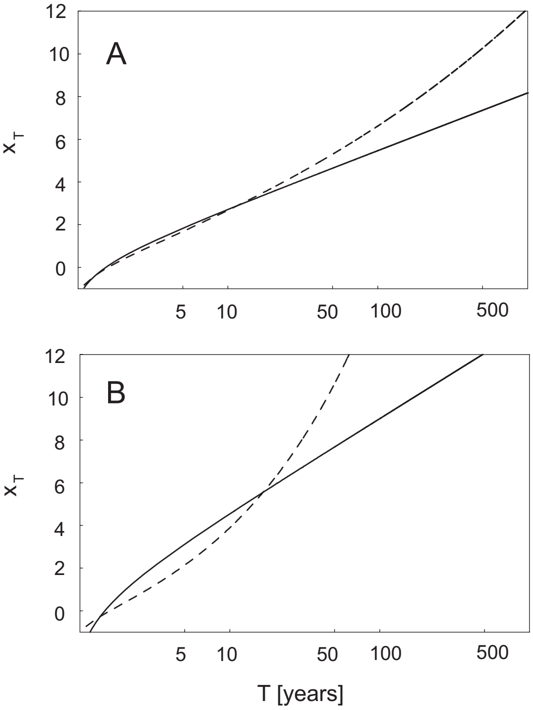

Fig. 2. Comparison of a GEV(dashed line) and a Gumbel (solid line) distribution function sharing the first two L-moments. The design event x_T is plotted as a function of return period T (in logarithmic scale). The parameter values for the GEV are θ_1=0, θ_2=1, θ_3=0.15 in panel **(A)** and θ_1=0, θ_2=1, θ_3=0.45 in panel **(B)**. The parameters of the Gumbel distribution are found by matching the first two distribution L-moments to those of the GEV distribution.

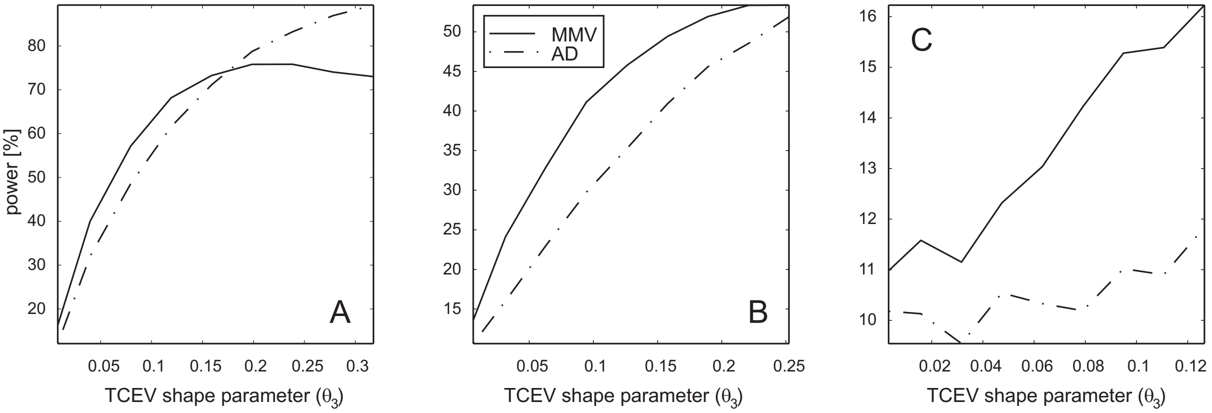

Fig. 3. Power of the modified maximum-value (MMV) and Anderson-Darling (AD) tests for a TCEV parent distribution. In panel **(A)** θ_4=10, in panel **(B)** θ_4=5, and in panel **(C)** θ_4=2. The sample size is n=20.

facilitate the comparison between the two tests (i.e., MMV and AD) that performed better with a GEV parent distribution. The results are shown in Fig. 3, where the parameters of the "basic component" of a TCEV (to use the notation by Rossi et al., 1984 and Fiorentino et al., 1985) are kept constant (θ_1=ln(10), θ_2=1) while the parameters of the "outlying component" are allowed to vary. The three panels refer to different values of the parameter θ_4, while the values of θ_3 vary continuously on the x-axis. Only the case of n=20 is examined. In this case the MMV test is found to be more powerful than the AD test for high θ_4 values and low θ_3 values; while the performances of both tests are poor for small θ_4 values (panel C). Conversely, the AD test is more powerful for high θ_3 values; in fact, the intersection of the two curves occurs for $\theta_3 \simeq 0.2$ (panel A).

A similar behavior is therefore found with the two different parent distributions (GEV and TCEV): the MV test performs better than the AD test when the parent and hypothetical distributions are rather similar. A possible explanation of this behavior is the following: when the discrepancies between the parent and hypothetical distribution are large, they significantly affect also the central part of the distribution, therefore increasing the discerning ability of standard testing techniques, as the AD test. In contrast, when the parent and hypothetical distribution are rather similar, one may spot the difference only by looking at the tails of the distribution, i.e., for example, at the maximum observed value.

4 Conclusions

Outliers in hydrological samples are often seen by the modelers as disturbing elements, because their very presence challenges well-established and convenient practices, by contributing to raise doubts on the correctness of the hypothesized probability distribution model (for example, the adoption of the Gumbel distribution to represent flood or rainfall annual maxima). In this paper we follow the principle that, at least in some cases, it is exactly in the "outlying" data that resides very important information for the validation of the statistical model to be used in the frequency analysis of extreme events. More in detail, we have described a procedure to perform a goodness-of-fit test based solely on the maximum recorded value in the sample. We have shown that the maximum value test, if correctly applied, represents a simple, analytically explicit, alternative to other commonly used goodness-of-fit tests: this test performs consistently better than the Chi-squared test, and it proves to be more powerful than the Anderson-Darling test to recognize the lack of fit when the parent distribution is similar to the distribution being tested. Since the test is based only on the maximum recorded value, it is particularly suited when small samples (e.g., $n \leq 20 \div 25$) are available. In the cases when the test hypothesis is falsified and no other data (systematic or non-systematic) can be added to the sample, our recommendation is to resort to alternative solutions as, for instance, regional analysis. Possible future developments of this work are in the direction of further investigations of the problem of the residual dependence of the censored L-moments estimators on $x_{(n)}$ and application of the test to the case when non systematic data are included in the sample.

Acknowledgements. The work was funded by the Italian Ministry of Education (grants 2007HBTS85 and 2008KXN4K8). We are grateful to D. Koutsoyiannis for providing the reference to the works by Hershfield (1961, 1965) and for pointing us to his teaching material at http://www.itia.ntua.gr/en/docinfo/116/. Comments and suggestions from two Anonymous Reviewers are gratefully acknowledged.

Edited by: S. Grimaldi

References

Ahmad, M., Sinclair, C., and Spurr, B.: Assessment of flood frequency models using empirical distribution function statistics, Water Resour. Res., 24, 1323-1328, 1988.

Barnett, V. and Lewis, T.: Outliers in statistical data, Springer Series in Statistics, John Wiley and Sons, 1994.

Bayliss, A. and Reed, D.: The use of historical data in flood frequency estimation, Tech. rep., Centre for Ecology and Hydrology, 2001.

Bryson, M.: Heavy-tailed distributions: properties and tests, Technometrics, 16, 61–68, 1974.

Chowdhury, J., Stedinger, J., and Lu, L.: Goodness-of-fit tests for regional generalized extreme value flood distributions, Water Resour. Res., 27, 1765–1776, 1991.

D'Agostino, R. and Stephens, M.: Goodness-of-Fit Techniques, Marcel Dekker Inc, New York, 1986.

Di Baldassarre, G., Laio, F., and Montanari, A.: Design flood estimation using model selection criteria, Phys. Chem. Earth, 34(10–12), 606–611, doi:10.1016/j.pce.2008.10.066, 2008.

Falk, M. and Reiss, R.: Independence of Order Statistics, Annals of Probability, 16, 854–862, 1988.

Fill, H. and Stedinger, J.: L-moment and probability plot correlation coefficient goodness-of-fit tests for the Gumbel distribution and impact of autocorrelation, Water Resour. Res., 31, 225–229, 1995.

Fiorentino, M., Versace, P., and Rossi, F.: Regional flood frequency estimation using the two-component extreme value distribution, Hydrolog. Sci. J., 30, 51–63, 1985.

Frances, F.: Using the TCEV distribution function with systematic and non-systematic data in a regional flood frequency analysis, Stoch. Hydrol. Hydraul., 12, 267–283, 1998.

Fung, K. and Paul, S.: Comparison of outlier detection procedures in Weibull or Extreme-Value distribution, Commun. Statist. Simula. Computa, 14, 895–917, 1985.

Grubbs, F.: Procedures for detecting outlying observations in samples, Technometrics, 11, 1–21, 1969.

Gumbel, E.: Discussion of the Papers of Messrs. Anscombe and Daniel, Technometrics, 2, 165–166, 1960.

Hershfield, D.: Estimating the probable maximum precipitation, J. Hydraul. Div. ASCE, 87(HY5), 99–106, 1961.

Hershfield, D.: Method for estimating probable maximum precipitation, J. Am. Water Works Assoc., 57, 965–972, 1965.

Hosking, J. and Wallis, J.: Regional Frequency Analysis: An Approach Based on L-Moments, Cambridge University Press, 1997.

Hosking, J., Wallis, J., and Wood, E.: Estimation of the Generalized Extreme Value distribution by the method of the probability weighted moments, Technometrics, 27, 251–261, 1985.

Kendall, M. and Stuart, A.: The Advanced Theory of Statistics, Charles Griffin and Company Limited, 1979.

Kottegoda, N. and Rosso, R.: Statistics, probability, and reliability for civil and environmental engineers, McGraw-Hill, International Edition, 1998.

Koutsoyiannis, D.: Probable maximum precipitation, http://www.itia.ntua.gr/getfile/116/5/documents/2000HydrometPMP.pdf, 2000.

Laio, F.: Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters, Water Resour. Res., 40, W09308, doi:10.1029/2004WR003204, 2004.

Laio, F., Di Baldassarre, G., and Montanari, A.: Model selection techniques for the frequency analysis of hydrological extremes, Water Resour. Res., 45, W07416, doi:10.1029/2007WR006666, 2009.

Laio, F., Allamano, P., and Claps, P.: Interactive comment on "Exploiting the information content of hydrological "outliers" for goodness-of-fit testing" by F. Laio et al., Hydrol. Earth Syst. Sci. Discuss., 7, C2227–C2230, 2010.

Mitosek, H., Strupczewski, W., and Singh, V.: Three procedures for selection of annual flood peak distribution, J. Hydrol., 323(1–4), 57–73, 2006.

Moore, D.: Goodness-of-Fit Techniques, chap. Tests of the chi-squared type, Marcel Dekker, New York, 1986.

Rossi, F., Fiorentino, M., and Versace, P.: Two-component extreme value distribution for flood frequency analysis, Water Resour. Res., 20, 847–856, 1984.

Stedinger, J., Vogel, R., and Foufoula-Georgiou, E.: Handbook of Hydrology, chap. 8: Frequency analysis of extreme events, McGraw-Hill, New York, 1992.

Strupczewski, W., Singh, V., and Weglarczyk, S.: Asymptotic bias of estimation methods caused by the assumption of false probability distributions, J. Hydrol., 258, 122–148, 2002.

Vogel, R.: The probability plot correlation coefficient test for the normal, lognormal, and Gumbel distributional hypotheses, Water Resour. Res., 22, 587–590, 1986.

Vogel, R. and McMartin, D.: Probability plot goodness-of-fit and skewness estimation procedures for the Pearson type 3 distribution, Water Resour. Res., 27, 3149–3158, 1991.

Wang, Q.: Unbiased estimation of probability weighted moments and partial probability weighted moments from systematic and historical flood information and their application to estimating the GEV distribution, J. Hydrol., 120, 115–124, 1990.

Wang, Q.: Approximate goodness-of-fit tests of fitted generalized extreme value distributions using LH moments, Water Resour. Res., 34, 3497–3502, 1998.