Model selection techniques for the frequency analysis of hydrological extremes

Francesco Laio,[1] Giuliano Di Baldassarre,[2,3] and Alberto Montanari[4]

Received 15 November 2007; revised 15 December 2008; accepted 7 May 2009; published 18 July 2009.

[1] The frequency analysis of hydrological extremes requires fitting a probability distribution to the observed data to suitably represent the frequency of occurrence of rare events. The choice of the model to be used for statistical inference is often based on subjective criteria, or it is considered a matter of probabilistic hypotheses testing. In contrast, specific tools for model selection, like the well-known Akaike information criterion (AIC) and the Bayesian information criterion (BIC), are seldom used in hydrological applications. The objective of this study is to verify whether the AIC and BIC work correctly when they are applied to identifying the probability distribution of hydrological extremes, i.e., when the available samples are small and the parent distribution is highly asymmetric. An additional model selection criterion, based on the Anderson-Darling goodness-of-fit test statistic, is here proposed, and the performances of the three methods are compared through an extensive numerical analysis. The capability of the three criteria to recognize the correct parent distribution from the available data samples varies from case to case, and it is rather good in some cases (in particular when the parent is a two-parameter distribution) and unsatisfactory in others. An application to flood peak time series from 1000 catchments located in the United Kingdom provides some further information on the qualities and drawbacks of the considered criteria. From the numerical simulations and data-based analyses it can be concluded that the three model selection techniques considered here produce results of comparable quality.

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

1. Introduction

[2] The frequency analysis of hydrological extremes aims at estimating the design event, i.e., the discharge (or precipitation, or wind velocity, etc.) value corresponding to a given return period [e.g., *Stedinger et al.*, 1992]. Schematically, frequency analysis at gauged stations reduces to the choice of a probabilistic model (i.e., a probability distribution) and to the estimation of its parameters. Many probability distributions have been considered, in different situations, for the probabilistic modeling of extreme events. Examples include the extreme value distributions (Generalized Extreme Value (GEV), Gumbel and Frechet distributions), the distributions of the transformed normal or gamma families, and many others less frequently used [e.g., *Kottegoda and Rosso*, 1997].

[3] While several books and articles in scientific journals deal with the parameter estimation for probabilistic models, not as much has been published in the hydrologic literature about probabilistic model selection. The choice of a model to be used for statistical inference is often based on subjective criteria, or it is considered a matter of probabilistic hypotheses testing. Within the frequency analysis of hydrological variables, for example, the hypothetical probability distributions undergo some goodness-of-fit tests: Those accepted with a prefixed level of significance are retained [e.g., *Stedinger et al.*, 1992], and the others are rejected as not being suitable for use in design value estimation. This procedure has some evident limitations, because the obtained results are subjective, depending, for example, on the significance level chosen, and ambiguous, as often more than one distribution pass the goodness-of-fit tests [see *Burnham and Anderson*, 2002, pp. 35–36]. Ambiguity can be resolved within statistical testing only when two candidate (nested) probability distributions are considered. In this case the simpler model (the one with fewer parameters) is a special case of the more complicated model (the Gumbel and GEV distribution are an example of nested models), and therefore the test is able to univocally select one of the two models if, as usual, the null hypothesis is taken that the parent is the simpler model, and the alternative hypothesis is taken that the parent is the more complicated model.

[4] One may note that ambiguousness in statistical testing can be resolved by identifying the probability distribution that passes an assigned null hypothesis test with the highest significance level [e.g., *Stephens*, 1986, p. 120]. However, this kind of methodology goes beyond standard statistical

[1] Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, Turin, Italy.

[2] School of Geographical Sciences, University of Bristol, Bristol, UK.

[3] Now at Department of Hydroinformatics and Knowledge Management, UNESCO-IHE Institute for Water Education, Delft, Netherlands.

[4] Dipartimento di Ingegneria delle Strutture, dei Trasporti, delle Acque, del Rilevamento, del Territorio, Università degli Studi di Bologna, Bologna, Italy.

0043-1397/09/2007WR006666

testing, which requires explicitly formulating a null hypothesis and fixing a significance level, and it has been sometimes criticized for lack of theoretical bases [e.g., *Lindsey*, 1999; *Burnham and Anderson*, 2002, p. 84]. We argue that this approach falls under the umbrella of model selection techniques and therefore needs to be properly tested (see, for example, the use of Anderson-Darling criterion in the present paper). Model selection can therefore be understood as an alternative (or complementary) procedure with respect to statistical testing. Its main advantages are objectivity (no significance level needs to be defined) and unambiguousness (a single distribution is selected). It is not guaranteed that the distribution identified through model selection passes a given statistical test for an assigned confidence level. In the case the user needs to check that the identified probability distribution satisfies a statistical test, the latter can be applied subsequently to model selection, with the additional advantage that only one distribution needs to be checked.

[5] The basic idea behind this paper is to define some objective criteria for model selection, i.e., for identifying the distribution closer to the one that is supposed to have generated the available data (the so-called parent distribution). Naturally, one does not necessarily believe that the data were actually generated in this way: The model is a convenient representation of a more complex phenomenon. Still, finding two- to three-parameter distributions that provide a satisfactory fit of the frequency of occurrence of the observed events is an essential step toward design value estimation. Objective procedures for probabilistic model selection can be found in the specific literature. The subject was first proposed by *Akaike* [1973], who introduced the principle of maximum entropy as the theoretical basis for model selection, and by *Schwarz* [1978], who, by developing a similar idea in a Bayesian context, proposed the Bayesian information criterion for model selection. Extensions of these methods include corrections to be used with small sample size [*Hurvich and Tsai*, 1989] and generalizations of the mentioned criteria [e.g., see *Bozdogan*, 1987; *Konishi and Kitagawa*, 1996; *Wasserman*, 2000]. Only with the recent increase of computer capabilities have other methods been proposed and developed for the nonasymptotic model selection, based on bootstrap [*Chung et al.*, 1996] or on cross-validation [*Browne*, 2000] techniques. Detailed descriptions of these model selection criteria can be found in the monographs by *Linhart and Zucchini* [1986] and *Burnham and Anderson* [2002] and in the review article by *Zucchini* [2000]. All model selection criteria implicitly use some notion of the principle of parsimony [*Box and Jenkins*, 1970] that describes the conceptual trade-off between bias and variance, i.e., the fact that the bias of the estimates decreases and their variance increases as the number of model parameters increases.

[6] Applications of model selection criteria within the field of the frequency analysis of hydrological extremes are rare and nonmethodical: In the works by *Turkman* [1985], *Mutua* [1994], *Hache et al.* [1999], *Strupczewski et al.* [2001, 2002] and *Cahill* [2003] the Akaike information criterion is applied to single case studies without, however, analyzing the properties of such criterion for small samples or comparing its selection efficiency with that of other criteria. *Mitosek et al.* [2006] applied three model discrimination procedures in order to find the best fitting distribution. Nevertheless, the study considered as alternative model only two-parameter distributions [*Mitosek et al.*, 2006]. A simple approach to the choice 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. The major problems with this approach are (1) the subjectivity inherent in the visual verification of the alignment of the empirical points, and (2) the fact that the method is available only for two-parameter distributions. Another approach occasionally used to discriminate among competing models is based on the use of the likelihood ratio test [e.g., *Strupczewski et al.*, 2006], which, however, can be applied when only two competing models are considered. Moreover, standard applications require that the models are nested, in which case the distribution of the test statistic under the null hypothesis is known (chi-square distribution).

[7] Perhaps the most common approach to the choice of the probabilistic model in hydrology is based on the use of L-moments plots, which are used to determine the probability distribution "closer" to the available sample of data [*Vogel et al.*, 1993a, 1993b; *Hosking and Wallis*, 1997, pp. 73–86]. Suitable L-moments plots are used to discern among two-parameter distributions [e.g., *Di Baldassarre et al.*, 2006a] or among three-parameter distributions [e.g., *Onoz and Bayazit*, 1995; *Di Baldassarre et al.*, 2006b]. This approach, however, is not fully objective as well, because the goodness of fit of a distribution to the data is often based only upon graphic judgment. This limitation may be overcome by using a criterion based on the distance between the real and theoretical L-kurtosis coefficient [*Pandey et al.*, 2001]. Also, *Kroll and Vogel* [2002] developed a performance measure for avoiding difficulties with the visual interpretation of the L-moment diagrams. In particular, they proposed the use of the average weighted orthogonal distance between sample and distribution L-moment ratios. However, comparing the descriptive ability of distributions with different number of parameters still remains an open question. For example, two- and three-parameter distributions are compared separately by *Kroll and Vogel* [2002].

[8] This study aims at verifying whether some objective model selection techniques proposed in the past few decades work correctly when applied to identify the probability distribution of extreme events. A comparison of various techniques is carried on, in order to verify which technique is more efficient in the case of small sample sizes or highly asymmetric distributions, which are typical conditions encountered within the frequency analysis of hydrological extremes. In particular, an extensive numerical analysis is performed to investigate the efficiency of the model selection criteria to select the real parent distribution. Additionally, the study investigates the capability to select the best operational model in terms of reliability of the design flood estimation.

2. Model Selection Criteria

[9] The problem of model selection can be formalized as follows: A sample of n data, $D = \{x_1, \ldots, x_n\}$, arranged in ascending order is available, sampled from an unknown parent distribution $f(x)$; N_m operating models, M_j, $j = 1, \ldots, N_m$, are used to represent the data. The operating models are in the form of probability distributions, $M_j = g_j(x, \hat{\vartheta})$, with

parameters $\hat{\vartheta}$ estimated from the available data sample D. The scope of model selection is to identify the model M_{opt} that is better suited to represent the data. The model selection techniques differ essentially for the metric $\Delta[M_j, f(x)]$, which is used to measure the discrepancy between the hypothetical and the parent distribution, with the additional difficulty that in practical applications the parent distribution is unknown, which implies that the discrepancy cannot be measured but it has to be estimated [*Linhart and Zucchini*, 1986].

[10] We consider three different model selection criteria, namely, the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the Anderson-Darling criterion (ADC). The basic operational characteristics of these criteria are described in the following, while some considerations regarding their theoretical bases and the discrepancy measure upon which they are based are reported in Appendix A (AIC), Appendix B (BIC), and Appendix C (ADC). Of the three methods, the first two belong to the category of classical literature approaches, while the third derives from a heuristic interpretation of the results of a standard goodness-of-fit test and is here proposed for the first time as a model selection criterion.

2.1. Akaike Information Criterion

[11] The Akaike information criterion [*Akaike*, 1973] is based on the use of Kullback-Leibler's information as the discrepancy measure between the true model $f(x)$ and the approximating model, $M_j = g_j(x, \vartheta)$; see Appendix A. The AIC for the jth operational model can be computed as

$$AIC_j = -2\ln\left(L_j\left(\hat{\vartheta}\right)\right) + 2p_j, \tag{1}$$

where $L_j(\hat{\vartheta}) = \prod_{i=1}^{n} g_j(x_i, \hat{\vartheta})$ is the likelihood function, evaluated at the point $\vartheta = \hat{\vartheta}$, corresponding to the maximum likelihood estimator of the parameter vector ϑ [*Linhart and Zucchini*, 1986], and p_j is the number of estimated parameters of the jth operational model. In practice, after the computation of the AIC_j, for all of the operating models, one selects the model with the minimum AIC value, $AIC_{\min}$. By analyzing equation (1), one can see that the first term on the right-hand side tends to decrease as more parameters are added to the approximating model, while the second term tends to increase. This is the trade-off between bias and variance that is the essence behind the principle of parsimony [*Box and Jenkins*, 1970].

[12] When the sample size n is small, with respect to the number of estimated parameters p, the AIC may perform inadequately [*Sugiura*, 1978]. Therefore *Sugiura* [1978] derived a second-order variant of AIC, called AICc:

$$AICc_j = -2\ln\left(L_j\left(\hat{\vartheta}\right)\right) + 2p_j\left(\frac{n}{n-p_j-1}\right). \tag{2}$$

Indicatively, *Burnham and Anderson* [2002] recommend to use AICc when $n/p < 40$. The performances of both AIC and AICc are checked in section 3.

2.2. Bayesian Information Criterion

[13] The Bayesian information criterion was proposed by [*Schwarz*, 1978] Bayesian framework (see Appendix B). The BIC for the jth operational model reads

$$BIC_j = -2\ln\left(L_j\left(\hat{\vartheta}\right)\right) + \ln(n)p_j. \tag{3}$$

In practical application, after the computation of the BIC_j, for all of the operating models one selects the model with the minimum BIC value, $BIC_{\min}$. Despite the completely different manner in which BIC is obtained (see Appendix B), it turns out that the final form of this criterion is rather similar to that of Akaike information criterion; see equation (1). In this case, however, the penalty term due to the number of parameters p_j in the model is multiplied by a factor $0.5\ln(n)$ with respect to the AIC method. As a consequence, the BIC leans more than the AIC toward lower-dimensional models when there are at least eight available observations (i.e., for $n \geq 8$).

[14] Several attempts to extend and generalize the BIC have been made in the literature [e.g., *Wasserman*, 2000; *Konishi and Kitagawa*, 1996], but none of these seems particularly attractive when dealing with small samples and highly asymmetrical distributions, which is the usual case in hydrological applications.

2.3. Anderson-Darling Criterion

[15] The AIC and BIC are standard model selection techniques, commonly used in many different fields; however, none of these methods was explicitly designed to deal with the small sample sizes and highly asymmetric distributions that are commonly encountered in hydrological application. For this reason we propose here another model selection criterion, which is based on the use of the Anderson-Darling test statistic. The Anderson-Darling test has demonstrated good skills when applied to hydrological samples [e.g., *Onoz and Bayazit*, 1995; *Laio*, 2004; *Viglione et al.*, 2007], and our aim in the present work is to verify if these positive results also apply when the statistic is used for model selection purposes. The Anderson-Darling criterion has the form (see Appendix C)

$$ADC_j = 0.0403 + 0.116\left(\frac{\Delta_{AD,j} - \xi_j}{\beta_j}\right)^{\frac{\eta_j}{0.861}} \quad \text{if } 1.2\xi_j \leq \Delta_{AD,j} \tag{4a}$$

$$ADC_j = \left[0.0403 + 0.116\left(\frac{0.2\xi_j}{\beta_j}\right)^{\frac{\eta_j}{0.861}}\right] \cdot \frac{\Delta_{AD,j} - 0.2\xi_j}{\xi_j} \quad \text{if } 1.2\xi_j > \Delta_{AD,j}, \tag{4b}$$

where $\Delta_{AD,j} = \Delta_{AD}[g_j(x, \hat{\vartheta}), f_n(x)]$ is the discrepancy measure characterizing the criterion, evaluated in practice with equation (C3); and ξ_j, β_j, and η_j are distribution-dependent coefficients that are tabled by *Laio* [2004, Tables 3 and 5] for a set of seven distributions commonly used for the frequency analysis of extreme events. In practice, after the computation of the ADC_j, for all of the operating models one selects the model with the minimum ADC value, $ADC_{\min}$. The Anderson-Darling criterion is

Table 1. Probability Models Considered in This Study[a]

Distribution	Acronym (Parameters)	Cumulative Distribution Function or Probability Distribution Function	Range
Gumbel or Extreme Value type I	EV1 (ϑ_1, ϑ_2)	$G(x, \vartheta) = \exp[-\exp(-(x - \vartheta_1)/\vartheta_2)]$	$-\infty < x < +\infty$
Normal or Gaussian	NORM (ϑ_1, ϑ_2)	$g(x, \vartheta) = (1/\sqrt{2\pi}\vartheta_2) \exp[-1/2((x - \vartheta_1)/\vartheta_2)^2]$	$-\infty < x < +\infty$
Generalized Extreme Value	GEV (ϑ_1, ϑ_2, ϑ_3)	$G(x, \vartheta) = \exp[-(1 - (\vartheta_3 (x - \vartheta_1))/\vartheta_2)^{1/\vartheta_3}]$	$(\vartheta_3(x - \vartheta_1))/\vartheta_2 < 1$
Gamma or Pearson type III	GAM (ϑ_1, ϑ_2, ϑ_3)	$g(x, \vartheta) = [1/(\lvert\vartheta_2\rvert\Gamma(\vartheta_3))]((x - \vartheta_1)/\vartheta_2)^{\vartheta_3 - 1} \exp(-[(x - \vartheta_1)/\vartheta_2])$	$(x - \vartheta_1)/\vartheta_2 > 0$

[a] The EV2, LN, and LP3 distributions are obtained as log-transforms of the EV1, NORM, and GAM distributions, respectively.

equivalent to calculating the p-values of the corresponding goodness-of-fit test and selecting the distribution that provides the best fit (maximum p-value). The principle of parsimony (i.e., the fact that more parameterized models are not favored) is in this case preserved by the fact that the coefficients ξ_j, β_j, and η_j in equation (4) depend on the considered distribution (i.e., of course, also on the number of parameters of the distribution). In particular, it is noticed from *Laio* [2004, Table 3] that the coefficient β_j is smaller for more parameterized distributions, which implies that the models with fewer parameters are favored in the application of the criterion (principle of parsimony).

3. Comparison of Model Selection Criteria

[16] The comparison of the model selection criteria described in section 2 is carried out through an extensive numerical analysis aimed at checking the performances of the considered methods when dealing with small sample sizes and highly asymmetric distributions.

[17] The analysis is performed by means of Monte Carlo simulations by using as operational models M_j a total of seven probability models commonly used in the frequency analysis of extreme events: Four of these models (Gumbel or Extreme Value 1 (EV1) distribution, Normal (NORM) distribution, Generalized Extreme Value (GEV) distribution, Gamma or Pearson type III (GAM) distribution) are defined in Table 1 in terms of their cumulative distribution function (cdf), $G_j(x, \vartheta)$, or probability density function (pdf), $g_j(x, \vartheta)$. Three other distributions, namely, the Frechet or Extreme Value 2 (EV2) distribution, the two-parameter lognormal (LN) distribution, and the log-Pearson type 3 (LP3) distribution, are converted to EV1, NORM, and GAM distributions, respectively, when the data are preliminarily log transformed. In detail, Monte Carlo experiments are structured as follows: (1) One of the seven probability distributions, M_{j^*}, with a specified set of parameters ϑ^* is set as parent distribution, $f(x) = g_{j^*}(x, \vartheta^*)$. (2) The $f(x)$ is used for generating a sample of length n. (3) The parameters of the seven probabilistic models are estimated by using the maximum likelihood method [see *Laio*, 2004]. For the GAM and GEV distributions, *Smith*'s [1985] estimators are used instead of maximum likelihood estimators when the latter do not exist or are not asymptotically efficient; see *Laio* [2004, Appendix A] for details. (4) The Akaike information criterion, Bayesian information criterion, and Anderson Darling criterion (see section 2) are calculated for each of the seven models; AIC_j, BIC_j, and ADC_j values are obtained, $j = 1, .., 7$. (5) The model M_{i^*} with the minimum value of AIC is selected, $AIC_{i^*} = AIC_{\min}$; if $i^* = j^*$, the AIC is successful, since it has recognized the correct parent distribution. The same procedure is repeated for BIC and ADC. (6) Steps 2–5 are repeated m times, and the percentage of times each criterion produces a successful selection are counted.

[18] A first set of Monte Carlo experiments is performed by taking $n = 50$ and $m = 1000$. For visualizing the obtained results, a type A diagram is constructed as follows (see Figure 1): The percentage of cases in which each candidate distribution is selected by each criterion is reported as a histogram bar; the title of a type A diagram, acronym (ϑ^*), indicates the adopted parent distribution and the parameter set ϑ^* used for generating 1000 samples of 50 data each. The diagram also shows a dashed line at a percentage value equal to 1/7 (around 14.28%), corresponding to a completely random model selection; this gives a simple "no-skill" model selection that can be used for comparison. A second set of Monte Carlo experiments is performed by varying the sample size n between 10 and 100 at steps of 10. Again, different probability distributions are used as parents, and the percentage of cases is calculated when the parent distribution is correctly selected by each criterion. A type B diagram is constructed (see Figure 2) by plotting the percentage of cases in which the real parent distribution is correctly selected by each criterion as a function of the sample size n. The title of the diagram again indicates the parent distribution and the adopted parameter set.

[19] Presentation of the results of these model selection experiments is complicated by the fact that the results depend on the parent distribution, the parameters of the parent, the sample size, and the statistical properties of the generated sample in general. To make the presentation clearer, we limit our attention to a few cases that we consider the most important ones for the purposes of the present paper. In particular, we concentrate our attention to the parent distributions that belong to the extreme-value family (EV1, EV2, or GEV distributions).

[20] By analyzing Figure 1 one can observe that when EV1 is used as the parent distribution, all three of the model selection criteria correctly select EV1 in around 40–60% of the cases; BIC turns out to be the best criterion, even when varying the sample size n (see, e.g., EV1(10,2), Figure 2). This result does not relevantly change when EV1 distributions with different ϑ_2 values (i.e., with different variances) are taken as parent distributions (see Figure 1 and Figure 2).

[21] When the EV2 distribution is used as the parent, the model selection criteria correctly recognize in most of the cases that the parent distribution belongs to the extreme-value family. However, the BIC is more effective than the AIC and ADC (in this order) in recognizing the EV2 from the GEV.

[22] When GEV is used as the parent distribution, the performance of the model selection criteria is more contradictory (see Figure 1). For samples with skewness close to 0, all three of the model selection criteria indicate NORM as

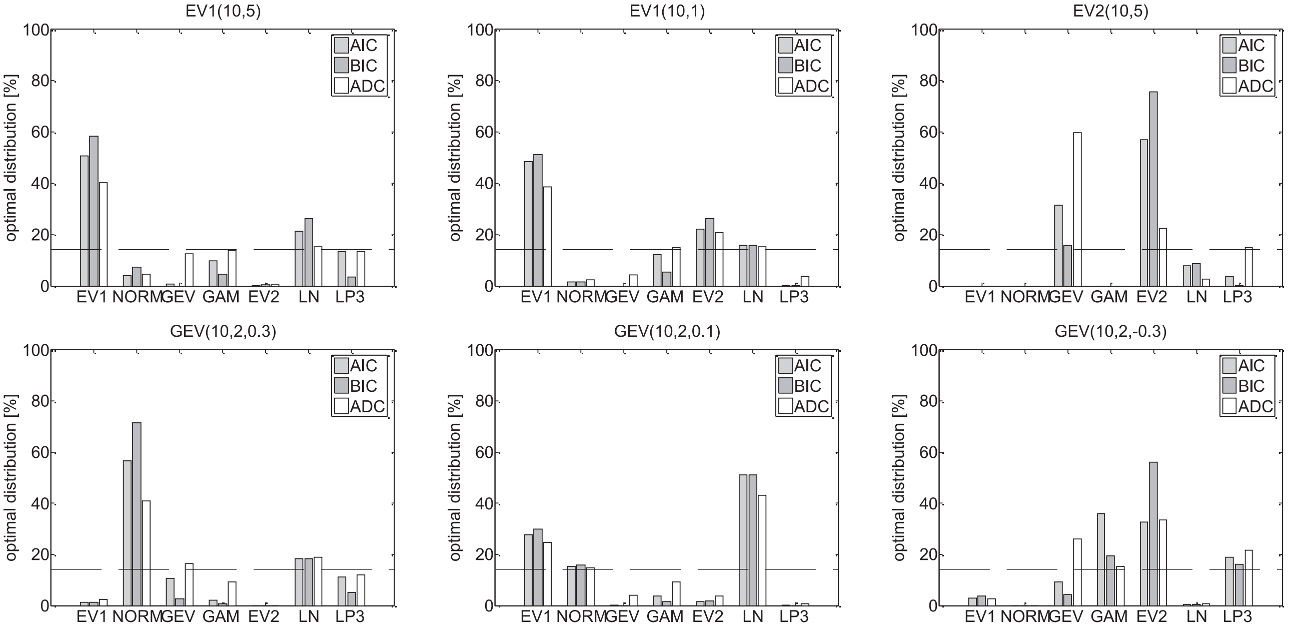

Figure 1. Type A diagrams: percentage of cases in which each candidate distribution is selected by each criterion (the title of each diagram indicates the parent distribution), with sample size $n = 50$; percentage value equal to 1/7 corresponding to random model selection (dashed lines).

the optimal distribution (e.g., GEV(10, 2, 0.3), Figure 1), while the EV2 is frequently selected when the parent distribution is heavy tailed (e.g., GEV(10, 2, −0.3), Figure 1). There is therefore a tendency of the criteria to select two-parameter distributions instead of the actual three-parameter parent. This outcome is not necessarily a limitation when estimating the flood quantile on the basis of small samples (see below). ADC turns out to be the best criterion in this case, even when varying the sample length n (see Figure 2), but the percentage of cases for which the GEV distribution is correctly recognized remains below 40%, even when dealing with large sample sizes. In order to better inspect the outcomes of the experiment with a GEV parent, we show in Figure 3 the percentage of cases in which the GEV is correctly selected by each criterion for varying shape parameter, ϑ_3 (positive ϑ_3 values correspond to upper bounded distributions, when $\vartheta_3 = 0$ the GEV degenerates into an EV1 distribution, and negative ϑ_3 correspond to heavy tailed distributions). ADC performs better than a no-skill model selection only for particular values of the parameter ϑ_3. Indeed, for certain parameter combinations the three-parameter distributions behave very similarly to the two-parameter ones, therefore making the model selection more complicated, especially for small sample sizes. Only for highly asymmetric samples the performances of the model selection criteria improve, as the three-parameter parent distribution becomes more clearly recognizable.

[23] The predisposition toward the selection of two-parameter distributions is the result of the principle of parsimony; see section 2. The ADC criterion has a less marked tendency toward model parsimony; as a consequence, it performs worse than AIC and BIC when a two-parameter distribution is used as parent distribution and better when a three-parameter distribution is used as parent. Summarizing, if a two-parameter distribution is used as the parent distribution, all three of the model selection criteria are able to recognize the correct parent in a large percentage of cases; among the criteria, BIC tends to perform better than AIC and ADC, even when varying the sample length n. On the

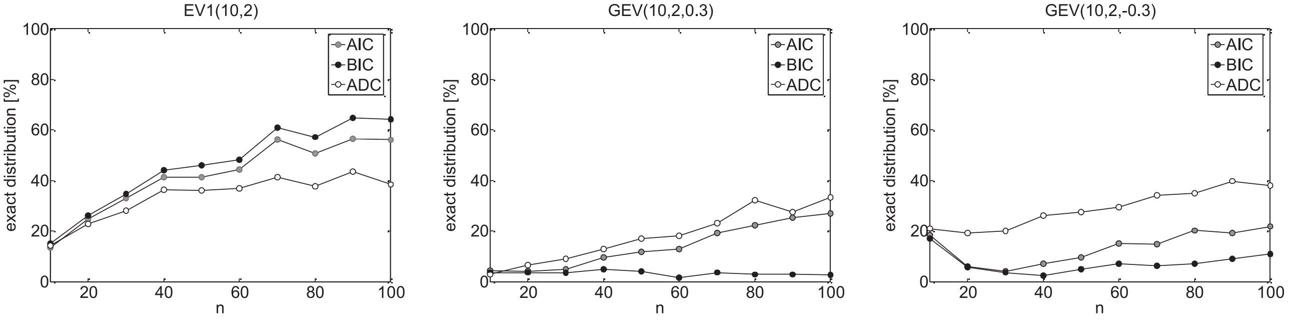

Figure 2. Type B diagrams: percentage of cases in which the real parent distribution is correctly selected by each criterion (the title of each diagram indicates the parent distribution) as a function of the sample size n.

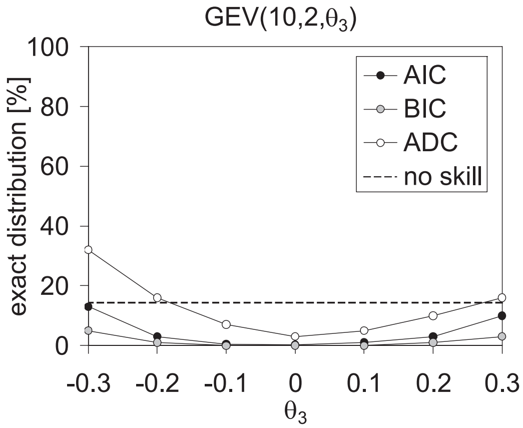

Figure 3. Percentage of cases in which the Generalized Extreme Value (GEV) (real parent) distribution is correctly selected by each criterion as a function of the shape parameter ϑ_3. The sample size is $n = 50$.

contrary, when a three-parameter distribution is used as the parent, the performance of the three model selection criteria depends on the set of parameters under consideration. In this case, ADC performs better than AIC and BIC even when varying the sample size n.

[24] From these analyses, however, it is not clear if a marked tendency toward parsimony is an advantage or a drawback of the criteria. For example, if the aim of flood frequency analysis is extrapolation to rare events with the smallest possible estimation error, it could be convenient to select a two-parameter distribution even when the parent is a three-parameter distribution. This changes the perspective of the model selection problem, which turns from a problem of recognition of the real parent to a problem of selection of the best operational model in terms of quality of the design event estimation. In this latter case, however, two further elements make the interpretation of the outcomes of the model selection exercise difficult: In fact, the results may change with the return period of the design events considered in the analyses, and with the error measure used to evaluate the quality of the selection. We postpone a full assessment of these nuisance factors to future research and present here the results of an additional analysis to better inspect the value of model selection criteria from the point of view of the accuracy of the estimated flood quantiles.

[25] A single case study is presented, wherein a GEV parent distribution is taken. A return period of 100 years is selected, and the target design value is found,

$$x_{100} = \vartheta_1 + \frac{\vartheta_2}{\vartheta_3}\left[1 - (-\log(0.99))^{\vartheta_3}\right].$$

[26] The model selection criteria are applied, and the selected distribution (with estimated parameters) is used to produce an estimate of the 100-year event, $\hat{x}_{100}$. This is repeated m times, with $m = 1000$, and an error measure E_2 is found by counting the number of times $\hat{x}_{100} < 0.85 \cdot x_{100}$ or $\hat{x}_{100} > 1.25 \cdot x_{100}$, i.e., when the estimated design value differs significantly from the real one. This error value is used due to its low sensitivity to outliers and for the possibility to account for the fact that in applied hydrology overestimation is less concerning than underestimation. The results of this analysis are shown in Figure 4, where the E_2 statistic is plotted against $\hat{\vartheta}_3$, i.e., the estimated shape parameter of the GEV (the corresponding ϑ_3 values for the parent distribution are varied between -0.5 and 0.5 at steps of 0.1). To further simplify the interpretation of the results, only two distributions (EV1 and GEV), rather than seven as in the previous analyses, are taken as candidate models. Since in this case the candidate models are only two, the likelihood ratio test is also used as a model selection criterion. For all criteria E_2 is found to decrease with $\hat{\vartheta}_3$, as an effect of the increased predictability of samples with large $\hat{\vartheta}_3$ values (lower skewness, light tails). Some differences in the E_2 values are apparent for $\hat{\vartheta}_3$ in the range $[-0.4{:}-0.2]$, where ADC performs better than the other criteria, and for positive $\hat{\vartheta}_3$ values, where in contrast ADC performs slightly worse than the other criteria. These results remain nearly unchanged when a different sample size is considered, except for a shift of the curves toward lower (higher) E_2 values when a larger (smaller) sample size is considered. Figure 4 also shows that by using a model selection criterion one obtains estimates of the flood quantiles that are almost always comparable or improved with respect to those obtained using the likelihood ratio test. This outcome confirms the potential utility of model selection techniques.

[27] A last points regards the use of corrected formulas to account for the small sample size. As mentioned (section 2), *Burnham and Anderson* [2002] recommend to use AICc when the ratio $n/p < 40$. Therefore a set of Monte Carlo experiments is performed for comparing the performance of AICc to that of AIC. Some results of these experiments are shown in Figure 5. We do not find significant differences between AIC and AICc: The corrected criterion AICc performs slightly better than AIC when a two-parameter distribution is used as the parent distribution (e.g., EV1,

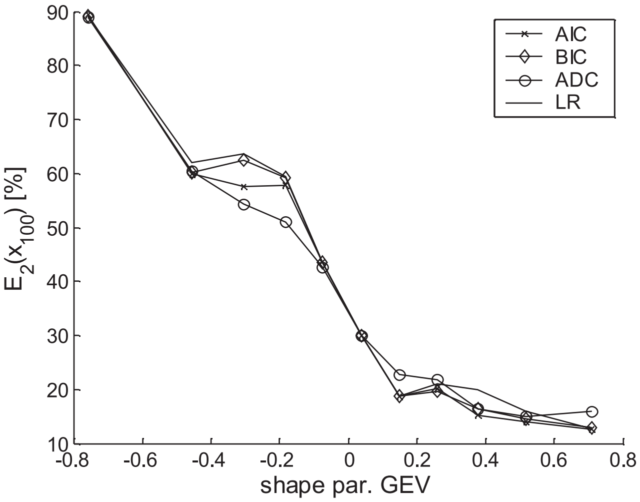

Figure 4. Percentage of wrong predictions, E_2, of the 100-year event (see text for details) as a function of the shape parameter of the parent GEV distribution. Model selection is carried out with two candidate models (EV1 and GEV), and also the likelihood-ratio (LR) test is used as a criterion. The sample size is $n = 20$.

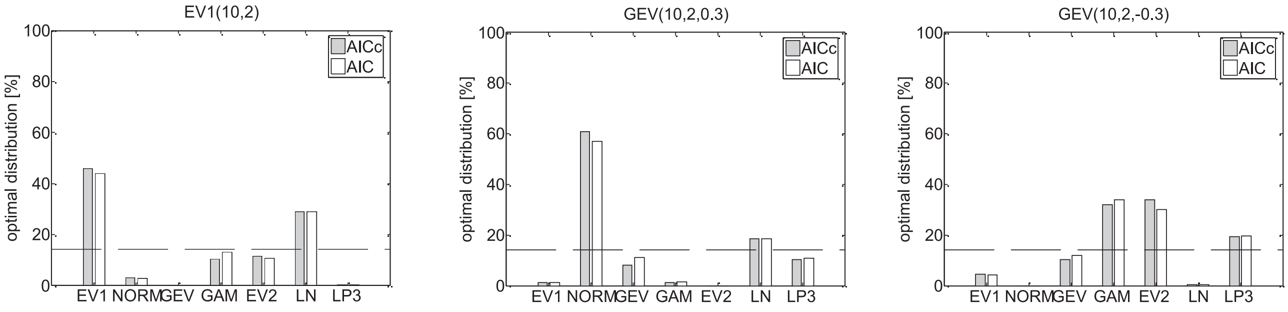

Figure 5. Type A diagrams (see Figure 1) for comparing Akaike information criterion (AIC) and second-order variant of AIC (AICc) (see equations (1) and (2)). The sample size is n = 50.

Figure 5) and vice versa when a three-parameter distribution is used as the parent (e.g., GEV, Figure 5). This effect may be attributed to the correction factor ($n/(n - p - 1)$; see section 2, equation (2)) which increases when the number of parameters p increases. Therefore the tendency of the AIC to select two-parameter distribution (principle of parsimony) is amplified when using the AICc. The differences between AIC and AICc are not significant also when varying the sample size n (not shown).

4. An Application to Real Data

[28] We have applied the three model selection criteria considered herein to a large data set of annual maxima of peak discharge pertaining to catchments located in the United Kingdom. The database is the one distributed with the *Flood Estimation Handbook* [*Institute of Hydrology*, 1999]. It contains 1000 annual maximum series of instantaneous peak flows with varying length, from a minimum of 5 years to a maximum of 112 years. A total of 23,378 data are available. The result of the application of the three model selection criteria to the UK data set is reported in Figure 6, using a type A diagram: The three model selection techniques are separately applied to each of the 1000 available samples, and the percentage of cases in which each candidate distribution is selected by each criterion is reported as a histogram bar. A first interesting result is that all distributions are selected in a rather large percentage of cases, which demonstrates that the seven distributions considered in this paper are all potentially suitable for the modeling of hydrological extremes. Only the GEV distribution is selected in a limited number of cases; the reason behind this lack of performance of the GEV is probably also in the relatively small sample sizes available in some basins. This result should not be interpreted as a clue of the inadequacy of the GEV distribution for modeling UK data: If a unique distribution was to be selected for all basins, three-parameter distributions would be preferred, but we are considering a different problem, that of selecting the "best" distribution for each of the 1000 basins. Our result in Figure 6 merely implies that in the majority of basins there is at least one distribution performing better than the GEV. When working with observations, the parent distribution is unknown, and it is then impossible to decide which criterion provides the best result. However, some interesting considerations derive from the analysis of Figure 6: From the comparison of the results obtained with the different model selection criteria, it turns hat the ADC leans more than the others toward the selection of distributions belonging to the extreme value family (EV1, GEV, and EV2), while the AIC and BIC select more frequently distributions of the log-transformed variable (EV2, LN, and LP3).

[29] The performances of model selection criteria when dealing with real-world data deserve further investigations. For example, it is of interest to consider the percentage of times each criterion selects a two-parameter or a three-parameter distribution, to better understand how the different criteria implement the principle of parsimony. The percentage of times a three- parameter distribution (GEV, GAM or LP3) is selected is reported in Figure 7 for the UK data, which have been subdivided in eight classes, based on their samples size. The boundaries of the classes are {10, 15, 20, 25, 30, 40, 50}, and the number of elements in each class is {97, 123, 175, 187, 159, 182, 48, 29}. Figure 7 demonstrates that the BIC tends to select more frequently the three-parameter distributions in smaller samples, while the reverse is true for the ADC. For the AIC the percentage of selections of three-parameter distributions remains almost constant with n. Only the ADC seems therefore to produce results that are consistent with a standard notion of the principle of parsimony (smaller samples require less parameterized distributions).

[30] Some other considerations are of interest regarding the capacity of the different criteria to select the same distribution. Over the entire data set, the BIC and AIC

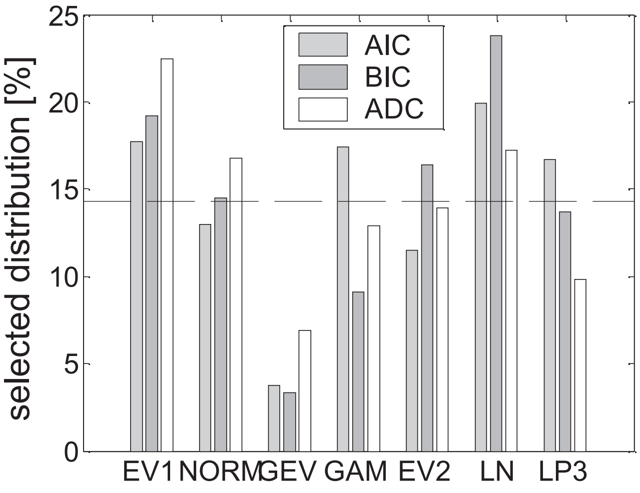

Figure 6. Type A diagram (see Figure 1) for the UK peak discharge data.

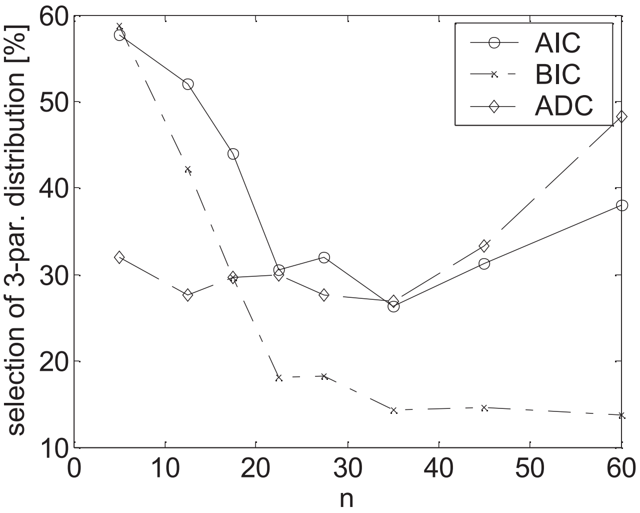

Figure 7. Percentage of selections of a three-parameter distribution (GEV, GAM, or LP3) by the three criteria for the UK peak discharge data, plotted as a function of the sample size *n*.

select the same distribution for 87% of the catchments. In contrast, the ADC selects the same probabilistic models as the AIC in 49% of the cases, and the same as the BIC in 51% of the cases: As a consequence, for about half of the time series the ADC selects a probability distribution different from the one selected by the AIC or BIC. On the one hand, this is a clue of the difficulty of model selection in the hydrologic field, and on the other it demonstrates that the ADC has different selection skills than the other methods. This point deserves some further analyses: In Figure 8 we plot the percentage of cases when the distribution selected by two criteria is the same as a function of the sample size (using the same classes as in Figure 7); it is found that the percentage of equal selections between ADC and AIC (or BIC) tends to increase with the sample size, as a consequence of the decreased uncertainty in selection. In contrast, the percentage of equal selections of AIC and BIC is always very large but slightly decreases with *n*; in fact, the two criteria have a very similar mathematical formulation (see equations (1) and (3)), and they provide exactly the same selection for $n = 8$. For larger *n* values, the penalty term in equation (1) and (3) is slightly different for the AIC and BIC, and the two criteria in some cases select different distributions. The fact that the ADC has different selection skills than the other methods can be an important quality; in fact, two probabilistic models can be equally good at representing the data (equifinality of the models, i.e., capacity to lead to an equally acceptable representation of the real process), and it is therefore important to have a method that tackles the model selection issue from a different perspective and carries different kind of information than the AIC and BIC.

5. Conclusions

[31] We have carried out an intensive numerical and data-based analysis of the performances of three model selection criteria, the Akaike information criterion (AIC), Bayesian information criterion (BIC), and Anderson-Darling criterion (ADC). The following considerations apply: (1) No model selection criterion performs consistently better than the others. (2) All criteria are rather effective at identifying the correct parent distribution when the parent is a two-parameter distribution; in contrast, they are less efficient when the real parent is a three-parameter distribution. However, this is not necessarily a limitation for the sake of estimating the flood quantile on the basis of small samples. (3) The BIC leans more than the AIC and the ADC (in this order) toward selecting more parsimonious (i.e., less parameterized) models. (4) If the aim of flood frequency analysis is extrapolation to rare events with the smallest possible estimation error, it could be convenient to select a two-parameter distribution even when the parent is a three-parameter distribution. This changes the perspective of the model selection problem, which turns from a problem of recognition of the real parent to a problem of selection of the best operational model in terms of quality of the design event estimation. Preliminary analyses have shown that model selection criteria are rather effective also under this ‘‘operational’’ perspective of the model selection problem. (5) The AIC and BIC usually produce the same model selection, while the ADC in many cases selects a different model than the others. The model selection criteria are an interesting alternative (or complement) to standard statistical testing, with respect to which they present the advantage of identifying with an objective procedure the best performing probability distribution.

[32] In summary, the model selection criteria provide promising results when applied to the frequency analysis of hydrological extremes; however, as often happens in statistical hydrology's applications, the obtained results are not completely conclusive because it remains unclear which criterion should actually be adopted in practical applications. In our opinion a good operational strategy could be to use the AIC (or BIC, which provides very similar information) in combination with the ADC. If the two criteria provide the same result, one can safely use the selected model for frequency analysis; otherwise one could think of

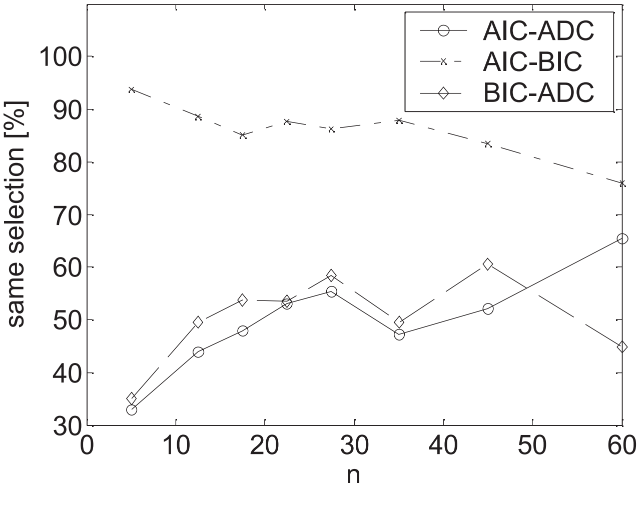

Figure 8. Percentage of times two criteria select the same distribution for the UK peak discharge data, plotted as a function of the sample size *n*.

using two different models, under the framework of equifinality, for the frequency analysis.

Appendix A: Akaike Information Criterion

[33] The Akaike information criterion uses the Kullback-Leibler's information as the discrepancy measure between the true model *f*(*x*) and the approximating model $M_j = g_j(x, \vartheta)$. For continuous functions, the Kullback-Leibler information between the two models is defined as [*Kullback and Leibler*, 1951]

$$I(M_j, f(x)) = \int \ln\left(\frac{f(x)}{g_j(x, \vartheta)}\right) f(x)\mathrm{d}x, \tag{A1}$$

where the notation $I(M_j, f(x))$ denotes the information lost when M_j is used to approximate *f*(*x*). By definition, the discrepancy of Kullback-Leibler is the expected value of the distance between the logarithms of the hypothetical and real probability distributions, where the expectation is taken over the real model *f*(*x*). As a heuristic interpretation, $I(M_j, f(x))$ is the distance from M_j to *f*(*x*). As a model selection criterion we seek an approximating model that loses as little information as possible; this is equivalent to minimizing $I(M_j, f(x))$ over M_j [*Cover and Thomas*, 1991; *Burnham and Anderson*, 2002]. The true model *f*(*x*) is considered to be fixed and *M* varies over a space of models indexed by *j*. The application of a model selection criterion based on the Kullback-Leibler information requires that the values of $I(M_j, f(x))$, obtained with different operational models, are compared. As a consequence the discrepancy can be simplified as

$$\Delta_{KL}[M_j, f(x)] = -\int \ln(g_j(x, \vartheta)) f(x)\mathrm{d}x, \tag{A2}$$

which is obtained from equation (1) by considering that the term $\int \ln (f(x)) f(x)\mathrm{d}x$ can be neglected because it remains constant when different candidate models are considered. As the sample size increases and if $\hat{\vartheta}$ is the "minimum discrepancy estimator" of ϑ [*Linhart and Zucchini*, 1986, Appendix A], the expected value of the discrepancy can be estimated as

$$C_{KL} = \Delta_{KL}\left[g_j\left(x, \hat{\vartheta}\right), f_n(x)\right] + \hat{K}/n, \tag{A3}$$

where the first term is the empirical discrepancy due to approximation, evaluated with the empirical density function $f_n(x)$ in place of *f*(*x*) and with $g_j(x, \hat{\vartheta})$ in place of $g_j(x, \vartheta)$; the second term in equation (A3) is the discrepancy due to estimation, $\hat{K}$ being an asymptotically unbiased estimator of the so-called trace term [see *Linhart and Zucchini*, 1986, Appendix A]. The empirical density function $f_n(x)$ in equation (A3) is the formal derivative of the empirical distribution function $F_n(x)$, defined as follows:

$$F_n(x) = 0 \quad x < x_i \tag{A4a}$$

$$F_n(x) = \frac{i}{n} \quad x_i \leq x < x_{i+1}, i = 1, 2, \ldots, n-1 \tag{A4b}$$

$$F_n \qquad 1 \;\; x_n \leq x. \tag{A4c}$$

Because $F_n(x)$ is a discontinuous function, its derivative can be formally represented as

$$f_n(x) = \frac{1}{n}\sum\nolimits_{i=1}^{n} \delta(x - x_i), \tag{A5}$$

where $\delta(x - x_i)$ is the Dirac delta function. By setting equation (A5) in (A2) one finds

$$\Delta_{KL}\left[g_j\left(x, \hat{\vartheta}\right), f_n(x)\right] = -\frac{1}{n}\sum\nolimits_{i=1}^{n} \ln\left[g_j\left(x_i, \hat{\vartheta}\right)\right]. \tag{A6}$$

As for the estimator $\hat{K}$ of the trace term in equation (A3), it reduces to the number p_j of parameters in the operating model, $\hat{K} = p_j$, if the parent and operating model coincide. As a consequence, equation (A3) becomes

$$C_{KL} = -\frac{1}{n}\sum\nolimits_{i=1}^{n} \ln\left[g_j\left(x_i, \hat{\vartheta}\right)\right] + \frac{p_j}{n}. \tag{A7}$$

Akaike [1973], who was the first to propose this model selection technique, obtained the criterion named after him by multiplying C_{KL} by 2*n*. It is easily seen that equation (1) is thus obtained form equation (A7).

Appendix B: Bayesian Information Criterion

[34] Bayesian inference is statistical inference in which observations are used to update the probability (or degree of belief) that a hypothesis may be true. In contrast with the frequentist approach to probability, in Bayesian statistics there is no need to think of a parent distribution that originated the data because the observations, and eventually some prior information on the involved processes, are all that serves for the inference. Bayesian methods condition on the data actually observed and are therefore able to assign posterior probabilities to any number of hypotheses directly [e.g., *Kass and Raftery*, 1995]. Also, the definition of a measure of the discrepancy between the model M_j and the parent distribution *f*(*x*) would be pointless under the Bayesian framework; in contrast, one could think of defining a discrepancy $\Delta_B[M_j, D]$ between the operational model and the data *D*. A natural choice in the Bayesian framework is to relate $\Delta_B[M_j, D]$ to the posterior probability of the model M_j, conditioned upon the data *D*, $\Pr(M_j|D)$: The larger is this posterior probability, the smaller is the discrepancy between the model and the data,

$$\Delta_B(M_j, D) = -\varphi(\Pr(M_j|D)), \tag{B1}$$

where $\varphi(\cdot)$ is a generic monotonically increasing function.

[35] Model selection can therefore be traced back to the definition of the following strategy: The model M_j is selected that maximizes $\Pr(M_j|D)$. This posterior probability can be expressed through the use of Bayes theorem as

$$\Pr(M_j|D) = \frac{\Pr(D|M_j)\Pr(M_j)}{\Pr(D)}, \tag{B2}$$

where $\Pr(D|M_j)$ is the probability of the data, given the model, $\Pr(M_j)$ is the prior probability of the model M_j, and $\Pr(D)$ is the (unconditional) probability of the data. If one considers a constant prior probability for all models (i.e., no model is favored a priori) and notes that Pr(D) is constant for all models, the relation

$$\Pr\left(M_j|D\right) \propto \Pr\left(D|M_j\right) \tag{B3}$$

is obtained. Since there are unknown parameters in the models, the densities Pr($D|M_j$) are obtained by integrating (not maximizing as in the usual frequentist approach to probability) over the parameter space [e.g., *Kass and Raftery*, 1995; *Konishi et al.*, 2004]:

$$\Pr\left(M_j|D\right) \propto \Pr\left(D|M_j\right) = \int \Pr\left(D|\vartheta, M_j\right)\pi\left(\vartheta|M_j\right)d\vartheta, \tag{B4}$$

where $\pi(\vartheta|M_j)$ is the prior distribution of the parameter vector ϑ in model M_j. If one further assumes that there is no prior information on the parameter values ($\pi(\vartheta|M_j)$ = const.), and considers that Pr($D|\vartheta$, M_j) is the likelihood function of ϑ,

$$\Pr\left(D|\vartheta, M_j\right) = L_j(\vartheta) = \prod_{i=1}^{n} g_j(x_i, \vartheta),$$

one finally finds out

$$\Pr\left(M_j|D\right) \propto \int L_j(\vartheta)d\vartheta. \tag{B5}$$

The integral in equation (B5) can be approximated through Laplace's method [e.g., *Kass and Raftery*, 1995; *Konishi et al.*, 2004], which entails approximating L_j (ϑ) through a multinormal density function with the mode in $\vartheta = \hat{\vartheta}$, where $\hat{\vartheta}$ is the maximum likelihood estimator of ϑ. Through Laplace's approximation one can solve the integral in (B5) to find

$$\Pr\left(M_j|D\right) \propto n^{-\frac{p_j}{2}} \prod_{i=1}^{n} g_j\left(x_i, \hat{\vartheta}\right) = n^{-\frac{p_j}{2}} L_j\left(\hat{\vartheta}\right). \tag{B6}$$

Once an estimator is obtained for Pr ($M_j|D$), it is sufficient to choose a suitable function $\varphi(\cdot)$ in equation (B1) to define a model selection criterion; by taking $\varphi(\cdot)$ = 2ln($\cdot$) one obtains the formulation of the Bayesian Information Criterion reported in equation (3).

Appendix C: Anderson-Darling Criterion

[36] We define the Anderson-Darling criterion starting from a discrepancy measure which is a weighted mean squared distance between the hypothetical, $G_j(x, \vartheta)$, and real, $F(x)$, cumulative probability distributions:

$$\Delta_{AD}\left[M_j, f(x)\right] = n \int \left[F(x) - G_j(x, \vartheta)\right]^2 \psi(x) f(x) dx, \tag{C1}$$

where $\psi(x)$ is a weight function,

$$\psi(x) = \left[G_j(x, \vartheta)\ \ 1 - G_j(x, \vartheta))\right]^{-1}, \tag{C2}$$

that allows one to have the discrepancies in the tails of the distribution weighted more than those in the central part. In goodness-of-fit applications, Δ_{AD} is called the Anderson-Darling test statistic. By setting the empirical cumulative function $F_n(x)$ (see equations (A4a), (A4b), and (A4c)) in place of $F(x)$, the empirical density function $f_n(x)$ in place of $f(x)$, and $G_j(x, \hat{\vartheta})$ in place of $G_j(x, \vartheta)$, one obtains a suitable estimator for $\Delta_{AD}[M_j, f(x)]$ [see e.g., *Stephens*, 1986]:

$$\Delta_{AD}\left[g_j\left(x, \hat{\vartheta}\right), f_n(x)\right] = -n - \frac{1}{n} \sum\nolimits_{i=1}^{i=n} \Big[(2i-1)\ln\left[G_j\left(x_i, \hat{\vartheta}\right)\right] + (2n+1-2i)\ln\left[1 - G_j\left(x_i, \hat{\vartheta}\right)\right]\Big]. \tag{C3}$$

[37] However, the discrepancy measure as defined in equation (C3) does not suitably account for the different number of parameters (estimated from the data sample) in the different operating models M_j. As a consequence, the more parameterized models would often produce lower discrepancies, without a suitable penalty term to account for overfitting. This same problem is encountered when the Anderson-Darling goodness-of-fit test is applied: If the parameter values need to be estimated form the same sample being tested, the distribution of Δ_{AD} significantly deviates from the one pertaining to the case when the parameter values are known a priori [e.g., *Stephens*, 1986]. More in detail, it turns out that the distribution of Δ_{AD} depends on the hypothetical model under consideration. This hampers the direct use of Δ_{AD} as a model selection criterion. Under the framework of goodness-of-fit tests, *Laio* [2004] proposed a method to overcome this difficulty, which is based on the use of a transformation $ADC_j = \phi_j(\Delta_{AD})$ which converts the estimated discrepancy values to ADC_j values. It has been demonstrated that the distribution of ADC_j (that of *Laio* [2004] is called ω) does not depend on the hypothetical model under consideration; as a consequence, ADC_j is suitable to be used as a model selection criterion. The transformation ϕ_j has the form reported in equation (4).

[38] **Acknowledgments.** The study has been supported by the Italian Ministry of Education through the grant 2006089189 (national research project entitled "Advanced techniques for estimating the magnitude and forecasting extreme hydrological events, with uncertainty analysis"). We are grateful to P. Naveau, D. Rosbjerg, D. Koutsoyiannis, and two anonymous referees for useful comments on an earlier version of the paper.

References

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csaki, pp. 267–281, Acad. Kiadó, Budapest.

Box, G. E. P., and G. M. Jenkins (1970), *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, Calif.

Bozdogan, H. (1987), Model selection and Akaike's information criterion (AIC): The general theory and its analytic extensions, *Psychometrika*, *52*, 270–345.

Browne, M. W. (2000), Cross-validation methods, *J. Math. Psychol.*, *44*, 132, doi:10.1006/jmps.1999.1279.

Burnham, K. P., and D. R. Anderson (2002), *Model Selection and Multimodel Inference*, 2nd ed., Springer, New York.

Cahill, A. T. (2003), Significance of AIC differences for precipitation intensity distributions, *Adv. Water Resour.*, *26*, 457–464, doi:10.1016/S0309-1708(02)00167-7.

Chung, H.-Y., K.-W. Lee, and J.-Y. Koo (1996), A note on bootstrap model selection criterion, *Stat. Probab. Lett.*, *26*, 35–41, doi:10.1016/0167-7152(94)00249-5.

Cover, T. M., and J. A. Thomas (1991), *Elements of Information Theory*, John Wiley, Hoboken, N. J.

Di Baldassarre, G., A. Brath, and A. Montanari (2006a), Reliability of different depth-duration-frequency equations for estimating short-duration design storms, *Water Resour. Res.*, *42*, W12501, doi:10.1029/2006WR004911.

Di Baldassarre, G., A. Castellarin, and A. Brath (2006b), Relationships between statistics of rainfall extremes and mean annual precipitation: An application for design storm in northern central Italy, *Hydrol. Earth Syst. Sci.*, *10*, 589–601.

Hache, M., L. Perrault, and L. Remillard (1999), An approach for statistical model selection: Application to the Saguenay-Lac-St-Jean hydrographic basin, *Can. J. Civ. Eng.*, *26*(2), 216–225, doi:10.1139/cjce-26-2-216.

Hosking, J. R. M., and J. R. Wallis (1997), *Regional Frequency Analysis*, Cambridge Univ. Press, Cambridge, U. K.

Hurvich, C. M., and C. Tsai (1989), Regression and time series model selection in small samples, *Biometrika*, *76*, 297–307, doi:10.1093/biomet/76.2.297.

Institute of Hydrology (1999), *Flood Estimation Handbook*, Wallingford, U. K.

Kass, R. E., and A. E. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, *90*, 773–795, doi:10.2307/2291091.

Konishi, S., and G. Kitagawa (1996), Generalized information criteria in model selection, *Biometrika*, *83*(4), 875–890, doi:10.1093/biomet/83.4.875.

Konishi, S., T. Ando, and S. Imoto (2004), Bayesian information criterion and smoothing parameter selection in radial basis function networks, *Biometrika*, *91*(1), 27–43, doi:10.1093/biomet/91.1.27.

Kottegoda, N. T., and R. Rosso (1997), *Statistics, Probability and Reliability for Civil and Environmental Engineers*, McGraw-Hill, New York.

Kroll, C. N., and R. M. Vogel (2002), Probability distribution of low streamflow series in the United States, *J. Hydrol. Eng.*, *7*(2), 137–146, doi:10.1061/(ASCE)1084-0699(2002)7:2(137).

Kullback, S., and R. A. Leibler (1951), On information and sufficiency, *Ann. Math. Stat.*, *22*, 79–86, doi:10.1214/aoms/1177729694.

Laio, F. (2004), 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.

Lindsey, J. K. (1999), Some statistical heresies, *J. R. Stat. Soc., Ser. D*, *48*(1), 1–4, doi:10.1111/1467-9876.00135.

Linhart, H., and W. Zucchini (1986), *Model Selection*, John Wiley, Hoboken, N. J.

Mitosek, H. T., W. G. Strupczewski, and V. P. Singh (2006), Three procedures for selection of annual flood peak distribution, *J. Hydrol.*, *323*, 57–73, doi:10.1016/j.jhydrol.2005.08.016.

Mutua, F. M. (1994), The use of the Akaike information criterion in the identification of an optimum flood frequency model, *Hydrol. Sci. J.*, *39*(3), 235–244.

Onoz, B., and M. Bayazit (1995), Best-fit distribution of largest available flood samples, *J. Hydrol.*, *167*, 195–204, doi:10.1016/0022-1694(94)02633-M.

Pandey, M. D., P. van Gelder, and J. K. Vrijling (2001), Assessment of an L-kurtosis-based criterion for quantile estimation, *J. Hydrol. Eng.*, *6*(4), 284–292, doi:10.1061/(ASCE)1084-0699(2001)6:4(284).

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*, 461–464, doi:10.1214/aos/1176344136.

Smith, R. L. (1985), Maximum likelihood estimation in a class of non-regular cases, *Biometrika*, *72*(1), 67–90, doi:10.1093/biomet/72.1.67.

Stedinger, J. R., R. M. Vogel, and E. Foufula-Georgiou (1992), Frequency analysis of extreme events, in *Handbook of Hydrology*, edited by R. Maidment, chap. 18, pp. 18.1–18.66, McGraw-Hill, New York.

Stephens, M. A. (1986), Tests based on EDF statistics, in *Goodness-of-Fit Techniques*, edited by R. B. D'Agostino and A. M. Stephens, pp. 97–194, Marcel Dekker, New York.

Strupczewski, W. G., V. P. Singh, and W. Feluch (2001), Non-stationary approach to at-site flood frequency modeling: I. Maximum likelihood estimation, *J. Hydrol.*, *248*, 123–142, doi:10.1016/S0022-1694(01)00397-3.

Strupczewski, W. G., V. P. Singh, and S. Weglarczyk (2002), Asymptotic bias of estimation methods caused by the assumption of false probability distributions, *J. Hydrol.*, *258*, 122–148, doi:10.1016/S0022-1694(01)00563-7.

Strupczewski, W. G., H. T. Mitosek, K. Kochanek, V. P. Singh, and S. Weglarczyk (2006), Probability of correct selection from lognormal and convective diffusion models based on the likelihood ratio, *Stochastic Environ. Risk Assess.*, *20*, 152–163, doi:10.1007/s00477-005-0030-5.

Sugiura, N. (1978), Further analysis of the data by Akaike's information criterion and the finite corrections, *Commun. Stat. Theory Methods*, *A7*, 13–26, doi:10.1080/03610927808827599.

Turkman, K. F. (1985), The choice of extremal models by Akaike's information criterion, *J. Hydrol.*, *82*, 307–315, doi:10.1016/0022-1694(85)90023-X.

Viglione, A., F. Laio, and P. Claps (2007), A comparison of homogeneity tests for regional frequency analysis, *Water Resour. Res.*, *43*, W03428, doi:10.1029/2006WR005095.

Vogel, R. M., O. Wilbert, and T. A. McMahon (1993a), Floodflow frequency model selection in southwestern United States, *J. Water Resour. Plann. Manage.*, *119*(3), 353–366, doi:10.1061/(ASCE)0733-9496(1993)119:3(353).

Vogel, R. M., T. A. McMahon, and F. Chiew (1993b), Floodflow frequency model selection in Australia, *J. Hydrol.*, *146*, 421–449, doi:10.1016/0022-1694(93)90288-K.

Wasserman, L. (2000), Bayesian model selection and model averaging, *J. Math. Psychol.*, *44*, 92–107, doi:10.1006/jmps.1999.1278.

Zucchini, W. (2000), An introduction to model selection, *J. Math. Psychol.*, *44*, 41–61, doi:10.1006/jmps.1999.1276.

G. Di Baldassarre, Department of Hydroinformatics and Knowledge Management, UNESCO-IHE Institute for Water Education, NL-2628 Delft, Netherlands. (g.dibaldassarre@unesco-ihe.org)

F. Laio, Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, Torino I-10129, Italy. (francesco.laio@polito.it)

A. Montanari, Dipartimento di Ingegneria delle Strutture, dei Trasporti, delle Acque, del Rilevamento, del Territorio, Università degli Studi di Bologna, Bologna I-40100, Italy. (alberto.montanari@unibo.it)