Canadian Journal of Civil Engineering
 Revue canadienne de génie civil

**DISCRIMINATION BETWEEN A GROUP OF THREE-
 PARAMETER DISTRIBUTIONS FOR HYDRO
 METEOROLOGICAL FREQUENCY MODELING**

Journal:	<i>Canadian Journal of Civil Engineering</i>
Manuscript ID	cjce-2017-0416.R1
Manuscript Type:	Article
Date Submitted by the Author:	26-Nov-2017
Complete List of Authors:	Ba, Ismaila; Université de Moncton, Dépt. de mathématiques et de statistique Ashkar, Fahim; Université de Moncton, Dépt. de mathématiques et de statistique
Is the invited manuscript for consideration in a Special Issue? :	N/A
Keyword:	hydrology < Hydrotechnical Eng., rivers-lakes-est-& reserv < Hydrotechnical Eng.

SCHOLARONE™
 Manuscripts

DISCRIMINATION BETWEEN A GROUP OF THREE-PARAMETER DISTRIBUTIONS FOR
HYDRO METEOROLOGICAL FREQUENCY MODELING

Ismaila Ba

Fahim Ashkar

Department of Mathematics and Statistics, Faculty of Sciences, Université de Moncton,
Moncton, N.B., Canada.

Corresponding author: Fahim Ashkar, Department of Mathematics and Statistics, Faculty of
Sciences, Université de Moncton, Moncton, N.B., E1A 3E9, Canada,
Telephone # 1 506 858-4312, fax # 1 506 858-4541, e-mail ashkarf@umoncton.ca

Abstract

We recommend methods of discrimination between some three-parameter distributions used in hydro-meteorological frequency modeling. Discriminations are between model pairs belonging to the group {generalized extreme value (GEV), Pearson Type III (P3), generalized logistic (GLO)}. To assess the fit of these distributions to data, the Akaike information criterion (AIC), Bayesian information criterion (BIC), and/or goodness-of-fit measures are commonly employed. However, it is difficult to estimate the discrimination power and bias of these methods when used with three-parameter distributions. Consequently, we propose two alternative tools and assess their performance. Both tools are based on a sample transformation to normality followed by applying a powerful statistic for testing normality, such as the Shapiro-Wilk or the probability plot correlation coefficient statistic. While arriving at recommendations for discriminating between the (GEV, GLO) and (P3, GLO) pairs of models, we show that the discrimination power between the P3 and GEV distributions can be rather low.

Keywords: Hydro-meteorological frequency modeling; Three-parameter distributions; Model selection; discrimination power; discrimination bias; Shapiro-Wilk statistic; Probability plot correlation coefficient statistic.

1. Introduction

Hydro-meteorological frequency analysis is concerned with analyzing the magnitudes of hydro-meteorological events and assessing their probability of occurrence, for use in risk assessment and management. These events can be floods, droughts, extreme rainfalls, or other extremes. In hydrology, the identification of a statistical distribution to model the frequency of occurrence of extreme events, and an efficient estimation of the distribution's parameters, are important. At present, a need exists for improved methods of selecting an appropriate statistical distribution to fit hydro-meteorological data. Such improved methods help reduce the error involved in quantile estimation. When it is reasonable to assume the data series to be independent and identically distributed (iid), hydrologists often fit several candidate distributions to the data and aim to choose the model with the best possible fit. The selection of a distribution is generally done on the basis of goodness-of-fit tests, in conjunction with graphical methods such as probability plots, but a fair degree of subjectivity still exists in the process involved in distribution selection. L-moment ratio diagrams (Hosking and Wallis 1997) are graphical methods used to explore the suitability of different statistical distributions, but the subjectivity involved in these methods limits their suitability for discriminating between candidate distributions.

After carefully assessing a group of candidate distributions, hydrologists often end up having to discriminate between a pair of models that appear to fit the data well, but only one of them is to be selected for the application at hand. Goodness-of-fit tests are generally used as a basis for rejecting some distributions but not to select the best distribution. This is why applying goodness-of-fit tests to two distinct models quite often leads to the rejection of neither of them, but the user still needs to make a unique model choice in an objective manner. Choosing the best model becomes especially important when estimating events in the distribution's tail(s), where the effect

of model assumption is critical. This makes it important to look for the most powerful and preferably most practical method(s) to discriminate between competing models.

Two-parameter probability distributions such as the Generalized Pareto, lognormal, gamma or Weibull are useful in fitting datasets in areas such as peaks-over-threshold hydro-meteorological extreme value modeling. Three-parameter distributions are also very important, such as in fitting annual maximum flood or precipitation series. These distributions include the generalized extreme value (GEV), Pearson Type III (P3), Log-Pearson Type III (LP3) and generalized logistic (GLO) distributions. Different countries have adopted some of these distributions for flood or precipitation frequency analysis. The United States Water Research Council (1977) recommends the LP3 distribution for flood frequency modeling, as is the case in Australia (Sangel and Kallio 1977; McMohan and SriKanthan 1981). China recommends the P3 distribution (Chen et al. 2010, 2012) and the British Natural Environment Research Council (NERC 1975) recommends the GEV for flood and rainfall analysis. The GLO distribution is recommended in the United Kingdom as their default flood frequency distribution (Robson and Reed 1999).

The present study focuses on a group of three-parameter distributions important in hydro-meteorological frequency modeling. It has the objective of recommending discrimination methods between these models. We will propose some discrimination procedures, justify their selection, and then test and compare them. We will base our comparisons on the procedures' "discrimination power" and "discrimination bias", which we will formally define in the next section.

2 Measuring model fit and discriminating between models

2.1 Measuring model fit

Based on a sample $\mathbf{X}_n = \{X_i\}_{i=1}^n$ of iid observations arising from an unknown distribution, goodness-of-fit tests serve to check the compatibility of $\mathbf{X}_n$ with a hypothesized cumulative distribution function F . Since some or all parameters of F are usually unknown, the user first needs to estimate them from the data. Hence, parameter estimation plays an important role in applying goodness-of-fit tests. Commonly used parameter estimation methods include those of maximum likelihood (ML), of moments and of probability weighted moments (PWM). Applying an estimation method to the sample, yields an estimate $\hat{F}$ of F .

Continuous probability distributions are particularly important in hydro-meteorological frequency modeling. When F is continuous, some popular goodness-of-fit tests make use of the empirical cumulative distribution function in assessing the fit of F to the data. The empirical cumulative distribution function is given by:

$$(1) \quad \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1\{X_i \leq x\}$$

Where:

$$(2) \quad 1\{X_i \leq x\} = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise} \end{cases}$$

Tests based on the empirical cumulative distribution function include the Kolmogorov-Smirnov, the Cramér-Von Mises and the Anderson-Darling tests (D'Agostino and Stephens 1986). What differentiates between these tests is the discrepancy measure they employ between the empirical cumulative distribution function $\hat{F}_n$ and the fitted cumulative distribution function $\hat{F}$. Other

goodness-of-fit tests are not based on the empirical cumulative distribution function, such as the Shapiro-Wilk (SW) test for normality (Shapiro and Wilk 1965) and the probability plot correlation coefficient (PPCC) test (Filliben 1975; Vogel 1986, and others). The Akaike information criterion and the Bayesian information criterion are also widely used to assess the fit of a hypothesized distribution to data.

2.2 Some remarks on the discrimination between models

The literature has given considerable attention to the discrimination between pairs of frequency models to fit data. Examples include such discriminations as:

lognormal vs. exponential (Cox 1961, 1962);

lognormal vs. gamma (Jackson 1968; Kappenman 1982; Quesenberry & Kent 1982);

lognormal vs. Weibull (Dumonceaux & Antle 1973; Kappenman 1982, 1989; Kundu & Manglick 2004);

Weibull vs. gamma (Bain & Engelhardt 1980; Kappenman 1982);

lognormal vs. Inverse Gaussian (Strupczewski et al. 2006);

lognormal vs. log-logistic (Dey & Kundu 2010; Aucoin & Ashkar 2010; Ashkar & Aucoin 2012b);

Weibull vs. log-logistic (Ashkar & Aucoin 2012a);

Generalized Pareto vs Kappa (Ashkar & Ba 2017).

It is noted that past studies have largely focused on discriminating between *two-parameter* models, which are important in hydrology as already mentioned. However, there is a clear need to pay more attention to discrimination between distributions with more than two parameters.

The ratio of maximized likelihood (RML) statistic is the most widely investigated method for choosing between frequency models. Early statistical studies used this statistic for discrimination (see e.g., Cox 1961, 1962; Jackson 1968). However, the large-sample basis of this discrimination statistic left some doubts with regard to its applicability to small- or moderate-sized samples. As a result, several subsequent studies (e.g., Dyer 1973; Dumonceaux & Antle 1973; Bain & Engelhardt 1980; Kappenman 1982; Kundu & Manglick 2004; Strupczewski et al. 2006; Aucoin & Ashkar 2010) resorted to Monte Carlo simulations to test the performance of the RML statistic with both large and small sample sizes. It is important to note that when the discrimination is between models with the same number of unknown parameters, such as in the present study, the RML statistic, the Akaike information criterion and the Bayesian information criterion methods become equivalent (as noted, for example, by Ashkar and Ba 2017). Other procedures for choosing between frequency models use goodness-of-fit measures that are based on the empirical cumulative distribution function given in Eq. (1). These include the Kolmogorov-Smirnov, the Cramér-Von Mises and the Anderson-Darling procedures mentioned in the previous section. In applying these procedures, many users choose the model with the lowest goodness-of-fit measure, although this is not necessarily the best way to apply these procedures. We will add a remark concerning this in section 3.

Some goodness-of-fit statistics aim at assessing the adequacy of *one specific type* of frequency model to fit a dataset. For example, the Shapiro-Wilk (SW) statistic is widely used to assess the fit of the *normal* distribution. Ashkar et al. (1997) proposed a modification to this statistic in order to allow for discrimination between a pair of *non-normal* distributions. This modification calls for transforming the sample to approximate normality and then applying the Shapiro-Wilk statistic to the transformed sample. We will denote this procedure by “TN.SW”.

Another discrimination statistic for assessing normality is the probability plot correlation coefficient (PPCC) statistic, proposed in 1975 by Filliben. Researchers, mainly in hydrology, later expanded the use of this statistic to test other distributions (e.g., Vogel 1986; Vogel & Kroll 1989; Chowdhury et al. 1991; Vogel & McMartin 1991; Heo et al. 2008; Kim & Heo 2010). In the present study, we will investigate this discrimination statistic both in its original form and in a modified form similar to the one just mentioned for the Shapiro-Wilk (SW) statistic. This modified form calls for transforming the sample to approximate normality, then applying the PPCC statistic to the transformed sample. We will denote this procedure by “TN.PPCC”.

2.3 Discriminating between three-parameter models

The focus of this study is to propose and compare practical methods of discrimination between three-parameter distributions such as GEV, P3, and GLO, which are important in hydro-meteorological frequency modeling. In discriminating between a pair of models (M1, M2), we will focus on those discrimination statistics that do not cause numerical difficulties in fitting either model. The most important potential numerical difficulty lies in the non-convergence of the method used to fit model(s) M1 and/or M2. By Monte Carlo simulation, we will assess the performance of the discrimination statistics by analyzing their probabilities of correct selection (PCS) between the models M1 and M2. Based on such analysis, we will attempt to identify the discrimination statistics that possess the highest discrimination power between M1 and M2 and the lowest discrimination bias, where power and bias will be measured as follows:

- Let PCS_{M1} denote the PCS when M1 is the assumed true model, and PCS_{M2} be defined in an analogous manner. Details on how to calculate PCS_{M1} and PCS_{M2} will be presented in section 5.

- Calculate $PCS.mean$, as follows, and use it as a measure of discrimination power:

$$(3) PCS.mean = \frac{PCS_{M1} + PCS_{M2}}{2}$$

- Calculate $PCS.abs.diff$, as follows, and use it as a measure of discrimination (absolute) bias:

$$(4) PCS.abs.diff = |PCS_{M1} - PCS_{M2}|$$

A necessary first step in discriminating between models is to estimate their parameters from the data. The maximum likelihood estimation method is often favored based on its desirable properties of being consistent, asymptotically normal, and asymptotically efficient under the classical regularity conditions (Nagatsuka et al. 2014). Another attractive property of the maximum likelihood method is that it is invariant to reparameterization. However, for three-parameter distributions such as P3, the maximum likelihood method is often eliminated because it leads to problems such as the non-existence of estimates in some range of the parameters, large variability, and convergence problems pointed out by many authors (e.g., Bobée and Ashkar 1991; Nagatsuka et al. 2014). Moreover, in Monte Carlo simulations such as those to be employed in the present study, when the generated and the fitted distributions are not of the same type (for example, generated = GEV, fitted = GLO), the maximum likelihood method should be eliminated altogether because it consistently leads to non-convergence. For these reasons, we will not use maximum likelihood as the fitting method but will replace it by the method of probability weighted moments, which we found not to suffer from the aforementioned drawbacks exhibited by the maximum likelihood method.

3 The discrimination statistics to be tested

As mentioned earlier, the ratio of maximized likelihood (RML) statistic is widely used to choose between a pair of competing models. In previous RML studies, which focused largely on discriminations between two-parameter distributions, the RML method did not pose any numerical problems. It also performed relatively well, although it was not necessarily the best method compared to others; see, e.g., Ashkar and Aucoin (2012a, 2012b), Ashkar and Ba (2017). However, attempting to use RML to discriminate between three-parameter distributions, led to some serious numerical problems. The reason being that this method calls for using maximum likelihood for parameter estimation, and as mentioned earlier, applying maximum likelihood to 3-parameter models led to convergence problems that necessitated replacing it by the probability weighted moments method. These problems occurred when the Monte-Carlo-generated distribution and the fitted distribution were not of the same type. For this reason, the RML method, and consequently the Akaike information criterion and the Bayesian information criterion, will not be pursued further in the present study, due to difficulties in calculating their discrimination power and bias by Monte Carlo simulation.

In applying goodness-of-fit measures based on the empirical cumulative distribution function (e.g., Kolmogorov-Smirnov, Cramér-Von Mises or Anderson-Darling) to select between pairs of models (M1, M2), many users simply choose the model that provides the lowest goodness-of-fit measure (e.g., the lowest value of the Anderson-Darling statistic). However, it can be shown that applying these procedures in this manner could lead to a large discrimination bias, by favoring one competing model over the other (Ashkar and Aucoin 2012a, 2012b). To partially correct for this bias, a “p-value-based” approach for choosing between M1 and M2 could be proposed (Ashkar and Aucoin 2012a). However, this approach demands computer programming that

191 makes it unattractive in practice. Therefore, for practical reasons, we will not pursue methods that
 192 are based on the empirical cumulative distribution function (e.g. those that use the Anderson-
 193 Darling statistic); although they may be interesting to pursue in future research.

194 Some studies (e.g., Ashkar & Aucoin 2012a, 2012b; Ashkar and Ba 2017) have shown the
 195 TN.SW statistic (mentioned in subsection 2.2) to display a relatively high discrimination power
 196 when tested with two-parameter distributions. The TN.SW statistic employs an approximate
 197 sample transformation to normality, followed by an application of the Shapiro-Wilk (SW)
 198 statistic. Appendix B presents the steps involved in calculating the TN.SW statistic S_{M1}
 199 (resp. S_{M2}) when M1 (resp. M2) is the fitted model.

200 If s_{M1} (resp. s_{M2}) is a realization of S_{M1} (resp. S_{M2}) based on an observed sample $\mathbf{x}_n$, the decision
 201 rule is to choose M1 as the true model if $s_{M1} > s_{M2}$ and to choose M2 otherwise.

202 The probability plot correlation coefficient (PPCC) goodness-of-fit statistic is another statistic
 203 used in model testing. As already mentioned, Filliben (1975) developed this statistic to test
 204 normality, but later studies expanded its use to test other models. This statistic is based on the
 205 correlation R between the ordered observations $X_{(i)}$ and the corresponding fitted quantiles
 206 $W_i = \hat{F}^{-1}(p_i)$, where p_i is a plotting position of $X_{(i)}$. The plotting position formula used in the
 207 present study was the Hazen formula: $p_i = (i - 0.5)/n$. The statistic R is given by:

$$(5) \quad R = \frac{\sum \left(X_{(i)} - \bar{X} \right) \left(W_i - \bar{W} \right)}{\left[\sum \left(X_{(i)} - \bar{X} \right)^2 \sum \left(W_i - \bar{W} \right)^2 \right]^{0.5}}$$

209 where $\bar{X}$ and $\bar{W}$ are the means of $\mathbf{X}_n$ and of $\mathbf{W}_n = \{W_i\}_{i=1}^n$ respectively.

Essentially, R measures the linearity of the probability plot of W_i versus $X_{(i)}$, and values of R close to 1.0 indicate a good fit of the distribution to the sample. One may choose to use R to discriminate between two models (M1, M2). Let R_{M1} (resp. R_{M2}) be the PPCC statistic when M1 (resp. M2) is the fitted model. If r_{M1} (resp. r_{M2}) is a realization of R_{M1} (resp. R_{M2}) based on an observed sample $\mathbf{x}_n$, the decision rule would be to select M1 as the true model if $r_{M1} > r_{M2}$ and to select M2 otherwise.

However, our experience has shown that a modification of the R statistic tends to improve its performance. This modification, which we denoted in subsection 2.2 by TN.PPCC, employs a sample transformation to normality, followed by an application of the PPCC statistic. The calculation of the TN.PPCC statistics R^*_{M1} and R^*_{M2} is outlined in Appendix B. If r^*_{M1} (resp. r^*_{M2}) is a realization of R^*_{M1} (resp. R^*_{M2}) based on an observed sample $\mathbf{x}_n$, the decision rule is to choose M1 to be the true model if $r^*_{M1} > r^*_{M2}$ and to choose M2 otherwise.

Based on the foregoing discussion, the three discrimination statistics that we will focus on in the remainder of this study are TN.SW, PPCC, and TN.PPCC.

4 The distributions to consider

The discriminations that we will consider are between model pairs belonging to the group {GEV, P3, GLO}. Numerical problems encountered with the LP3 distribution prohibited it from being included in the analysis, as we will explain later in this section.

Hereafter, we present the probability density function, cumulative distribution function and quantile function for the GEV, GLO and P3 distributions. Appendix A provides additional information on these models, such as moment and probability-weighted-moments parameter estimators.

232 The cumulative distribution function, probability density function and quantile functions of the
 233 GEV distribution are respectively given by:

$$234 \quad (6) \quad \begin{cases} F_{GEV}(x) = \exp \left\{ - \left[1 - k \left(\frac{x-c}{b} \right) \right]^{1/k} \right\} \\ f_{GEV}(x) = b^{-1} \left[1 - k \left(\frac{x-c}{b} \right) \right]^{1/k-1} F_{GEV}(x) \\ x(F) = c + \frac{b \left[1 - (-\ln F)^k \right]}{k} \end{cases}$$

235 where b , c , and k are scale, location, and shape parameters, respectively; and where:

$$236 \quad (7) \quad \begin{cases} b > 0, k \neq 0, \\ k < 0 \Rightarrow x > c + b/k \\ k > 0 \Rightarrow x < c + b/k \end{cases}$$

237 For the GLO distribution, we have:

$$238 \quad (8) \quad \begin{cases} F_{GLO}(x) = \left\{ 1 + \left[1 - k \frac{(x-c)}{b} \right]^{1/k} \right\}^{-1} \\ f_{GLO}(x) = b^{-1} \left[1 - k \frac{(x-c)}{b} \right]^{(1/k)-1} \left\{ 1 + \left[1 - k \frac{(x-c)}{b} \right]^{1/k} \right\}^{-2} \\ x(F) = c + \frac{b}{k} \left\{ 1 - \left[\frac{(1-F)}{F} \right]^k \right\} \end{cases}$$

239 where b , c , and k are scale, location, and shape parameters, respectively. The support (or sample
 240 space) of the GLO distribution is the same as that of the GEV distribution, given in Eq. (7).

241 For the P3 distribution, we have:

$$(9) \quad \begin{cases} f_{GAM3}(x) = \frac{\exp\left(-\frac{x-m}{\alpha}\right)\left(\frac{x-m}{\alpha}\right)^{\lambda-1}}{|\alpha|\Gamma(\lambda)} \text{ where } \Gamma(.) \text{ denotes the gamma function} \\ F_{GAM3}(x) \text{ and } x(F) \text{ are calculated numerically} \end{cases}$$

where α , m , and λ are scale, location, and shape parameters, respectively; and where:

$$(10) \quad \begin{cases} \lambda > 0, -\infty < m < \infty, \alpha \neq 0 \\ \alpha > 0 \Rightarrow x > m \\ \alpha < 0 \Rightarrow x < m \end{cases}$$

As we mention in Appendix A, the *lmom* package of the computing language *R* was the one used to estimate the P3 distribution parameters by probability weighted moments. Using this package with the LP3 distribution produced some convergence problems, which led to the exclusion of the LP3 distribution in this study.

5 Comparing the discrimination statistics by Monte Carlo simulation

Our objective is to compare three discrimination statistics: PPCC, TN.PPCC and TN.SW, with respect to their ability to correctly select between three model pairs (M1, M2). These model pairs are (GEV, GLO), (P3, GLO) and (GEV, P3). We will use the probability of correct selection (PCS) to compare the performance of the three discrimination statistics. The following steps describe the Monte Carlo algorithm that serves to calculate PCS_{M1} , the PCS when M1 is the assumed true model (when M2 is the true model, a similar algorithm yields PCS_{M2}):

1. Fix the sample size n and generate a sample $\mathbf{x}_n$ of n iid observations from model M1;
2. Fit both models M1 (the correct model) and M2 (the wrong model) to $\mathbf{x}_n$ by probability weighted moments;

3. Apply each of the three discrimination statistics and record the test result as "1" if model M1 is selected (*correct decision*) or "0" if model M2 is selected (*wrong decision*);
4. Repeat the preceding steps $N = 1000$ times, and store the test results ($I_j; j = 1, \dots, N$) in a vector I , which would contain only zeros and ones;
5. Calculate $\sum_{j=1}^N I_j / N$ and use it as an estimate of PCS_{M1} .

In selecting the sample sizes to include in the Monte Carlo experiment, it is useful to focus on those commonly encountered in hydrological practice. In this study, we chose sample sizes $n = 20$ (20) 100, which we consider to be representative of those frequently encountered in hydrology.

The PCS results to obtain from the Monte Carlo experiment depend on the parameter(s) of the distributions being considered, especially the shape parameter(s) values of these distributions. Since we are dealing with three-parameter distributions, we need to incorporate several 3-dimensional input vectors θ for each distribution into the Monte Carlo experiment. To select input parameters θ that are representative of those found in hydrological practice, we chose to base the selection on real hydrologic data. To this end, we considered annual maximum streamflow series recorded at 220 hydrometric stations from across Canada, chosen from the Water Survey of Canada's HYDAT database. These stations have been screened for regulation, diversions, or land use influences or changes; they are considered to have good quality data. For each series, we calculated the mean, the coefficient of variation c_v , and the coefficient of skewness c_s . Figure 1 depicts a plot of c_s versus c_v for these 220 series. The three quartiles of the observed c_s values are $Q_1 = 0.55$, Median = 0.97 and $Q_3 = 1.55$. From the 220 series, we selected a subset of 20 series with c_s values more or less uniformly distributed among the 220 observed c_s values. The objective was to choose a set of stations that exhibit a wide range of skewness in their

frequency distribution. Table 1 presents the 20 selected series, ordered according to their c_s values from smallest to largest. The series are identified as #1 to #20. It is noted that the c_s values range from 0.17 to 2.62, with the first half having $c_s < 1.00$ and the second half having $c_s > 1.00$. The table displays for each series, the mean, the standard deviation s , c_v and c_s . In choosing the 20 series, no consideration was given to any potential numerical difficulties to be encountered later in the analysis. However, with the parameter estimation method used, convergence problems were in fact later encountered with the last two stations in Table 1, which are the ones with $c_s > 2.00$. We therefore eliminated these two stations and limited consideration in our Monte Carlo simulations to the first 18 stations for which the estimation method converged. From each of the 18 $(\bar{x}, c_s, c_v)$ combinations, parameter vector estimates $\hat{\theta}$ were obtained for each distribution (GEV, P3, GLO) by the method of moments, using equations presented in Appendix A. For each distribution, we used these 18 parameter vector estimates as input into the Monte Carlo experiment. We will refer to these as: θ_{input} Vectors #1, #2, ..., #18.

5.1 Comparing PPCC and TN.PPCC

We will start by comparing the PPCC and TN.PPCC discrimination methods, where Table 2 presents some PCS results from the Monte Carlo experiment. To save space, we only present results for three sample sizes ($n = 20, 60, 100$) and five θ_{input} vectors (#4, #7, #10, #13, #16). However, we can confirm that these results are representative of the complete Monte Carlo results obtained for PPCC and TN.PPCC.

It is readily noted from Table 2 that the probabilities of correct selection, PCS_{M1} and PCS_{M2} , vary considerably. If we take as an example the case: [θ_{input} vector #4, discrimination statistic = PPCC, (M1, M2) = (GEV, GLO), $n = 100$], the reported PCS values are 90(68). This means that

the PPCC method correctly identifies the GEV distribution as the true model 90% of the time, while correctly identifying GLO as the true model only 68% of the time. The average PCS value (Eq. 3): $PCS.mean = (90\% + 68\%) / 2 = 79\%$ is a measure of discrimination power.

Additionally, the PCS absolute difference (Eq. 4): $PCS.abs.diff = |90\% - 68\%| = 22\%$ is a measure of discrimination absolute bias. Note that the bias in this case favors $M1 = \text{GEV}$ over $M2 = \text{GLO}$, because $PCS_{M1} = 90\% > PCS_{M2} = 68\%$.

From the information presented in Table 2, the two measures of discrimination power and discrimination absolute bias were calculated and are presented in Table 3. This table simplifies the comparison between PPCC and TN.PPCC. It shows that the difference in performance between these two discrimination statistics lies more in their discrimination absolute bias than in their power. This is clearly seen by referring to the last two rows marked “Average” in Table 3, as well as by referring to Figure 2 (A & B). In fact, Figure 2 (A) shows the difference in discrimination power between PPCC and TN.PPCC to be small, but Figure 2 (B) shows the difference in discrimination absolute bias to be much more substantial. From Figure 2 (B), and the last two rows marked “Average” in Table 3, the following conclusions may be drawn:

1) In the (GEV, GLO) discrimination, TN.PPCC decisively outperforms PPCC by producing lower discrimination absolute bias;

2) In the (P3, GLO) discrimination, TN.PPCC generally outperforms PPCC, except for the small sample size of $n = 20$;

3) In the (GEV, P3) discrimination, both TN.PPCC and PPCC fail to produce PCS.mean values much larger than 50%, which points to a difficulty in discriminating between the GEV and P3 distributions based on the population parameters and sample sizes included in the Monte Carlo experiment.

From the complete results obtained from the Monte Carlo experiment, we found no clear reason to continue investigating the PPCC discrimination statistic, because it was quite steadily outperformed by TN.PPCC with regard to discrimination absolute bias. Therefore, in the remainder of this section, we will shift attention to comparing the TN.PPCC and TN.SW discrimination statistics.

5.2 Comparing TN.SW and TN.PPCC

Table 4 and Figure 3 present the Monte Carlo simulation results for (GEV, GLO) discrimination. Figure 3 shows the difference between TN.SW and TN.PPCC to lie mainly in their discrimination absolute bias (Figure 3 (B)) than in their discrimination power (Figure 3 (A)). The same conclusion is arrived at by referring to the last two rows in Table 4 marked “Average”. These two rows show TN.PPCC to outperform TN.SW by producing lower discrimination absolute bias when $n \leq 80$, but TN.SW outperforms TN.PPCC for $n = 100$. The average PCS.mean value to expect from applying either of the two discrimination statistics is $\approx 59\%$ for $n = 20$ and $\approx 77\%$ for $n = 100$. Figure 3 (A) shows PCS.mean to decrease as the population skewness increases for both TN.SW and TN.PPCC. Figure 3 (B) shows TN.PPCC to outperform TN.SW by producing lower discrimination absolute bias when $n = 20, 40$, but for $n \geq 60$ both discrimination statistics show comparable discrimination absolute bias performance. As a practical recommendation, it seems reasonable to propose the use of TN.PPCC when $n \leq 50$ and to propose either of the two methods when $n > 50$. Under such a recommendation, the expected PCS.abs.diff value (a measure of discrimination absolute bias) should be less than 10% (i.e. a PCS mean error less than $\pm 5\%$), as can be seen from the last two rows marked “Average” in Table 4.

Table 5 and Figure 4 present the results for (P3, GLO) discrimination. Figure 4 shows the difference between TN.SW and TN.PPCC to lie more in their discrimination absolute bias (Figure 4 (B)) than in their discrimination power (Figure 4 (A)). The same conclusion is arrived at by referring to the last two rows marked “Average” in Table 5. These two rows show TN.SW to outperform TN.PPCC by producing lower discrimination absolute bias for all sample sizes considered. Therefore, it seems reasonable to consistently recommend TN.SW for (P3, GLO) discrimination, while keeping in mind that for small sample size (e.g., $n = 20$) and when there is an indication that the population skewness is quite large (e.g., $cs > 1.0$), TN.PPCC may be preferable to TN.SW. From such a recommendation, the expected PCS.mean value should be $\approx 61\%$ for $n = 20$ and $\approx 75\%$ for $n = 100$, as can be seen from the last two rows marked “Average” in Table 5. The expected PCS.abs.diff should be less than about 10% (PCS mean error less than $\pm 5\%$) for $n = 20$ and less than about 6% (PCS mean error less than $\pm 3\%$) for $n \geq 40$.

Table 6 and Figure 5 present the results for (GEV, P3) discrimination. Once again, Figure 5 shows TN.SW and TN.PPCC to have more differences in their discrimination absolute bias (Figure 5 (B)) than in their discrimination power (Figure 5 (A)). The same conclusion is arrived at by referring to the two rows marked “Average” in Table 6. These two rows and Figure 5 (A) show that both TN.SW and TN.PPCC fail to produce PCS.mean values much larger than 50%. In fact, one should not expect to obtain a PCS.mean value greater than 60% except when there is an indication that the population skewness is quite large; e.g., $cs > 1.2$. These results point to a difficulty in discriminating between the GEV and P3 distributions, at least from what our Monte Carlo experiments could show.

6 Application

As an application, we shall revisit the 18 stations considered in the previous sections. Table 7 contains the statistics computed for these 18 stations. The column marked “Chosen Model” shows the selected model for each station by applying the TN.SW and TN.PPCC statistics. In this column, it is seen that GEV appears six times, P3 seven times and GLO three times. These are the cases where both the TN.SW and TN.PPCC statistics lead to the same model choice. However, for two of the stations, which are marked by the symbol *, the two statistics do not lead to selecting the same model. These are stations # 5 and # 12. For these two stations, the TN.SW method recommends choosing GEV whereas TN.PPCC chooses GLO. So, the choice of model for these two stations requires revisiting our conclusions drawn from the simulation experiments. For this purpose, we will focus on Figure 3 (B) to find which model to recommend for these two stations. Since the sample size of station # 5 is equal to 42 and that of station # 12 is equal to 41, we will refer in Figure 3 (B) to the case $n = 40$. In this case, it is seen that TN.PPCC exhibits smaller discrimination bias as compared to TN.SW. Therefore, for these two stations, we will choose GLO as a model, in accordance with the TN.PPCC method.

As a further comparison of the distributions being investigated, we estimated the 100-year event by all three of them for the 18 stations by probability weighted moments. The GLO model consistently gave the highest 100-year-event estimates in comparison to the other two models. This may indicate a tendency of the GLO distribution to overestimate the 100-year event. The absolute difference between the highest and the lowest 100-year-event estimates was also calculated and is presented in the last column of Table 7. We express this absolute difference as a percentage of the highest 100-year-event estimate; i.e.,

$$Max. Abs. Diff = (|highest estimate - lowest estimate| / highest estimate) \times 100\%.$$
 It

is seen that the calculated absolute difference for the 18 stations varies between 4.8% and 12.5%:
a difference that could be considered quite substantial.

7 Discussion and conclusion

We dealt with discriminations between pairs of statistical distributions belonging to the group {GEV, P3, GLO}. We focused on sample sizes commonly encountered in hydro-meteorological applications and aimed at finding discrimination methods that are relatively powerful and easy to use. In assessing the methods' performance, we paid attention to both discrimination power and discrimination bias. While seeking high discrimination power is usually a key goal, seeking low discrimination bias is also essential, because we need to avoid favoring one competing model over the other in the discrimination. In our view, the assessment discrimination bias has not been given the attention that it deserves in previous research.

To assess the fit of distributions to data, the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and/or a goodness-of-fit measure such as the Anderson-Darling (AD) statistic are widely used in hydrological applications. However, we uncovered some inconveniences of these model choice methods, mainly relating to difficulties in assessing their performance in terms of discrimination power and discrimination bias. In applying the AD method, one should not simply choose the model that gives the lowest AD statistic, because this generally leads to a large discrimination bias between the competing models, as mentioned in section 3. For better discrimination, the "p-value-based" approach is more appropriate, but this approach requires a degree of computer programming that limits its appeal in practice. For choosing between models with a different number of unknown parameters, the AIC and BIC criteria have the appeal of offering a trade-off between model goodness of fit and complexity. However, because of their strict reliance on maximum likelihood estimation, they had to be

eliminated in this study due to serious difficulties in assessing their performance by Monte Carlo simulation, as we specified in subsection 2.3. For these reasons, we proposed TN.SW and TN.PPCC as discrimination tools, and were able to assess their performance by Monte Carlo simulation. Both of these tools are based on the idea of transforming the sample to approximate normality prior to performing model discrimination. TN.SW is a recently developed tool that needed further validation, while TN.PPCC is a new one that has been proposed and tested for the first time in the present study. An important practical advantage of both these methods is that they are easy to use.

For (GEV, GLO), (P3, GLO) and (P3, GEV) discriminations, TN.SW and TN.PPCC differed mainly in their discrimination bias, but not in their discrimination power. For (GEV, GLO) discrimination, our recommendation was to use TN.PPCC when $n \leq 50$, and either TN.SW or TN.PPCC when $n > 50$. With such a recommendation, the expected PCS mean error should be no more than $\approx \pm 5\%$, with a PCS.mean value $\approx 59\%$ for $n = 20$, and $\approx 77\%$ for $n = 100$. For (P3, GLO) discrimination, our general recommendation was to use TN.SW. Associated with such a recommendation is an expected PCS mean error of no more than $\approx \pm 3\%$ and a PCS.mean value $\approx 61\%$ for $n = 20$ and $\approx 75\%$ for $n = 100$. In the (P3, GEV) discrimination, both TN.SW and TN.PPCC failed to produce PCS.mean values much larger than 50%. This pointed to a difficulty in discriminating between the P3 and GEV distributions for the range of population parameters and sample sizes covered by our Monte Carlo simulations, which we chose based on observed annual maximum streamflow series from 220 Canadian hydrometric stations. The discrimination between P3 and GEV may prove to be less difficult if the population parameters are changed. This needs to be further investigated in future research.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the NSERC Canadian FloodNet strategic network. The hydrometric data used in this study is available from Environment Canada's HYDAT Database (<https://ec.gc.ca/rhc-wsc/default.asp?lang=En&n=9018B5EC-1>).

References

- Ashkar, F., and Aucoin, F. 2012a. Choice between competitive pairs of frequency models for use in hydrology: A review and some new results, *Hydrological Sciences Journal*, 57(6), 1092-1106.
- Ashkar, F., and Aucoin, F. 2012b. Discriminating between the lognormal and the log-logistic distributions for hydrological frequency analysis, *Journal of Hydrologic Engineering*, 17(1), 160-167.
- Ashkar, F., and Ba, I. 2017. Selection between the generalized Pareto and Kappa distributions in peaks-over-threshold hydrologic frequency modeling, *Hydrological Sciences Journal*, 62(7), 1167-1180.
- Ashkar, F., Arsenault, M., and Zoglat, A. 1997. On the discrimination between statistical distributions for hydrological frequency analysis, In: *Proceedings of the Canadian Society of Civil Engineers' Annual Conference: "Major Public Works: Key Technologies for the 21st Century"*, Sherbrooke, Quebec, Canada, 3, 169-178.
- Aucoin, F., and Ashkar, F. 2010. Discriminating between the logistic and the normal distributions based on likelihood (Online), *Interstat*. Available from <http://interstat.statjournals.net/YEAR/2010/abstracts/1008003.php> (Accessed 14 July 2016).

- 460 Bain, L.J., and Engelhardt, M. 1980. Probability of correct selection of Weibull versus gamma
461 based on likelihood ratio, *Communication in Statistics: Theory and Methods*, 9(4), 375-381.
- 462 Bobée, B., and Ashkar, F. 1991. The gamma family and derived distributions applied in
463 hydrology, Water Resources Publications.
- 464 Chen, L., Guo, S., Yan, B., Liu, P., and Fang, B. 2010. A new seasonal design flood method
465 based on bivariate joint distribution of flood magnitude and date of occurrence, *Hydrological
466 Sciences Journal*, 55(8), 1264-1280.
- 467 Chen, L., Singh, V.P., Shenglian, G., Hao, Z., and Li, T. 2012. Flood coincidence risk analysis
468 using multivariate copula functions, *Journal of Hydrologic Engineering*, 17(6), 742-755.
- 469 Chowdhury, J.U., Stedinger, J.R., and Lu, L-H. 1991. Goodness-of-fit tests for regional GEV
470 flood distributions, *Water Resources Research*, 27(7), 1765-1776.
- 471 Cox, D.R. 1961. Tests of separate families of hypotheses, In: *Proceedings of the Fourth Berkeley
472 Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley,
473 CA, 105-123.
- 474 Cox, D.R. 1962. Further results on tests of separate families of hypotheses, *Journal of Royal
475 Statistical Society*, 24, 406-424.
- 476 D'Agostino, R.B., and Stephens, M.A. 1986. Goodness-of-fit techniques. *Statistics, Textbooks
477 and Monographs* vol. 68, Marcel Dekker Inc.
- 478 Dey, A.K., and Kundu, D. 2010. Discriminating between the log-normal and log-logistic
479 distributions, *Communication in Statistics: Theory and Methods*, 39(2), 280-292.

- 480 Dumonceaux, R., and Antle, C.E. 1973. Discrimination between the lognormal and the Weibull
481 distributions, *Technometrics* (American Statistical Association and American Society for
482 Quality), 15(4), 923-926.
- 483 Dyer, A.R. 1973. Discrimination procedures for separate families of hypotheses, *Journal of the*
484 *American Statistical Association*, 68(344), 970-974.
- 485 Filliben, J.J. 1975. The probability plot correlation test for normality, *Technometrics*, 17(1), 111-
486 117.
- 487 Heo, J., Kho, Y., Shin, H., Kim, S., and Kim, T. 2008. Regression equations of probability plot
488 correlation coefficient test statistics from several probability distributions, *Journal of Hydrology*,
489 355, 1– 15.
- 490 Hosking, J.R.M., and Wallis, J.R. 1997. *Regional frequency analysis: an approach based on L-*
491 *moments*. Cambridge University Press, Cambridge
- 492 Hosking, J.R.M., Wallis, J.R., and Wood, E.F. 1985. Estimation of the generalized extreme-value
493 distribution by the method of probability-weighted moments, *Technometrics*, 27(3), 251-261.
- 494 Jackson, O.A.Y. 1968. Some results on tests of separate families of hypotheses, *Biometrika*,
495 55(2), 355-363.
- 496 Kappenman, R.F. 1982. On a method for selecting a distributional model, *Communication in*
497 *Statistics: Theory and Methods*, 11, 663-672.
- 498 Kappenman, R.F. 1989. A simple method for choosing between the lognormal and Weibull models,
499 *Statistics & Probability Letters*, 7(2), 123-126.

- 500 Kim, S., and Heo, J. 2010. Comparison of the Probability Plot Correlation Coefficient Test
501 Statistics for the General Extreme Value Distribution, World Environmental and Water
502 Resources Congress 2010: Challenges of Change, 2456-2466.
- 503 Kundu, D., and Manglick, A. 2004. Discriminating between the Weibull and log-normal
504 distributions, Naval Research Logistics, 51, 893-905.
- 505 Landwehr, J.M., Matalas, N.C., and Wallis, J.R. 1979. Probability weighted moments compared
506 with some traditional techniques in estimating Gumbel parameters and quantiles, Water
507 Resources Research, 15(5), 1055-1064.
- 508 McMohan, T.A., and Srikanthan, R. 1981. Log Pearson III distribution—Is it applicable to flood
509 frequency analysis of Australian streams? Journal of Hydrology, 52(1-2), 139-147.
- 510 Nagatsuka, H., Balakrishnan, N., and Kamakura, T. 2014. A consistent method of estimation for
511 the three-parameter gamma distribution, Communications in Statistics-Theory and
512 Methods, 43(18), 3905-3926.
- 513 NERC 1975. Flood Studies Report, Natural Environ. Res. Council HMSO, London, Vol. I.
- 514 Quesenberry, C.P., and Kent, J. 1982. Selecting among probability distributions used in reliability,
515 Technometrics, 24, 59-65.
- 516 Robson, A.J., and Reed, D.W. 1999. Statistical procedures for flood frequency estimation,
517 Volume 3 of the Flood Estimation Handbook, Oxfordshire: Centre for Ecology and Hydrology.
- 518 Royston, P. 1982a. An extension of Shapiro and Wilk's W test for normality to large samples,
519 Applied Statistics, 31, 115-124.
- 520 Royston, P. 1982b. Algorithm AS 181: the W test for normality, Applied Statistics, 31, 176-180.

- 521 Royston, P. 1995. Remark AS R94: a remark on algorithm AS 181: the W test for normality,
522 Applied Statistics, 44, 547-551.
- 523 Sangel, B.P., and Kallio, R.W. 1977. Magnitude and frequency of floods in Southern Ontario,
524 Inland Waters Directorate, Water Planning and Management Branch, Canada, ISBN:
525 0662016173, pp: 336.
- 526 Shapiro, S.S., and Wilk, M.B. 1965. An analysis of variance test for normality (complete
527 samples), Biometrika, 52(3), 591-611.
- 528 Strupczewski, W.G., Mitosek, H.T., Kochanek, K., Singh, V.P., and Weglarczyk, S. 2006.
529 Probability of correct selection from lognormal and convective diffusion models based on the
530 likelihood ratio, Stochastic Environmental Research and Risk Assessment, April 2006, Volume
531 20, Issue 3, 152-163.
- 532 Vogel, R.M. 1986. The probability plot correlation coefficient test for normal, lognormal and
533 Gumbel distributional hypotheses, Water Resources Research, 22(4), 587-590.
- 534 Vogel, R.M., and Kroll, C.N. 1989. Low-flow frequency analysis using probability-plot
535 correlation coefficients, Journal of Water Resources Planning and Management, ASCE, 115(3),
536 338-357.
- 537 Vogel, R.M., and McMartin, D.E. 1991. Probability plot goodness-of-fit and skewness estimation
538 procedures for the Pearson Type III distribution, Water Resources Research, 27(12), 3149-3158.
- 539 Water Resources Council U.S. 1977. Hydrology Committee, Guidelines for determining flood
540 flow frequency.
- 541

542 **Table 1.** Information on the 20 chosen annual maximum streamflow series

ID	Station	Province	Drainage Area (km ²)	Years	Mean	sd	cv	cs
#1	02VC001	Quebec	13000	52	1518	433	0.285	0.166
#2	07EE009	British Columbia	310	37	48.1	16	0.334	0.336
#3	02FC001	Ontario	3953.52	100	494.4	172.3	0.348	0.392
#4	03QC001	Newfoundland	10900	45	1756	519.7	0.296	0.458
#5	02UC002	Quebec	19000	42	2166	554.2	0.256	0.533
#6	08NH131	British Columbia	119	32	31.3	6.3	0.202	0.576
#7	04KA001	Ontario	4250	43	393.5	198.3	0.504	0.663
#8	08GA010	British Columbia	173	96	227.7	71.4	0.314	0.752
#9	04NA001	Quebec	3680	78	183	43	0.235	0.836
#10	06DA004	Saskatchewan	7730	48	94.6	28.2	0.298	0.954
#11	01BC001	New Brunswick	3160	50	613.4	223.5	0.364	1.012
#12	10RC001	Nunavut	93900	41	4557	1390	0.305	1.114
#13	08MH016	British Columbia	335	86	72.6	22.2	0.306	1.182
#14	02ZB001	Newfoundland	205	52	174.1	84.5	0.485	1.314
#15	11AA026	Alberta	569	24	10.4	10.2	0.984	1.451
#16	08NH130	British Columbia	585	42	129.4	34.4	0.266	1.574
#17	05FB002	Alberta	3500.3	20	5.3	5.2	0.981	1.732
#18	06AB002	Alberta	725.7	40	5.8	4.9	0.848	1.877
#19	08MB006	British Columbia	1010	39	47.6	28	0.588	2.189
#20	08NE006	British Columbia	330	51	97.2	38.1	0.391	2.62

543

Table 2. Probability of correct selection (PCS), in percentage, rounded to the nearest integer, for comparing PPCC and TN.PPCC. Outside of the brackets are PCS values under GEV as the true model; within brackets are PCS values under GLO as the true model.

Θ_{input} Vector	Discrimination Statistic	(M1, M2)								
		(GEV, GLO)			(P3, GLO)			(P3, GEV)		
		Sample size n			Sample size n			Sample size n		
		20	60	100	20	60	100	20	60	100
#4	PPCC	75(47)	85(61)	90(68)	60(60)	77(64)	83(68)	57(40)	57(44)	56(44)
	TN.PPCC	63(57)	66(75)	74(85)	50(69)	62(78)	68(82)	59(39)	56(45)	55(47)
#7	PPCC	75(45)	83(59)	86(66)	65(58)	82(62)	84(66)	53(45)	51(49)	51(51)
	TN.PPCC	63(56)	66(74)	73(84)	54(68)	66(77)	71(82)	58(40)	57(46)	56(47)
#10	PPCC	74(44)	79(58)	83(63)	72(56)	81(61)	86(65)	51(55)	49(62)	49(66)
	TN.PPCC	63(55)	65(73)	72(82)	60(67)	68(76)	74(81)	60(44)	61(54)	62(57)
#13	PPCC	73(43)	76(57)	80(61)	72(54)	81(60)	88(64)	52(59)	50(69)	51(76)
	TN.PPCC	64(54)	65(71)	70(79)	62(66)	66(76)	70(81)	62(50)	65(57)	67(64)
#16	PPCC	72(42)	74(54)	77(59)	76(52)	82(59)	87(64)	51(66)	51(75)	53(80)
	TN.PPCC	65(52)	66(68)	70(75)	64(64)	64(75)	65(82)	63(56)	71(61)	76(63)

Table 3. PCS means and absolute differences, both in percentage rounded to the nearest integer, for comparing PPCC and TN.PPCC. Values of *PCS.mean* (a measure of discrimination power) are outside the brackets; values of *PCS.abs.diff* (a measure of discrimination absolute bias) are within brackets. In the last two table rows, the “best” PCS results are highlighted.

Θ_{input} Vector	Discrimination Statistic	(GEV, GLO)			(P3, GLO)			(P3, GEV)		
		Sample size n			Sample size n			Sample size n		
		20	60	100	20	60	100	20	60	100
#4	PPCC	61(28)	73(24)	79(22)	60(0)	70(13)	76(15)	48(17)	50(13)	50(12)
	TN.PPCC	60(6)	70(9)	80(11)	60(19)	70(16)	75(14)	49(20)	50(11)	51(8)
#7	PPCC	60(30)	71(24)	76(20)	62(7)	72(20)	75(18)	49(8)	50(2)	51(0)
	TN.PPCC	60(7)	70(8)	78(11)	61(14)	72(11)	76(11)	49(18)	52(11)	52(9)
#10	PPCC	59(30)	68(21)	73(20)	64(16)	71(20)	76(21)	53(4)	56(13)	58(17)
	TN.PPCC	59(8)	69(8)	77(10)	64(7)	72(8)	78(7)	52(16)	58(7)	60(5)
#13	PPCC	58(30)	66(19)	70(19)	63(6)	72(6)	77(8)	56(7)	60(19)	64(25)
	TN.PPCC	59(10)	68(6)	74(9)	63(18)	70(21)	76(24)	56(12)	61(8)	66(3)
#16	PPCC	57(30)	64(20)	68(18)	64(24)	70(23)	76(23)	58(15)	63(24)	66(27)
	TN.PPCC	58(13)	67(2)	72(5)	64(0)	70(11)	74(17)	60(7)	66(10)	70(13)
Average	PPCC	59(29.6)	68.4(21.6)	73.2(19.8)	62.6(10.6)	71(16.4)	76(17)	52.8(10.2)	55.8(14.2)	57.8(16.2)
	TN.PPCC	59.2(8.8)	68.8(6.6)	76.2(9.2)	62.4(11.6)	70.8(13.4)	75.8(14.6)	53.2(14.6)	57.4(9.4)	59.8(7.6)

Table 4. Comparing TN.SW and TN.PPCC for (GEV, GLO) discrimination. This table gives PCS means and absolute differences, both in percentage rounded to the nearest integer. *PCS.mean* values are outside the brackets; *PCS.abs.diff* values are within brackets. In the last two rows of the table, the “best” PCS results are highlighted.

Θ_{input} Vector	Discrimination Statistic	Sample size n				
		20	40	60	80	100
#1	TN.SW	62(25)	68(12)	74(7)	78(4)	82(1)
	TN.PPCC	60(5)	69(6)	73(10)	78(11)	81(12)
#2	TN.SW	61(26)	67(14)	72(7)	76(5)	82(3)
	TN.PPCC	60(4)	68(6)	72(9)	76(12)	80(12)
#3	TN.SW	61(26)	67(14)	72(7)	76(5)	81(2)
	TN.PPCC	60(5)	68(5)	72(9)	76(11)	80(12)
#4	TN.SW	60(27)	66(15)	72(7)	76(5)	80(3)
	TN.PPCC	60(6)	68(5)	70(9)	76(11)	80(11)
#5	TN.SW	60(26)	66(14)	72(8)	76(7)	80(4)
	TN.PPCC	60(6)	67(4)	70(9)	75(10)	80(11)
#6	TN.SW	60(27)	66(15)	72(8)	75(6)	80(5)
	TN.PPCC	60(7)	67(4)	70(9)	75(10)	78(11)
#7	TN.SW	60(27)	66(16)	72(7)	74(7)	79(4)
	TN.PPCC	60(7)	67(4)	70(8)	75(10)	78(11)
#8	TN.SW	59(28)	65(16)	71(8)	74(7)	78(5)
	TN.PPCC	60(7)	66(3)	70(9)	74(10)	78(12)
#9	TN.SW	59(28)	64(17)	70(9)	74(8)	78(6)
	TN.PPCC	59(8)	66(3)	69(8)	74(8)	78(11)
#10	TN.SW	58(29)	64(17)	70(10)	74(9)	77(6)
	TN.PPCC	59(8)	66(2)	69(8)	73(8)	77(10)
#11	TN.SW	58(29)	64(18)	70(10)	74(9)	76(7)
	TN.PPCC	59(8)	66(2)	68(7)	73(8)	77(10)
#12	TN.SW	58(30)	64(18)	69(10)	72(9)	76(7)
	TN.PPCC	58(9)	66(1)	68(7)	72(7)	76(9)
#13	TN.SW	58(30)	64(19)	68(11)	72(10)	76(7)
	TN.PPCC	59(10)	66(1)	68(6)	72(6)	74(9)
#14	TN.SW	58(32)	64(19)	68(12)	72(11)	74(9)
	TN.PPCC	58(11)	64(0)	68(4)	72(5)	74(8)
#15	TN.SW	58(33)	63(20)	68(13)	70(11)	74(9)
	TN.PPCC	59(12)	64(2)	68(3)	71(4)	74(7)
#16	TN.SW	58(33)	62(21)	67(14)	70(11)	73(10)
	TN.PPCC	58(13)	64(3)	67(2)	70(3)	72(5)
#17	TN.SW	58(34)	62(22)	66(15)	69(14)	72(11)
	TN.PPCC	58(14)	63(4)	66(1)	70(2)	72(4)
#18	TN.SW	58(35)	62(23)	66(17)	69(14)	71(12)
	TN.PPCC	58(15)	63(6)	66(1)	70(1)	72(3)
Average	TN.SW	59.1(29.2)	64.7(17.2)	69.9(10)	73.4(8.4)	77.2(6.2)
	TN.PPCC	59.2(8.6)	66(3.4)	69.1(6.6)	73.4(7.6)	76.7(9.3)

Table 5. Comparing TN.SW and TN.PPCC for (P3, GLO) discrimination. This table gives PCS means and absolute differences, both in percentage rounded to the nearest integer. *PCS.mean* values are outside the brackets; *PCS.abs.diff* values are within brackets. In the last two rows of the table, the “best” PCS results are highlighted.

Θ_{input} Vector	Discrimination Statistic	Sample size n				
		20	40	60	80	100
#1	TN.SW	58(13)	63(10)	68(11)	71(8)	74(9)
	TN.PPCC	57(28)	62(30)	66(28)	70(25)	72(24)
#2	TN.SW	59(6)	65(2)	70(3)	72(1)	76(0)
	TN.PPCC	58(25)	64(22)	70(19)	72(17)	76(16)
#3	TN.SW	60(4)	64(3)	70(2)	74(3)	76(0)
	TN.PPCC	60(21)	64(21)	69(18)	72(16)	76(15)
#4	TN.SW	60(2)	66(1)	72(1)	74(3)	76(2)
	TN.PPCC	60(19)	65(18)	70(16)	73(14)	75(14)
#5	TN.SW	62(4)	66(3)	71(2)	74(3)	76(3)
	TN.PPCC	60(17)	66(15)	70(16)	73(14)	76(12)
#6	TN.SW	61(2)	68(5)	72(4)	75(6)	76(5)
	TN.PPCC	60(17)	66(15)	71(14)	74(12)	76(11)
#7	TN.SW	61(4)	66(5)	72(5)	74(6)	76(3)
	TN.PPCC	61(14)	67(14)	72(11)	74(10)	76(11)
#8	TN.SW	62(8)	68(7)	72(4)	75(8)	76(5)
	TN.PPCC	62(13)	68(11)	70(11)	74(10)	78(8)
#9	TN.SW	62(9)	66(7)	72(6)	74(8)	77(6)
	TN.PPCC	62(12)	68(10)	72(9)	76(7)	78(9)
#10	TN.SW	62(11)	67(8)	72(7)	74(8)	76(3)
	TN.PPCC	64(7)	68(9)	72(8)	75(6)	78(7)
#11	TN.SW	62(11)	67(8)	72(7)	74(7)	76(3)
	TN.PPCC	63(8)	68(8)	72(6)	75(6)	77(8)
#12	TN.SW	63(12)	68(10)	71(6)	72(4)	77(2)
	TN.PPCC	63(6)	70(5)	72(6)	74(8)	77(8)
#13	TN.SW	63(14)	68(9)	70(4)	72(3)	74(3)
	TN.PPCC	64(4)	68(7)	71(10)	74(9)	76(11)
#14	TN.SW	62(11)	66(7)	69(2)	70(1)	74(6)
	TN.PPCC	64(5)	68(7)	70(9)	72(11)	74(13)
#15	TN.SW	62(12)	66(6)	68(1)	70(1)	73(8)
	TN.PPCC	64(3)	68(6)	70(8)	73(10)	75(14)
#16	TN.SW	62(14)	64(5)	68(1)	69(6)	72(11)
	TN.PPCC	64(0)	66(7)	70(11)	72(13)	74(17)
#17	TN.SW	61(12)	63(2)	66(4)	68(9)	70(15)
	TN.PPCC	64(0)	66(7)	68(13)	71(16)	72(19)
#18	TN.SW	62(13)	63(2)	66(5)	67(10)	69(18)
	TN.PPCC	64(2)	66(6)	69(12)	70(18)	71(22)
Average	TN.SW	61.3(9)	65.8(5.6)	70.1(4.2)	72.2(5.3)	74.7(5.7)
	TN.PPCC	61.9(11.2)	66.6(12.1)	70.2(12.5)	73(12.3)	75.4(13.3)

Table 6. Comparing TN.SW and TN.PPCC for (GEV, P3) discrimination. This table gives PCS means and absolute differences, both in percentage rounded to the nearest integer. *PCS.mean* values are outside the brackets; *PCS.abs.diff* values are within brackets. In the last two rows of the table, the “best” PCS results are highlighted.

Θ_{input} Vector	Discrimination Statistic	Sample size n				
		20	40	60	80	100
#1	TN.SW	49(30)	52(30)	55(26)	56(24)	60(20)
	TN.PPCC	48(25)	51(18)	54(11)	56(9)	58(3)
#2	TN.SW	48(25)	52(20)	53(18)	53(18)	56(16)
	TN.PPCC	48(23)	51(12)	52(8)	53(6)	54(3)
#3	TN.SW	48(24)	51(18)	52(19)	53(18)	54(15)
	TN.PPCC	48(22)	50(11)	52(9)	52(7)	53(4)
#4	TN.SW	49(18)	51(14)	50(17)	52(14)	52(14)
	TN.PPCC	49(20)	50(11)	50(11)	52(9)	51(8)
#5	TN.SW	48(15)	50(11)	50(14)	50(12)	52(11)
	TN.PPCC	49(18)	49(12)	50(12)	50(11)	52(7)
#6	TN.SW	49(14)	50(9)	51(10)	51(10)	50(13)
	TN.PPCC	50(19)	48(11)	52(11)	52(9)	50(11)
#7	TN.SW	48(11)	50(8)	52(7)	52(3)	52(7)
	TN.PPCC	49(18)	50(11)	52(11)	52(8)	52(9)
#8	TN.SW	50(8)	52(4)	52(7)	54(0)	56(1)
	TN.PPCC	50(17)	52(13)	52(12)	54(9)	54(7)
#9	TN.SW	51(6)	52(5)	56(1)	56(5)	57(4)
	TN.PPCC	52(16)	52(13)	54(12)	55(8)	56(5)
#10	TN.SW	52(5)	54(2)	58(1)	59(6)	60(6)
	TN.PPCC	52(16)	54(13)	58(7)	59(6)	60(5)
#11	TN.SW	54(4)	54(1)	58(3)	60(7)	62(8)
	TN.PPCC	54(13)	55(14)	58(7)	60(6)	62(4)
#12	TN.SW	54(5)	58(5)	60(4)	62(7)	65(10)
	TN.PPCC	54(15)	58(9)	59(8)	62(5)	65(2)
#13	TN.SW	56(0)	58(3)	60(3)	63(6)	64(7)
	TN.PPCC	56(12)	58(10)	61(8)	63(4)	66(3)
#14	TN.SW	56(1)	60(3)	62(5)	63(4)	66(3)
	TN.PPCC	56(9)	58(9)	64(7)	64(8)	68(6)
#15	TN.SW	58(1)	62(5)	64(4)	66(3)	68(1)
	TN.PPCC	58(10)	62(6)	66(7)	67(8)	70(9)
#16	TN.SW	60(2)	62(4)	64(2)	65(2)	68(7)
	TN.PPCC	60(7)	63(6)	66(10)	67(10)	70(13)
#17	TN.SW	61(4)	62(2)	64(4)	66(6)	68(11)
	TN.PPCC	62(3)	64(6)	66(11)	68(14)	70(16)
#18	TN.SW	62(6)	63(2)	66(5)	66(9)	68(16)
	TN.PPCC	62(1)	66(7)	68(11)	68(15)	70(19)
Average	TN.SW	52.9(9.9)	55.2(8.1)	57.2(8.3)	58.2(8.6)	59.9(9.4)
	TN.PPCC	53.2(14.7)	55.1(10.7)	57.4(9.6)	58.6(8.4)	60.1(7.4)

Table 7. Use of the TN.SW statistic s and the TN.PPCC statistic r^* to choose between the GEV, GLO and P3 models for the 18 data series discussed in the Application

ID	Sample size	s.GEV	s.P3	s.GLO	r^* .GEV	r^* .P3	r^* .GLO	Chosen Model	100-Yr Event Max.Abs.Diff (%)
#1	52	0.981	0.98	0.971	0.993	0.992	0.989	GEV	5.9
#2	37	0.979	0.978	0.968	0.991	0.991	0.987	GEV	6.0
#3	100	0.989	0.989	0.987	0.995	0.995	0.995	P3	6.4
#4	45	0.975	0.976	0.964	0.99	0.99	0.985	P3	5.4
#5	42	0.986	0.985	0.985	0.991	0.991	0.993	*	5.0
#6	32	0.971	0.973	0.962	0.99	0.99	0.986	P3	4.8
#7	43	0.98	0.983	0.971	0.992	0.993	0.989	P3	7.9
#8	96	0.989	0.99	0.982	0.995	0.995	0.992	P3	5.7
#9	78	0.981	0.98	0.99	0.99	0.989	0.995	GLO	4.7
#10	48	0.977	0.98	0.969	0.991	0.992	0.988	P3	6.5
#11	50	0.987	0.971	0.988	0.992	0.983	0.994	GLO	6.7
#12	41	0.991	0.966	0.989	0.995	0.979	0.995	*	6.6
#13	86	0.994	0.972	0.991	0.997	0.984	0.996	GEV	6.7
#14	52	0.981	NA	0.973	0.992	NA	0.989	GEV	9.9
#15	24	0.95	0.979	0.942	0.98	0.992	0.976	P3	12.5
#16	42	0.979	0.976	0.982	0.989	0.987	0.992	GLO	5.7
#17	20	0.945	NA	0.938	0.978	NA	0.974	GEV	11.6
#18	40	0.976	NA	0.972	0.991	NA	0.989	GEV	12.3

574

Figure 1. plot of c_s versus c_v for the 220 annual maximum streamflow series.

575

576

Figure 2. PCS means (2A) and absolute differences (2B), for comparing PPCC and TN.PPCC.

Figure 3. PCS means (3A) and absolute differences (3B), for comparing TN.PPCC and TN.SW. The discrimination is between GEV and GLO

583

584

Figure 4. PCS means (4A) and absolute differences (4B), for comparing TN.PPCC and TN.SW. The discrimination is between P3 and GLO

Figure 5. PCS means (5A) and absolute differences (5B), for comparing TN.PPCC and TN.SW. The discrimination is between P3 and GEV

Appendix A

Here we give the moment and probability weighted moments estimators of the parameters of the GEV, P3 and GLO distributions.

A1 moments estimators

The estimators by the method of moments of the GEV distribution parameters have the following expressions:

$$(A1) \quad \hat{b} = |\hat{k}| \sqrt{(\hat{g}_2 - \hat{g}_1^2)s}$$

$$(A2) \quad \hat{c} = \bar{x} - \frac{\hat{b}}{\hat{k}}(1 - \hat{g}_1)$$

where $\hat{k}$ is obtained by using the method of Newton-Raphson to the GEV skewness function:

$$(A3) \quad c_s = -\frac{|k|}{k} \frac{g_3 - 3g_1g_2 + 2g_1^3}{(g_2 - g_1^2)^{3/2}}$$

and where:

$$(A4) \quad g_i = \Gamma(1 + ik)$$

For the GLO distribution, the parameter estimators by the method of moments have the following expressions:

$$(A5) \quad \hat{b} = \frac{s|\hat{k}|}{\left[2\hat{\alpha} \csc(\hat{\alpha}) - (\hat{\alpha} \csc(\hat{\alpha}))^2\right]^{1/2}}$$

(A6)
$$\hat{c} = \bar{x} - \frac{\hat{b}}{\hat{k}} \left(1 - \hat{\alpha} \csc(\hat{\alpha}) \right)$$

where $\hat{k}$ is obtained by applying the Newton-Raphson method to the GLO skewness function:

(A7)
$$c_s = \frac{-2\pi^2 \csc(\alpha) + 6\frac{\pi}{k} \csc(\alpha) \csc(2\alpha) - \frac{3}{k^2} \csc(3\alpha)}{\sqrt{\pi} \left[\frac{2}{k} \csc(2\alpha) - \pi \csc^2(\alpha) \right]^{3/2}}$$

and where:

(A8)
$$\alpha = \pi k$$

The moment estimators of the P3 distribution parameters have the following expressions:

(A9)
$$\hat{\lambda} = \frac{4}{(c_s)^2}$$

(A10)
$$|\hat{\alpha}| = \frac{s}{\sqrt{\hat{\lambda}}}$$

(A11)
$$\hat{m} = \bar{x} - \hat{\alpha} \hat{\lambda}$$

A2 probability weighted moments estimators

The probability weighted moments estimators of the GEV parameters are the solution of the following system of equations:

$$(A12) \quad \begin{cases} \beta_0 = c + \frac{b}{k} [1 - \Gamma(1+k)] \\ 2\beta_1 - \beta_0 = \frac{b}{k} [\Gamma(1+k)(1-2^{-k})] \\ \frac{3\beta_2 - \beta_0}{2\beta_1 - \beta_0} = \frac{1-3^{-k}}{1-2^{-k}} \end{cases}$$

where:

$$(A13) \quad \beta_r = \frac{1}{1+r} \left\{ c + \frac{b}{k} \left[1 - \frac{\Gamma(1+k)}{(1+r)^k} \right] \right\}, \quad k > -1, \quad r = 0, 1, 2, \dots \quad [\text{Hosking et al., 1985}]$$

β_0, β_1 and β_1 have to be replaced by the unbiased estimator proposed by Landwehr et al. (1979)

$$(A14) \quad b_r = \frac{1}{n} \sum_{i=1}^n \frac{(j-1)(j-2)\dots(j-r)}{(n-1)(n-2)\dots(n-r)} x_{(j)} \quad r = 0, 1, 2, \dots$$

where $x_{(j)}$ represents the ordered GEV distributed sample.

The probability weighted moments estimators are then:

$$(A15) \quad \hat{b} = \hat{k} \left[\frac{2b_1 - b_0}{\Gamma(1+\hat{k})(1-2^{-\hat{k}})} \right]$$

$$(A16) \quad \hat{c} = b_0 + \frac{\hat{b}}{\hat{k}} [\Gamma(1+\hat{k}) - 1]$$

where $\hat{k}$ is obtained from:

$$(A17) \quad \frac{3b_2 - b_0}{2b_1 - b_0} = \frac{1-3^{-k}}{1-2^{-k}}$$

The probability weighted moments estimators of the GLO parameters have the following expressions:

(A18)
$$\hat{k} = -\frac{\lambda_3}{\lambda_2}$$

(A19)
$$\hat{b} = \frac{\lambda_2}{\left\{ \Gamma(1-\hat{k}) \Gamma(1+\hat{k}) \right\}}$$

(A20)
$$\hat{c} = \lambda_1 - \hat{b} \left\{ 1 - \Gamma(1-\hat{k}) \Gamma(1+\hat{k}) \right\} / \hat{k}$$

where:

(A21)
$$\lambda_{r+1} = \sum_{j=0}^r P_{r,j} \beta_j, \quad r=0,1,2,\dots$$

(A22)
$$r\beta_{r-1} = c + \frac{b}{k} \left\{ 1 - \frac{\Gamma(1+k) \Gamma(1-k)}{\Gamma(r)} \right\}, \quad |k| < 1 \text{ et } r=1,2,\dots$$

(A23)
$$P_{r,j} = (-1)^{r-j} \binom{r}{j} \binom{r-j}{j}$$

We used the R package **lmom** to compute the needed probability weighted moments estimators of the P3 distribution.

Appendix B

Here we outline the steps involved in calculating the TN.SW statistics S_{M1} and S_{M2} and the TN.PPCC statistics R^*_{M1} and R^*_{M2}

B1 TN.SW

To calculate the TN.SW statistics S_{M1} and S_{M2} , start by assuming M1 to be the true model for the sample $\mathbf{X}_n = \{X_i\}_{i=1}^n$. Let F_{M1} be the cumulative distribution function of M1 and Φ^{-1} the quantile function of the standard normal distribution. If F_{M1} were known, the following variable transformation:

$$(B1) \quad Z_i = \Phi^{-1}(F_{M1}(X_i))$$

would transform $\mathbf{X}_n$ to an exact standard normal sample $\mathbf{Z}_n = \{Z_i\}_{i=1}^n$. However, since F_{M1} is unknown, it is estimated by $\hat{F}_{M1}$ using a parameter estimation method such probability weighted moments. Consequently, Eq. (B1) would be estimated by:

$$(B2) \quad Z_i^* = \Phi^{-1}(\hat{F}_{M1}(X_i))$$

This implies that $Z_i^* \sim \text{approx } N(0, 1)$. The foregoing transformed sample $\mathbf{Z}_n^* = \{Z_i^*\}_{i=1}^n$ is then used to calculate the TN.SW statistic in the same way the SW statistic is calculated:

$$(B3) \quad S_{M1} = \frac{\left[\sum_{i=1}^n v_i Z_{(i)}^* \right]^2}{\sum_{i=1}^n \left[Z_{(i)}^* - \bar{Z}^* \right]^2}$$

where $Z_{(i)}^*$ and $\bar{Z}^*$ are the i^{th} order statistic and the mean of $\mathbf{Z}_n^*$, respectively, and the v_i 's are coefficients approximated as in (Royston 1982a, b, 1995). Next, the random sample is assumed to arise from the model M2, and the statistic S_{M2} is calculated in a similar manner.

B2 TN.PPCC

Calculation of the TN.PPCC statistics R_{M1}^* and R_{M2}^* is done as follows: *a)* Start by assuming M1 to be the true model for the sample $\mathbf{X}_n$ and use $Z_i^* = \phi^{-1}(\hat{F}_{M1}(X_i))$ to transform $\mathbf{X}_n$ to an approximate $N(0, 1)$ sample $\mathbf{Z}_n^*$; *b)* calculate $\mathbf{W}_n^* = \{W_i^*\}_{i=1}^n$ where $W_i^* = \phi^{-1}(\hat{F}(W_i)) = \phi^{-1}(p_i)$; *c)* calculate the TN.PPCC statistic R_{M1}^* using the formula:

$$(B4) \quad R_{M1}^* = \frac{\sum (Z_i^* - \bar{Z}^*) (W_i^* - \bar{W}^*)}{\left[\sum (Z_i^* - \bar{Z}^*)^2 \sum (W_i^* - \bar{W}^*)^2 \right]^{0.5}}$$

d) Assume M2 to be the true model and repeat steps *a)* to *c)* to calculate R_{M2}^*