A bivariate extension of the Hosking and Wallis goodness-of-fit measure for regional distributions

T. R. Kjeldsen,[1] and I. Prosdocimi[2]

Corresponding author: T. R. Kjeldsen, Department of Architecture and Civil Engineering, University of Bath, Claverton Downs, Bath, BA2 7AY, United Kingdom. (t.r.kjeldsen@bath.ac.uk)

[1]Department of Architecture and Civil Engineering, University of Bath, Claverton Downs, Bath, BA2 7AY, United Kingdom.

[2]Centre for Ecology & Hydrology, Maclean Building, Wallingford, OX10 8BB, United Kingdom.

Abstract. This study presents a bivariate extension of the goodness-of-fit measure for regional frequency distributions developed by *Hosking and Wallis* [1993] for use with the method of L-moments. Utilising the approximate joint normal distribution of the regional L-skewness and L-kurtosis, a graphical representation of the confidence region on the L-moment diagram can be constructed as an ellipsoid. Candidate distributions can then be accepted where the corresponding theoretical relationship between the L-skewness and L-kurtosis intersects the confidence region, and the chosen distribution would be the one that minimises the Mahalanobis distance measure. Based on a set of Monte Carlo simulations it is demonstrated that the new bivariate measure generally selects the true population distribution more frequently than the original method. Results are presented to show that the new measure remains robust when applied to regions where the level of inter-site correlation is at a level found in real world regions. Finally the method is applied to two different case studies involving annual maximum peak flow data from Italian and British catchments to identify suitable regional frequency distributions.

1. Introduction

The seminal work of *Hosking* [1990], *Hosking and Wallis* [1993, 1997] and others [e.g. *Vogel and Fennessey*, 1993; *Institute of Hydrology*, 1999] popularised the use of L-moments and L-moment ratios in regional frequency analysis of environmental extremes such as floods. In particular, *Hosking and Wallis* [1997] presented a seemingly complete and robust framework for using the index flood method in combination with the method of L-moment, including measures for identifying discordant data series, assessing the homogeneity of regions, and evaluation of the goodness-of-fit of regional statistical distributions. This framework has been used by numerous researchers to develop regional flood frequency tools for many different geographical regions, e.g. *Vogel et al.* [1993], *Mkhandi et al.* [2000], and *Kumar et al.* [2003].

The results from simulation experiments reported by *Hosking and Wallis* [1997] showed that regional frequency analysis is generally more accurate than at-site analysis, especially for design events with very high return periods in excess of 1000 years. At the same time *Hosking and Wallis* [1997] reported that misspecification of the underlying regional frequency distribution becomes an important factor when considering design events with return periods in excess of 100 years. Thus, correctly specifying the regional distribution is a key task in order to fully capitalise on the benefits of regional frequency analysis.

Different methods for elucidating regional frequency distributions have been developed based on L-moment diagrams. Examples include the goodness-of-fit (GOF) measure presented by *Hosking and Wallis* [1993] in the form of a test statistic of a normal variate where the significance of the difference between a sample value of the regional L-kurtosis and a set of theoretical values corresponding to different 3-parameter distributions is assessed using Monte Carlo simulations. *Vogel et al.* [1993] recommended using the location of the regional mean of the L-moment ratio on the L-moment ratio diagram as a guide for the choice of an appropriate model. *Peel et al.* [2001] compared two different graphical methods for assessing the regional distribution based on L-moment diagrams, a sample average and a line of best fit through the sample L-moment ratios. They concluded that the sample mean was the most reliable method. *Madsen et al.* [1997] found that use of partial duration series data led to a less ambiguous interpretation of the L-moment diagram than the application of annual maximum series. Other researchers [e.g. *Liou et al.*, 2008; *Wu et al.*, 2012; *Wang and Hutson*, 2013] have utilised the approximate normal distribution of the L-moment ratios to develop graphical representations on a L-moment diagram of the confidence regions obtained from a single site. Based on the work of these researchers, the objective of this paper is to develop a graphical bivariate extension of the Hosking and Wallis (HW) GOF measure for selecting regional distributions. Where the original Hosking and Wallis GOF measure considered only the variability of the L-kurtosis, the new bivariate version introduced in this paper will consider variability in both L-skewness and L-kurtosis, as well as the correlation between the two. In addition, the new measure has a more direct visual interpretation on the L-moment diagram. First, the assumptions underpinning the index flood method will be discussed and used for developing the new bivariate GOF measure. Next, a series of Monte Carlo experiments will be conducted to assess the ability of the new measure to detect the correct distribution, especially when compared to the original HW measure. Finally, the new measure will be applied to two case studies; a homogeneous region of peak flow series from Italy, and a national study using pooling groups formed using annual maximum (AMAX) series of peak flow from gauging stations located in the United Kingdom.

2. A general framework for the index flood method

2.1. The statistical model of a homogeneous region

The starting point for the statistical model underpinning the index flood method is to assume that N sites form a homogeneous region, and that at each site n_i years of independent annual maximum (AMAX) data are available, from which the sample L-moment ratios can be derived. The definition of L-moments is well documented by *Hosking and Wallis* [1997] and others and therefore not repeated here. The observed r-th order L-moment ratio at the i-th site, $t_r^{(i)}$, is defined as the true, but unknown, value for the homogeneous region, τ_r, plus an error, ε_i, because the sample value is derived from a finite number (n_i) of observations, i.e.

$$t_r^{(i)} = \tau_r + \varepsilon_i, \ r = 2, 3, 4, \ i = 1, \dots N \tag{1}$$

This study will consider only $r = 2, 3, 4$ denoted L CV, L-skewness and L-kurtotis. The variance-covariances of the sample L-moment ratios are assumed inversely proportional to the sample size (record-length) [*Hosking*, 1986], and are given as a set of the covariance matrices $\mathbf{\Sigma}_{rq}$ with elements (i, j) defined as

$$\mathbf{\Sigma}_{rq,ij} = \mathrm{cov}\left(t_r^{(i)}, t_q^{(j)}\right) \tag{2}$$

where diagonal elements $(i = j)$ represent the variance of the r-th L-moment ratio at each of the N sites, and the non-diagonal elements represent the covariance between the r-th and q-th L-moment ratios at different sites $(i \neq j)$. Estimating the elements of these covariance matrices will be discussed later.

The regional estimate of the r-th order L-moment ratio is derived as a weighted average

$$t_r^R = \sum_{i=1}^{N} \omega_r^{(i)} t_r^{(i)} = \omega_{\mathbf{r}}^{\mathbf{T}} \mathbf{t}_r \tag{3}$$

where $\mathbf{t_r}$ is a vector containing the r-th order L-moment ratio for each of the N sites, and ω_r is a $n \times 1$ vector of weights assigned to each individual site in the region and which sum to one, i.e. $\sum \omega_r^{(i)} = 1$. The variance of the regional L-moment ratio of the r-th order (Eq. 3) is a scalar but can be expressed as a matrix multiplication as

$$\sigma_r^2 = \mathrm{var}(t_r^R) = \mathrm{var}\left(\omega^T \mathbf{t}_r\right) = \omega_r^T \mathbf{\Sigma}_{rr} \omega_r \tag{4}$$

where the covariance matrix $\mathbf{\Sigma}_{rr}$ is defined in Eq. (2). Similarly, the covariance between the regional L-moment ratios can be derived as

$$\sigma_{rq} = \mathrm{cov}\left(t_r^R, t_q^R\right) = \mathrm{cov}\left(\omega_r^T \mathbf{t}_r, \omega_q^T \mathbf{t}_q\right) = \omega_r^T \mathbf{\Sigma}_{rq} \omega_q^T \tag{5}$$

Using the method of Lagrange multipliers for constraint optimisation the set of weights which gives the minimal variance of the regional L-moment ratio can be derived from Eq.(4) as

$$\omega_r = \mathbf{\Sigma}_{rr}^{-1} \mathbf{i} (\mathbf{i}^T \mathbf{\Sigma}_{rr}^{-1} \mathbf{i})^{-1} \tag{6}$$

where $\mathbf{i}$ is a vector where all elements equal one. In the simplest case where no correlation exists between AMAX records across sites, and the samples are drawn from a homogeneous region, then the weights reduce to the record-length weighting procedure suggested by *Hosking and Wallis* [1997], and also used in this study. Next, the joint distribution of the L-skewness and the L-kurtosis is discussed, which will subsequently be used to developed a graphical version of the GOF measure presented by *Hosking and Wallis* [1993].

2.2. Bivariate distribution of L-skewness and L-kurtosis

In line with other researchers, notably *Hosking and Wallis* [1997] and *Liou et al.* [2008], it is assumed that the joint distribution of L-skewness and L-kurtosis is a bivariate normal distribution. As the regional L-moment ratios (t_3^R, t_4^R) are weighted averages of the at-site L-moment ratios, it follows by virtue of the central limit theorem that (t_3^R, t_4^R) is approximately distributed according to a bivariate normal distribution with a covariance matrix $\mathbf{\Omega}$ whose elements are defined by the expressions in Eqs.(4) and (5).

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_3^2 & \sigma_{34} \\ \sigma_{34} & \sigma_4^2 \end{bmatrix} \quad (7)$$

For selected one and two parameter distributions, *Hosking* [1986] provided analytical expressions for the variance and covariance of L-skewness and L-kurtosis, i.e. the elements of $\mathbf{\Sigma_{rq}}$ in Eq.(2), and thus by extension $\mathbf{\Omega}$ in Eq.(7) . However, for distributions of more than two parameters, the analytical expressions quickly become intractable; if they exist at all. Alternative analytical expressions can be derived using approximations, but they have generally been found to be inaccurate for sample sizes typically used in hydrology. Thus, a purely analytical approach to the specification of $\mathbf{\Omega}$ appears to have limited practical utility and will not be pursued further here. Other researchers have used extensive Monte Carlo simulations to derive approximations of the sampling variability of L-moment ratios [*Sankarasubramanian and Srinivasan*, 1999], but these are only available for a specific subset of distributions. The *Hosking and Wallis* [1993] goodness-of-fit measure, hereafter referred to as the HW measure, was developed specifically to enable assessment of the goodness-of-fit of several candidate three parameter distributions, and resorted to the use of Monte Carlo simulations from a 4-parameter Kappa distribution to obtain the variance of the regional L-moment ratios. This method has the advantage that it makes no explicit prior assumption on the type of distribution being assessed. *Wang and Hutson*

[2013] suggest that a well-defined GOF test based on a distribution-specific null-hypothesis might be more powerful than a more general model selection procedure such as the HW measure. However, the widespread use of the HW measure in the analysis of environmental extreme data is a testament to the usefulness of such a procedure for screening of noisy environmental data before committing to a particular distribution model; a point also emphasised by *Wang and Hutson* [2013].

3. Goodness of fit measures for regional distributions

3.1. The Hosking and Wallis Goodness-of-fit measure

Assuming a homogeneous region, the scatter of points on the L-moment diagram around the regional average values represents only sampling variability as per Eq. (1). The HW measure reduces the two-dimensional scatter (in both L-skewness and L-kurtosis directions) to a one dimension problem by assessing the bias corrected difference between the regional average L-kurtosis, i.e. t_4^R, with the notionally true value of L-kurtosis, τ_4^{DIST}, which can be calculated as a function of L-skewness for a range of distributions using the polynomial approximations provided by *Hosking and Wallis* [1997] in their Table A.3. A schematic representation of the measure, adopted from *Hosking and Wallis* [1993], is shown in Figure 1 in the left panel. Utilising that the L-moment ratios are approximately normally distributed, the HW measure takes the form of a univariate significance test

$$Z^{DIST} = \frac{\tau_4^{DIST} - t_4^R + B_4}{\sigma_4} \tag{8}$$

where B_4 is the bias correction of t_4^R, and σ_4 is the standard deviation of t_4^R which is assumed known. It then follows that Z^{DIST} is a standardised normal distribution, and *Hosking and Wallis* [1993] suggested using a 90% confidence level for accepting a particular distribution, i.e. if $|Z^{DIST}| \leq 1.64$, a distribution is considered an acceptable candidate distribution for the region. Although not strictly part of the HW method, the distribution with the Z^{DIST} score closest to zero is often chosen, but other distributions could be selected based on other considerations.

FIGURE 1 ABOUT HERE

The bias and standard deviation of the regional L-kurtosis value were obtained via Monte Carlo simulations. First a four parameter kappa distribution was specified using the first four L-moment ratios l, τ^R, τ_3^R and τ_4^R. From this kappa distribution a large number, N_{sim}, of homogeneous regions are generated, each representing AMAX data from $i = 1 \ldots N$ sites with individual record length n_i. For the m-th simulated region the regional average L skewness, $t_3^{[m]}$, and L kurtosis, $t_4^{[m]}$, are derived, and the bias B_4 and standard deviation σ_4 derived as

$$B_4 = N_{sim}^{-1} \sum_{m=1}^{N_{sim}} \left(t_4^{[m]} - t_4^R \right) \tag{9}$$

$$\sigma_4 = \left[(N_{sim} - 1)^{-1} \left\{ \sum_{m=1}^{N_{sim}} \left(t_4^{[m]} - t_4^R \right)^2 - N_{sim} B_4^2 \right\} \right]^{1/2} \tag{10}$$

Hosking and Wallis [1993] used $N_{sim} = 500$, and this was found to be an adequate number also for this study. The bias correction is likely to be important for short record lengths and for very skewed data series; see for example Figure 2.7 in *Hosking and Wallis* [1997].

Hosking and Wallis [1993] emphasised that the assumptions underpinning their GOF measure are unlikely to be met by real regions, and emphasised therefore that the measure should not be interpreted as a formal statistical test of goodness of fit. The same qualifier applies to the new bivariate extension presented in the next section.

3.2. A bivariate extension of the HW measure

The new bivariate extension of the HW measure proposed here is illustrated in the right panel on Figure 1. It is based on the interpretation of a confidence interval as a form of statistical test, and the approximate bivariate normal distribution of L-skewness and L-kurtosis as also utilised by *Liou et al.* [2008]. The confidence region for the bivariate distribution of L-skewness and L-kurtosis is defined by the measure T with a set of bias corrected regional L-moment ratios $\mathbf{t}^R$ against which the theoretical $\boldsymbol{\tau}$ L-moments are compared:

$$T = (\boldsymbol{\tau} - \mathbf{t}^R)^T \Omega^{-1} (\boldsymbol{\tau} - \mathbf{t}^R) \tag{11}$$

where $\boldsymbol{\tau}$ is the null hypothesis mean values of L-skewness and L-kurtosis. The components of the $\boldsymbol{\Omega}$ covariance matrix in Eq.(11) are estimated by means of N_{sim} synthetic samples generated from a kappa distribution with third and fourth L-moment equal to $\mathbf{t}^R$. In the case of perfectly independent observations and homogeneous regions the quantity $(N_{sim} - 2)/(2(N_{sim} - 1))\ T$ is distributed according to a F-distribution with $(2, N_{sim} - 2)$ degrees of freedom, i.e. $\frac{N_{sim}-2}{2(N_{sim}-1)} T \sim F_{2,N_{sim}-2}$. For N_{sim} sufficiently large the approximation $2F_{2,N_{sim}-2} \approx \chi^2_2$ holds, so that the quantity in Eq.(11) can be approximated by a chi-square distribution: $T \sim \chi^2_2$. The key assumptions behind this approximation are that the region under study is homogeneous, that a sufficient number of site-years are available and that a large number of N_{sim} synthetic samples are employed in the procedure to estimate $\boldsymbol{\Omega}$.

Utilising the same set of of Monte Carlo simulations deployed for calculating the variance of L-kurtosis in connection with the HW measure, the corresponding bias and variance of L-skewness, B_3 and σ_3^2, can be estimated using a similar set of equations as those used for L-kurtosis in Eqs. (9) and (10). The covariance between L-skewness and L-kurtosis can be estimated as

$$\sigma_{34} = (N_{sim} - 1)^{-1} \left\{ \sum_{m=1}^{N_{sim}} \left(t_3^{[m]} - t_3^R \right) \left(t_4^{[m]} - t_4^R \right) - N_{sim} B_3 B_4 \right\} \tag{12}$$

For a given significance level α, the $(1-\alpha)100\%$ confidence ellipse for the bias-corrected regional L-skewness and L-kurtosis, $\mathbf{t}_B^R = (t_3^R - B_3, t_4^R - B_4)$, can be constructed, using the estimated $\boldsymbol{\Omega}$ covariance matrix and the χ_2^2 approximation discussed above. The $(1 - \alpha)100\%$ confidence ellipse is plotted on the L-moment diagram along with the theoretical relationships between L-skewness and L-kurtosis as used previously in the calculation of the HW measure. If segments of the theoretical line of a specific distribution are located within the circumference of the confidence ellipse then this distribution should be considered as a candidate for the regional distribution. Taking $\tau^{DIST} = (\tau_3, \tau_4^{DIST}(\tau_3))$ to be the vector of possible (τ_3, τ_4) values for a distribution, if the minimum value of the Mahalanobis distance

$$\mathrm{D}^{DIST} = \left(\boldsymbol{\tau}^{DIST} - \mathbf{t}_B^R \right)^T \boldsymbol{\Omega}^{-1} \left(\boldsymbol{\tau}^{DIST} - \mathbf{t}_B^R \right) \tag{13}$$

is smaller than the critical $\chi^2_{2,1-\alpha}$ quantile, the distribution $DIST$ can be considered to be a possible candidate distribution at a significance level α. The final choice of distribution is determined by selecting from among all the theoretical curves, $(\tau_3, \tau_4^{DIST}(\tau_3))$ which lie within the $(1 - \alpha)100\%$ ellipsoid, the point with the shortest D^{DIST} value. As with the HW measure, other accepted distributions could be chosen if there were any particular reason to do so.

The concept is also illustrated on Figure 2 where the right panel shows the difference between the regional L-moment ratios and the theoretical lines representing various 3-parameter distributions. In Figure 2 the minimum distance is obtained for the GEV distribution, which is chosen as the regional distribution accordingly.

FIGURE 2 ABOUT HERE

Only distributions with theoretical lines located within the $(1-\alpha)\,100\%$ confidence region can be chosen as candidate distributions. Thus, in some cases the new bivariate measure may fail to accept any of the considered distributions as suitable for a particular region for the given confidence level.

4. Comparison of GOF measures

The performance of the new bivariate measure was evaluated and compared to the original HW measure using a set of Monte Carlo simulations and a significance level of $\alpha = 10\%$. Firstly, three different homogeneous regions were defined to mimic the regions used by *Hosking and Wallis* [1993] in their evaluation of the HW measure. Each region consists of $N = 21$ sites and each site has a record length of $n = 30$ years. Each of the three region is defined by a specified set of regional values for L CV and L-skewness ($(\tau = 0.10, \tau_3 = 0.05)$, $(\tau = 0.20, \tau_3 = 0.20)$ and $(\tau = 0.30, \tau_3 = 0.30)$), and one of four different parent distributions: Generalised Logistic (GLO), Generalised Extreme Value (GEV), Generalised Normal (GNO) or a Pearson Type III (PE3) distribution. The twelve resulting regions are listed in the first four columns in Tables 1 and 2. For each region, Monte Carlo simulations are used to generate 1000 replicas of the region from the specified parent distribution. For each one of the 1000 replica regions, both the original HW measure and the new bivariate measure were evaluated. Both the number of times a particular distribution was accepted as a parent distribution and the number of times each distribution was chosen as the best fitting distribution were recorded. The results are shown in Table 1 (original HW measure), and Table 2 (new bivariate measure).

TABLE 1 ABOUT HERE

TABLE 2 ABOUT HERE

The results obtained for the original HW measure in Table 1 are very similar to the results presented by *Hosking and Wallis* [1997], but differ in one aspect. By design the new bivariate version cannot choose a particular distribution without first accepting it as a possible candidate. However, this distinction was not enforced by *Hosking and Wallis* [1993] who reported that in some cases the GLO distribution had been chosen more times than it had been accepted. Thus, to enable a direct comparison of the two measures in this study, the original HW measure was only allowed to choose a distribution if this distribution had first been accepted by the same measure as a possible candidate.

For eleven of the twelve considered regions the new bivariate measure performs better than the original HW measure, meaning that the correct regional distribution is chosen more often by the new measure. While the differences are consistent they are not necessarily large, varying from 1% to 14%.

Given that no additional simulation effort is required when evaluating the GOF using the new measure compared to the original HW measure, the results shown in Table 1 and 2 suggest that the new bivariate measure should be used in preference to the original HW measure. However, it is necessary to discuss possible situations where the original HW measure appears to outperform the new bivariate measure.

Similarly to *Hosking and Wallis* [1993], the new bivariate version was found to accept the GLO distribution less frequent than other parent distributions. *Hosking and Wallis* [1993] suggested that this might be caused by underestimation of σ_4, but did not investigate further. An alternative explanation might relate to the asymmetric influence of the bias correction on the L-moment ratios. Figure 2.7 in *Hosking and Wallis* [1997] shows that the effect of the bias is more pronounced for higher value of L-skewness and L-kurtosis. As the GLO distribution is characterised by higher L-kurtosis values than the other 3 parameter distributions (the theoretical GLO lines is located above the other 3 parameter distribution lines in the L-moment diagram), sample values generated from a GLO distribution are therefore more likely to be moved even further up on the L-moment diagram as a result of the bias correction. Figure 3 shows the regional average L-skewness and L-kurtosis for five Monte Carlo generated regions from each of the three regions defined in Table 1. The plot on the left side shows the result when the AMAX events are generated from a GEV distribution, and the right side shows the results when generating AMAX events from a GLO distribution. The points represent the bias corrected values, and the arrows point to the location of the initial uncorrected sample values.

FIGURE 3 HERE

From the Figures it can be seen that samples generated from the GLO distribution are located higher on the L-moment diagram, and therefore are subject to a larger degree of bias correction. In some instances the bias correction is so large that the ellipse corresponding to the 90% confidence region (not shown) is moved so far that it no longer bisects the GLO line, suggesting that the GLO distribution is no longer considered suitable. This might be the reason why the GLO distribution is chosen less frequently than the other distributions, but it does not explain why the performance of the new bivariate measure is not as good as the original HW measure for the third region, consisting of a very skewed GLO distribution ($\tau_3 = 0.30$).

5. Assessing the effect of intersite correlation

The importance of intersite correlation between AMAX series from different sites within a region has been discussed by several authors, e.g *Stedinger* [1983], *Hosking and Wallis* [1988], *Kjeldsen and Jones* [2006], *Castellarin et al.* [2008]. From these studies it is well-understood that the effect of intersite correlation is primarily to increase the variance of the regional L-moment ratios. For the goodness-of-fit measures discussed in this study, the effect of increased variance of L-moment ratios should lead to a decrease in the ability of these measures to discriminate between distribution types.

A set of Monte Carlo simulations was used to investigate the effect of intersite correlation on the power of the original and new bivariate measure. The algorithm used for generating cross-correlated AMAX events from the N sites within a homogeneous region was adopted from *Hosking and Wallis* [1997], and also used by *Castellarin et al.* [2008] in a study of effects of intersite correlation on the performance of a measure for homogeneity. Repeated Monte Carlo simulations were conducted assuming an average cross correlation between 0.0 to 0.80 with a step-length of 0.10 (i.e. nine repetitions) using the same three regions as for the independent case discussed above, i.e. ($\tau = 0.10, \tau_3 = 0.05$), ($\tau = 0.20, \tau_3 = 0.20$) and ($\tau = 0.30, \tau_3 = 0.30$) assuming one of the four distributions: GLO, GEV, GNO or PE3. This experimental setup results in a total of 108 different regions. For each region, 1000 replica regions were generated, then the two GOF measures were evaluated, and the rate of choosing the correct regional distribution recorded. Figure 4 shows the percentage of the 1000 regions where the correct distribution type was selected by each of the two measures.

FIGURE 4 ABOUT HERE

For all four distributions (GLO, GEV, GNO and PE3) both measures are reasonably robust to the existence of intersite correlation when this is below about 0.40. For higher degrees of correlation the success rate of both measures starts to decline.

In general the new bivariate measure proposed in this study performs better than the original HW measure, except for the case of the GLO distribution for the region with high L-CV and L-skewness population parameters as already discussed. The performance of the two measures declines at a similar rate for higher intersite correlations: so for all levels of intersite variation the new bivariate measure is preferable.

6. Case studies

6.1. Example 1: Regional distribution of flood flow data in Central Italy

The new bivariate measure is applied to AMAX peak flow series from 22 flow gauging stations located in a Central part of Italy. These stations correspond to the catchments of region E described in *Castellarin* [2007], and have a record length between 15 and 74 years, with an average record length of 33.5 years. The original HW and the new bivariate measures are both applied to these series. Figure 5 shows the L-moments diagram with the ellipse corresponding to the 90% confidence region obtained from the bivariate measure.

FIGURE 5 ABOUT HERE

TABLE 3 ABOUT HERE

The results in Figure 5 and Table 3 show that for this dataset, both the GEV and the GNO distributions could be accepted as the regional distributions, but the the GEV distribution is the more likely candidate.

6.2. Example 2: A national distribution of UK flood data

The new new bivariate measure was applied on annual maximum (AMAX) series of peak flow from 564 rural catchments located through-out the UK. For each catchment, a site specific hydrological region (e.g. a pooling group) was formed based on hydrological similarity using the similarity measure developed by *Kjeldsen and Jones* [2009] and calculated using four different catchment descriptors: the catchment area (km^2), the standard annual average rainfall as measured between 1961 and 1990 (mm), an index of flood attenuation from upstream lakes and reservoirs, and the areal extent of floodplains in the upstream catchment defined by the 100-year flood level adopted from an existing national floodplain map.

A pooling group for each of the 564 catchments is formed by adding catchments from the entire database, starting with the most similar and continuing to add catchments until the total sum of AMAX events included in the pooling group exceeds 500. With an average record length of 36 years, a pooling group typically consists of between 12-15 catchments.

A first visual assessment of candidate distributions can be obtained by plotting the pairs of average L-skewness and L-kurtosis for each of the 564 pooling groups on a L-moment diagram as shown in Figure 6. From the L-moment diagram it is evident that the regional L-moment ratios generally plot between the two lines representing the GLO and the GEV distributions, both of which have previously been adopted as standard distributions for regional and pooled flood frequency estimation in the UK [*Natural Environment Research Council*, 1975; *Institute of Hydrology*, 1999]. Generally, the average correlation between the overlapping AMAX series within each pooling group is below 0.4 suggesting, with reference to the results in Figure 4, that the performance of the new GOF measure should not be unduly influenced by cross-correlation.

FIGURE 6 ABOUT HERE

A more quantitative assessment of the distribution type was undertaken by comparing the rate of accepting and choosing different distribution types using both the original HW measure and the new bivariate extension presented in this study. Applying the two measures to each of the 564 pooling groups, the percentage of pooling groups where a particular distribution is accepted and chosen is shown in Table 4.

TABLE 4 ABOUT HERE

The results in Table 4 show that both measures select the GLO distribution most frequently as the most suitable regional distribution. For the GNO and PE3 distributions, the selection rates are very similar for the two measures, and in any case much lower than for the GLO and GEV distributions. The new bivariate measure shows that the GLO and GEV distributions are accepted as candidate distributions an almost identical number of times, but that the GLO distribution is the preferred distribution as it is chosen more often than the GEV distribution. For 28 out of the 564 catchments ($\approx 5\%$), the new bivariate measure found that none of the four distributions adequately fitted the data. The original HW measure selects the GLO more often than the new bivariate measure, and thus gives more support to the GLO distribution as the default choice in UK catchments; for example if conducting a regional analysis in an ungauged catchment. The results shown in Table 4 combined with a visual inspection of the scatter of pooled L-moment ratios in Figure 6 suggests that the GLO distribution might not always be the best choice for UK catchments, and that the GEV distribution could also be considered in most cases.

7. Conclusion

This paper presented a new GOF measure for regional frequency distributions based on L-moment ratios and with a direct graphical interpretation using the L-moment ratio diagram. Based on a series of Monte Carlo simulations from homogeneous regions the new measure was found to provide a modest, but consistent, improvement in the ability to detect the underlying regional distribution when compared to the performance of the original one-dimensional GOF measure presented by *Hosking and Wallis* [1993]. This additional power was obtained utilising exactly the same set of Monte Carlo simulations as the original HW measure. Additional Monte Carlo simulations from regions where AMAX events are correlated across sites demonstrated that the performance of the new measure is sustained for regions with a level of correlation akin to that found in most UK pooling groups. As these pooling groups are made up of data from a relatively confined geographical region, it is expected that similar or less correlation is found in many other real world regions.

Further research should investigate the relatively poor performance of the new measure for detecting the GLO distribution in regions characterised by high values of L-skewness. Another important topic to investigate is if the more generalised set of weights in Eq. (6) can be developed to improve performance in cross-correlated and heterogeneous regions.

Acknowledgments. The authors would like to thank Attilio Castellarin (attilio.castellarin@unibo.it) and the UK measuring authority's HiFlows-UK database (http://www.ceh.ac.uk/data/nrfa/peakflow_overview.html) for providing access to the Italian and UK flood flow data, respectively. Three anonymous reviewers are acknowledged for helpful criticism of an earlier version of the manuscript.

References

Castellarin, A. (2007), Probabilistic envelope curves for design flood estimation at ungauged sites, *Water Resources Research*, *43*(4), 1944–7973.

Castellarin, A., D. Burn, and A. Brath (2008), Homogeneity testing: How homogeneous do heterogeneous cross-correlated regions seem?, *Journal of Hydrology*, *360*(1), 67–76.

Hosking, J. R. (1990), L-moments: analysis and estimation of distributions using linear combinations of order statistics, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 105–124.

Hosking, J. (1986), The theory of probability weighted moments, *IBM Res. Rep. RC12210, IBM, Yorktown Heights, NY.*

Hosking, J. R. M., and J. R. Wallis (1997), *Regional frequency analysis: an approach based on L-moments*, Cambridge University Press.

Hosking, J., and J. Wallis (1993), Some statistics useful in regional frequency analysis, *Water Resources Research*, *29*(2), 271–281.

Hosking, J., and J. Wallis (1988), The effect of intersite dependence on regional flood frequency analysis, *Water Resources Research*, *24*(4), 588–600.

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

Kjeldsen, T. R., and D. A. Jones (2006), Prediction uncertainty in a median-based index flood method using L moments, *Water resources research*, *42*(7). *W07414*.

Kjeldsen, T. R., and D. A. Jones (2009), A formal statistical model for pooled analysis of extreme floods, *Hydrology Research*, *40*(5), 465–480.

Kumar, R., C. Chatterjee, S. Kumar, A. Lohani, and R. Singh (2003), Development of regional flood frequency relationships using L-moments for Middle Ganga Plains Subzone 1 (f) of India, *Water Resources Management*, *17*(4), 243–257.

Liou, J.-J., Y.-C. Wu, and K.-S. Cheng (2008), Establishing acceptance regions for L-moments based goodness-of-fit tests by stochastic simulation, *Journal of Hydrology*, *355*(1), 49–62.

Madsen, H., C. P. Pearson, and D. Rosbjerg (1997), Comparison of annual maximum series and partial duration series methods for modeling extreme hydrological events 2. Regional modeling, *Water Resources Research*, *33*(4), 759–769.

Mkhandi, S., R. Kachroo, and T. Gunasekara (2000), Flood frequency analysis of southern Africa: II. identification of regional distributions, *Hydrological sciences journal*, *45*(3), 449–464.

Natural Environment Research Council (1975), *Flood Studies Report, (5 Volumes)*, Natural Environment Research Council, London, UK.

Peel, M. C., Q. Wang, R. M. Vogel, and T. A. McMahon (2001), The utility of L-moment ratio diagrams for selecting a regional probability distribution, *Hydrological sciences journal*, *46*(1), 147–155.

Sankarasubramanian, A., and K. Srinivasan (1999), Investigation and comparison of sampling properties of L-moments and conventional moments, *Journal of Hydrology*, *218*(1), 13–34.

Stedinger, J. R. (1983), Estimating a regional flood frequency distribution, *Water Resources Research*, *19*(2), 503–510.

Vogel, R. M., and N. M. Fennessey (1993), L moment diagrams should replace product moment diagrams, *Water Resources Research*, *29*(6), 1745–1752.

Vogel, R. M., W. O. Thomas Jr, and T. A. McMahon (1993), Flood-flow frequency model selection in Southwestern United states, *Journal of Water Resources Planning and Management*, *119*(3), 353–366.

Wang, D., and A. D. Hutson (2013), Joint confidence region estimation of L-moment ratios with an extension to right censored data, *Journal of Applied Statistics*, *40*(2), 368–379.

Wu, Y.-C., J.-J. Liou, Y.-F. Su, and K.-S. Cheng (2012), Establishing acceptance regions for L-moments based goodness-of-fit tests for the Pearson type III distribution, *Stochastic Environmental Research and Risk Assessment*, *26*(6), 873–885.

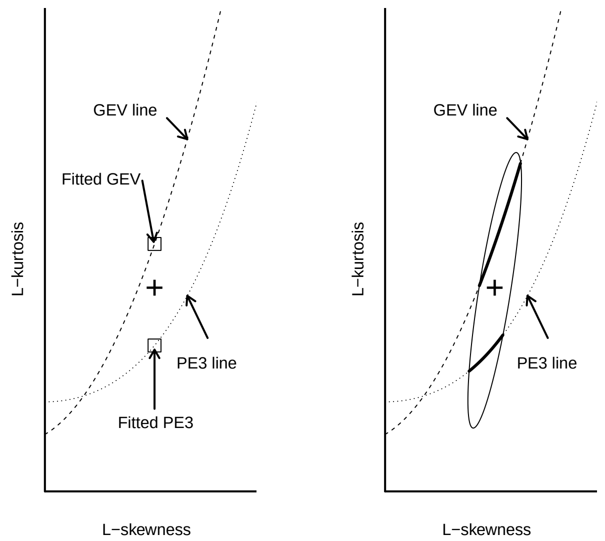

Figure 1. Explanatory sketches for HW GOF measure (left), adopted from *Hosking and Wallis* [1993], and (right) the new bivariate GOF measure. In the right panel, the bold line segments located with the circumsphere of the ellipsoid are within the 90% confidence region of the regional L-moment ratios, and thus potentially accepted as regional distributions. In both figures the bold cross represents the average sample values of L-skewness and L-kurtosis.

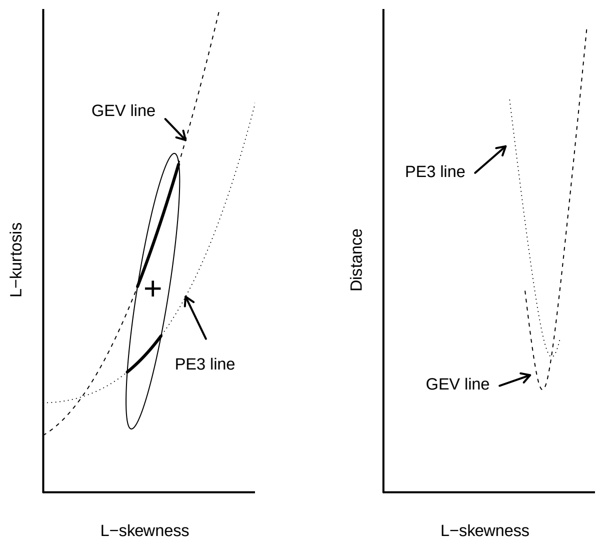

Figure 2. Accepted candidate distributions are identified where segments of the theoretical distribution lines are located within the confidence region, shown as bold line-segments on the left figure. The final choice of distribution is based on the minimum distance between regional L-moment ratios and the theoretical distribution within the region of acceptance as shown on the right figure.

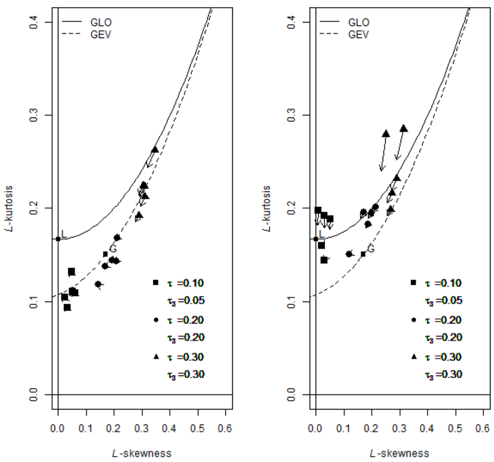

Figure 3. Regional estimates of L-skewness and L-kurtosis using AMAX data generated from a GEV distribution (left) and a GLO distribution (right) for three different homogeneous regions. The points represent the bias correct values of t_3^R and t_4^R and the arrows point to the initial uncorrected sample values.

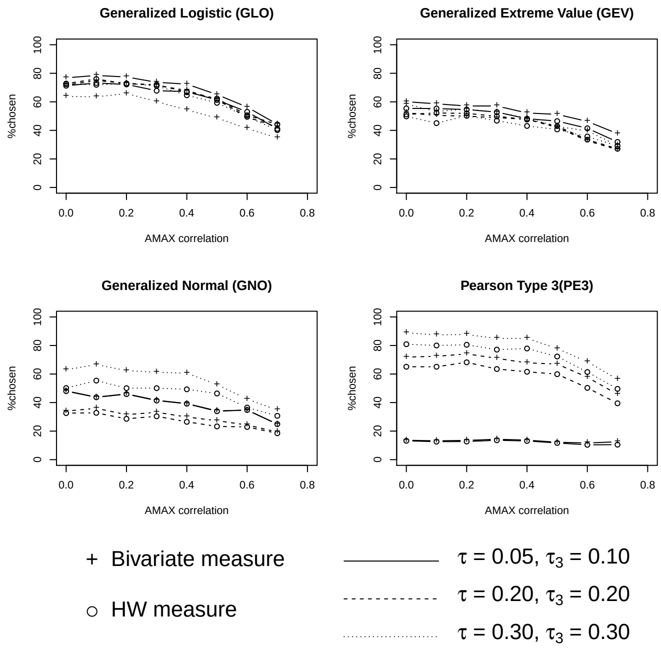

Figure 4. Comparison of the performance of the original HW and the new bivariate GOF measures in three different regions shown as a function of intersite correlation between AMAX series within each region.

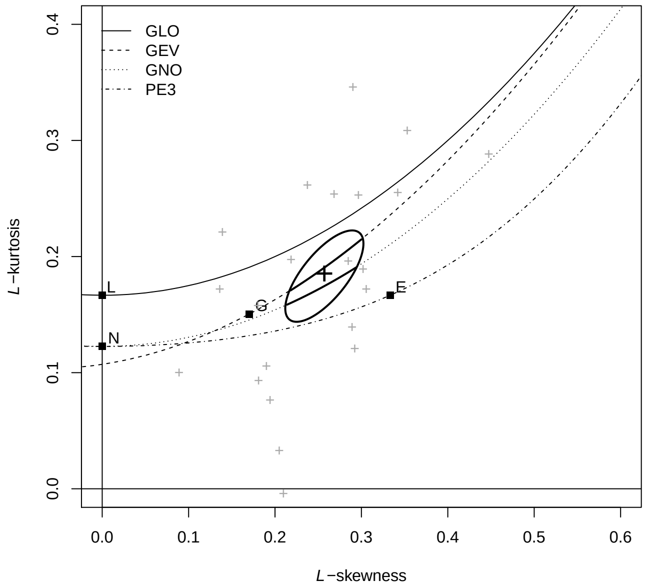

Figure 5. L-moment diagram showing L-moment ratios for the 22 Italian catchments and the corresponding 90% confidence region. The thick line segments represent the segments of the theoretical distributions that fall within the 90% confidence region.

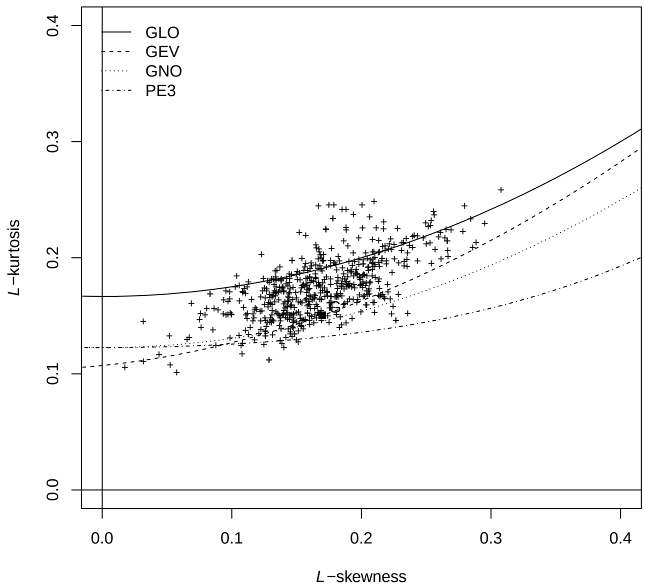

Figure 6. Regional L-moment ratios for each of the 564 UK pooling groups plotted on a L-moment diagram.

Table 1. Simulation results for the original HW GOF measure showing percentage of simulations where a particular distribution is accepted and chosen.

			% Accepted				% Chosen			
τ	τ_3		GLO	GEV	GNO	PE3	GLO	GEV	GNO	PE3
		GLO	**75**	3	12	10	**73**	0	9	0
		GEV	2	**87**	79	81	2	**52**	24	14
0.1	0.05	GNO	9	81	**88**	88	7	35	**45**	13
		PE3	7	83	88	**88**	6	35	41	**16**
		GLO	**78**	26	15	4	**72**	12	1	0
		GEV	34	**93**	86	51	17	**51**	19	13
0.2	0.2	GNO	15	90	**92**	73	5	35	**33**	27
		PE3	1	52	72	**89**	0	8	21	**65**
		GLO	**84**	54	17	1	**73**	13	1	0
		GEV	74	**94**	67	10	38	**47**	14	1
0.3	0.3	GNO	31	88	**95**	34	5	33	**54**	9
		PE3	0	6	38	**93**	0	0	14	**81**

Table 2. Simulation results for the new bivariate GOF measure showing percentage of simulations where a particular distribution is accepted and chosen.

			% Accepted				% Chosen			
τ	τ_3		GLO	GEV	GNO	PE3	GLO	GEV	GNO	PE3
		GLO	**85**	9	25	22	**79**	0	9	0
		GEV	5	**96**	88	90	1	**56**	24	15
0.1	0.05	GNO	18	92	**96**	96	6	34	**45**	14
		PE3	13	94	95	**95**	6	36	41	**16**
		GLO	**82**	38	24	8	**73**	14	1	0
		GEV	31	**95**	93	67	11	**50**	23	16
0.2	0.2	GNO	12	90	**96**	84	3	30	**37**	30
		PE3	0	50	75	**95**	0	5	20	**72**
		GLO	**84**	65	24	2	**64**	24	2	0
		GEV	54	**94**	75	15	20	**54**	25	2
0.3	0.3	GNO	11	66	**96**	47	1	21	**64**	14
		PE3	0	1	23	**95**	0	0	8	**87**

Table 3. Comparison of the original HW and the new bivariate GOF measures on a homogeneous region consisting of 22 Italian catchments. Numbers indicate values of the GOF measures and bold fonts highlight the chosen distributions.

	GLO	GEV	GNO	PE3
Original HW	1.86	**0.10**	-0.69	-2.14
New Bivariate	-	**0.32**	0.39	-

Table 4. Comparison of the new bivariate GOF measure and the original HW measure. Numbers represent percentages of the 564 pooling groups accepted and chosen by the new measure. Numbers in () refer to the corresponding results obtained using the original HW measure.

	GLO	GEV	GNO	PE3
Accepted	**74(70)**	79(67)	71(58)	50(36)
Chosen	**49(53)**	31(27)	12(11)	4(4)