Evaluating the uncertainty and reliability of standardized indices

L. Vergni, F. Todisco and F. Mannocchi

ABSTRACT

Standardized indices are widely used in the spatio-temporal monitoring of several hydrological variables. The estimation of these indices is affected by uncertainty which depends on the methods adopted for their quantification and on the characteristics (i.e., size and variability) of the available sample of observations. In this paper various uncertainty measures, applicable to any kind of standardized index, are proposed. These measures derive from bootstrap-based confidence intervals expressed in years of return period and are effective for assessing both the uncertainty and the reliability of the index estimate. In the illustrative case study the indices considered are the Standardized Precipitation Index and the Standardized Precipitation Evapotranspiration Index. Their time series have been quantified by both nonparametric and parametric approaches, using the weather data of a single station in central Italy. For the parametric approach, two possible types of distributions have been assumed for each index. The results are discussed in order to analyze the behavior of the proposed uncertainty measures in relation to: sample size, type of approach (parametric or nonparametric), time scale, type of standardized index, and type of anomaly (excess or deficit).

Key words | bootstrap, confidence intervals, L-moments, parametric and nonparametric drought indices, return period

L. Vergni (corresponding author)
F. Todisco
F. Mannocchi
Department of Agricultural, Food and Environmental Sciences,
University of Perugia,
Borgo XX Giugno 74,
Perugia 06121,
Italy
E-mail: *lorenzo.vergni@unipg.it*

INTRODUCTION

The spatio-temporal variability of hydrological variables is often described by defining and quantifying specific standardized indices (Mishra & Singh 2010). Meteorological standardized indices, such as the Standardized Precipitation Index (SPI) (Mckee *et al.* 1993) and the Standardized Precipitation Evapotranspiration Index (SPEI) (Vicente-Serrano *et al.* 2010), represent the most common applications, but there are also examples of indices relating to more complex hydrological variables such as the streamflow (Standardized Discharge Index (SDI), Nalbantis & Tsakiris 2009) or the runoff (Standardized Runoff Index, Shukla & Wood 2008). Recently, multivariate standardized indices have also been proposed (Hao & AghaKouchak 2013). Standardized indices express any observed value of the reference hydrological variable (i.e., precipitation, climatic water balance, streamflow, etc.) in terms of standard normal deviate, thus allowing a rapid and effective identification and quantification of the anomalies with respect to the 'normal' condition. The main advantage of standardization lies in the possibility of comparing the anomalies both in space and time, independently of any climatic or seasonal difference.

In the traditional parametric approach, the fundamental initial step in the quantification of a standardized index is the choice of a probability distribution that is suitable for describing the specific reference variable. Sometimes this step does not have a clear and univocal solution (Guttman 1999), thus introducing a first problematic issue: indeed, if different distributions are used, a single value of the reference variable corresponds to different standardized values

doi: 10.2166/nh.2016.076

(with a consequent reduction of the comparability feature). Several studies have investigated this aspect, mainly in relation to the SPI (Guttman 1999; Kumar *et al.* 2009; Blain 2011; Angelidis *et al.* 2012) and SPEI (Stagge *et al.* 2015).

To avoid these problems and minimize the uncertainty associated with the selection and estimation of parametric distribution functions, it is possible to derive the probability distribution of a certain hydrological variable by a nonparametric approach. This is typically obtained by using an empirical distribution function (Farahmand & AghaKouchak 2015) or a nonparametric kernel density estimator (Kumar *et al.* 2016).

Whatever the approach is, the values of the standardized indices are affected by uncertainty because the corresponding cumulative probabilities are estimated on the basis of a limited sample of observations. The quantification of such type of uncertainty can be very useful for practical applications (water resources management, prevention, etc.), because it makes it possible to define the reliability (or reasonableness) of the index estimate and of the dependent actions. However, this topic has rarely been addressed in the literature. No studies have discussed the uncertainty associated with nonparametric standardized indices. For the parametric approach, no studies are available for SPEI or RDI, and the only example for SDI can be found in Hong *et al.* (2015). More studies can be found for the SPI. For example, Naumann *et al.* (2012) and Hu *et al.* (2015) have applied bootstrap-based procedures to describe and quantify the confidence intervals (CI) associated with SPI estimates. Vergni *et al.* (2015) performed a more comprehensive bootstrap-based analysis to evaluate, for central Italy, the variability of the mean size of the CI of parametric SPI varying the time scale, the length of the time series, and the underlying distribution. In that analysis, the size (ΔS) of the CI was simply expressed in standard deviate units. Of course, this method is useful for comparing the contribution of different factors to the overall uncertainty, but it is not very effective for understanding the reliability of the estimate.

For this reason, a first objective of the present paper is to improve the methodology illustrated in Vergni *et al.* (2015) by proposing some more effective measures of both uncertainty and reliability. These measures are defined on the basis of CI expressed in years of return period. Second, the paper illustrates a possible method to extend this type of uncertainty analysis to nonparametric standardized indices.

The illustrative case study refers to SPI and SPEI indices, whose time series were calculated for a single station in central Italy. The results are discussed in order to analyze the variability of the proposed measures of uncertainty in relation to: sample size, type of approach (parametric or nonparametric), time scale, type of standardized index, and type of anomaly (excess or deficit).

MATERIALS AND METHODS

Standardization of hydrological variables

The first step in the calculation of a generic standardized index S consists in the quantification, for each year i ($i = 1, \ldots, N$), of the values, $X_{i,j}^k$, i.e., the values describing the reference variable X in relation to a particular period of interest j (typically the month), and to k-1 past consecutive months (k is called time scale). Then, a parametric or nonparametric probability function is fitted to homogenous (by j and k) samples of the values $X_{i,j}^k$. The fitting is performed for each calendar month, in order to take into account the climatic differences due to seasonality. After this step, the cumulative probability of an observed value $X_{i,j}^k$ can be estimated and, by an equiprobability transformation (Abramowitz & Stegun 1965), it is possible to derive the corresponding standard normal deviate, which represents the standardized index S.

The standardized indices considered in this paper are the SPI (McKee *et al.* 1993) and the SPEI (Vicente-Serrano *et al.* 2010). For SPI, the reference hydrological variable X is the cumulative precipitation, while for SPEI, X is the climatic water balance (difference between precipitation and reference evapotranspiration ET0).

In the parametric calculation, the underlying distributions assumed for the cumulative precipitation (SPI) are both the two-parameter gamma (GAM) and the Pearson type III (PE3) distributions, since they are those most used in the literature (Blain 2011). For the climatic balance (SPEI), the analysis is based on both the generalized logistic (GLO) distribution and the generalized extreme value (GEV) distribution. The GLO was considered in the original algorithm proposed by Vicente-Serrano *et al.* (2010), but Stagge *et al.* (2015), in a

recent analysis for Europe, showed that the goodness-of-fit of the GEV can be better than that of the GLO. The parameters of the distributions selected were estimated from L-moments (Hosking 1990). For both indices, each computation routine was performed assuming a unique probability distribution with parameters depending on the month of the year. As an alternative, it would be possible to assume a probability distribution dependent on the month of the year.

The nonparametric calculation of both SPI and SPEI was performed by fitting the empirical distributions of X with a kernel density function $\hat{f}_k(x)$:

$$\hat{f}_k(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\frac{(x - x_i)}{h} \quad (1)$$

where $K(x)$ is the smoothing kernel function, N is the sample size and h is the bandwidth that controls the variance of the Kernel function. A Gaussian kernel was used and the bandwidth was estimated by a direct plug-in selector (Wand & Jones 1995). For precipitation (SPI), kernel densities were estimated with a zero lower bound.

This type of nonparametric approach was preferred over that based on the empirical cumulative distribution functions (ECDF) (Farahmand & AghaKouchak 2015), which are less suitable for the calculation of the bootstrap CI. Concerning this, more details will be provided in the specific section.

Both SPI and the SPEI were analyzed at the 1-, 3-, 6- and 12-month time scales.

Evaluation of the goodness-of-fit

Two criteria were used to evaluate the goodness-of-fit of the different distributions (parametric and nonparametric). A first evaluation was based on the root-mean square error (RMSE) computed as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}} \quad (2)$$

where x_i and $\hat{x}_i$ are, respectively, the observed and estimated quantiles of X. $\hat{x}_i$ was estimated from the fitted cumulative distribution functions, using in input the empirical cumulative probabilities computed by the Gringorten plotting position, whose calculation details can be also found in Farahmand & AghaKouchak (2015).

A second criterion was based on the normality test (Wu *et al.* 2007) applied to the resulting time series of the standardized indices. According to Wu *et al.* (2007), the series of a generic standardized index S are considered non-normal when the following three criteria are met: (1) Shapiro-Wilk statistic, W, is less than 0.96; (2) p-value associated to the W statistic is less than 0.1; and (3) the absolute value of the median S is greater than 0.05. The rationale behind this test is that standardized indices were designed to represent drought and wet anomalies in a similar way, and this requirement can be achieved only by normally distributed series. Non-normality frequently occurs when the distribution selected exhibits a poor fitting (Vergni *et al.* 2015).

Recently, Stagge *et al.* (2015) have applied the normality test to SPI and SPEI series computed on the basis of reanalysis data in Europe to assess the goodness-of-fit of several parametric distributions. They demonstrated that the normality test produces results similar to (or even more restrictive than) those obtained by the more traditional Kolmogorov–Smirnov and Anderson–Darling tests.

CI of standardized indices

The CI for a generic standardized index S can be determined by using a bootstrap resampling technique. The theoretical details of the bootstrap approach can be found in Efron & Tibshirani (1993), and an example of its application in the estimation of the CI for parametric SPI is available in Hu *et al.* (2015). A brief description of the steps required for a generic index S is provided here in the following.

Input data consist of N values $X_{i,j}^k$, i ($i = 1,\ldots, N$) of the reference hydrological variable X for a given month j, time scale k, and location.

Then, M bootstrapped samples of size N with replacement from the original N values $X_{i,j}^k$ are generated. The further sequence of operations varies according to the type of calculation of the index S (parametric or nonparametric).

For parametric standardized indices, the steps are the following:

(1) Choice of the probability distribution to be fitted to X_j^k and estimation of the distribution parameters for each

of the M bootstrapped samples, to obtain a family of M possible populations.

(2) Calculation of M values of cumulative probability $P(X_{i,j}^k)$ for each single value $X_{i,j}^k$.

(3) Transformation of the M values of $P(X_{i,j}^k)$ in M values of the corresponding standardized index S through an equiprobability transformation (Abramowitz & Stegun 1965).

(4) Analysis of the ECDF of the M values of S to identify the $(\alpha/2)^{th}$ and $(1\text{-}\alpha/2)^{th}$ percentiles, which are assumed, respectively, as the lower (S_{min}) and upper (S_{max}) limits of the CI for S with probability $P = (1\text{-}\alpha)$.

For nonparametric standardized indices, steps 1 and 2 are replaced by the following ones:

(1) Choice of a nonparametric method for the estimation of the cumulative probability of the variable X_j^k.

(2) Calculation of M values of the nonparametric cumulative probability $F(X_{i,j}^k)$ for each single value $X_{i,j}^k$.

The other steps are analogous to those of the parametric approach but it is necessary to replace $P(X_{i,j}^k)$ with $F(X_{i,j}^k)$.

For this type of application, we found that a kernel density function (Equation (1)) works better than an ECDF, in particular for the determination of the CI associated with the extreme observations. In fact, using an ECDF, the minimum and maximum observations have constant cumulative probabilities in all the bootstrapped samples, thus not allowing construction of their CI. Instead, the kernel density function provides a smoothed and continuous probability distribution whose characteristics (e.g., the bandwidth) are different in each bootstrapped sample.

All the analyses in this paper were based on CI with $P = 90\%$ and on $M = 1{,}000$.

The sizes of the CI associated with the point estimates of a standardized index have been used in this paper to derive some measures of uncertainty and reliability, whose definitions and calculation details are provided in the next section.

Measures of uncertainty and reliability

The basic measure of uncertainty is represented by the size, ΔS, of the CI at a given probability P:

$$\Delta S = |S_{\max} - S_{\min}| \tag{3}$$

Of course, as also illustrated in Vergni *et al.* (2015), the greater is ΔS, the greater is the uncertainty of a given point estimate of the standardized index S. However, in order to obtain a more effective evaluation of the reliability and reasonableness associated with S, it is worthwhile to express the size of the CI in years of return period. In general, the return period is the expected time between hazard events of a certain magnitude S. It can be defined for both the non-exceedance and exceedance probabilities associated with the event magnitude (Tallaksen & van Lanen 2004). The return period T_D (months or years) for the non-exceedance probability is calculated as:

$$T_D = \frac{1}{P(X_{i,j}^k)} = \frac{1}{\Phi^{-1}(S)} \tag{4}$$

where φ^{-1} is the inverse cumulative normal standard distribution.

The return period T_W (months or years) for the exceedance probability is calculated as:

$$T_W = \frac{1}{1 - P(X_{i,j}^k)} = \frac{1}{1 - \Phi^{-1}(S)} \tag{5}$$

T_D is suitable to describe the risk associated with the occurrence of drought anomalies (i.e., $S \leq 0$), while T_W is suitable to describe the risk associated with the occurrence of wet anomalies (i.e., $S \geq 0$). Of course, Equations (4) and (5) can also be applied for the nonparametric approach, replacing $P(X_{i,j}^k)$ with $F(X_{i,j}^k)$.

On the basis of the above-defined return periods, two measures of uncertainty/reliability are proposed in this paper. The first is represented by the difference, ΔT, between the return periods associated with the limits S_{min} and S_{max} of the CI; ΔT can be calculated by the following equations:

$$\Delta T = T_D(S_{\min}) - T_D(S_{\max}) \text{ for } (S_{\min} <= 0 \text{ and } S_{\max} <= 0) \tag{6a}$$

$$\Delta T = T_W(S_{\max}) - T_W(S_{\min}) \text{ for } (S_{\min} >= 0 \text{ and } S_{\max} >= 0) \tag{6b}$$

ΔT is not calculated for the events characterized by opposite upper and lower limits of the CI (i.e., $S_{min} < 0$

and $S_{max} > 0$). In fact, under these circumstances, the return periods associated with the lower and upper limits are not comparable (the former is the return period for the non-exceedance probability, the latter is the return period for the exceedance probability). However, the events having $S_{min} < 0$ and $S_{max} > 0$ correspond to near-normal conditions ($S \approx 0$), for which the uncertainty is lower and the practical interest is limited.

As for ΔS, the greater the ΔT, the greater is the uncertainty. However, thanks to its unit of measure, ΔT is more informative than ΔS. Of course, it is not possible to define absolute ΔT thresholds, beyond which the estimate should be considered unacceptable or unreliable. Indeed, this depends on several factors, such as the return period associated with the point estimate S and the aims of the monitoring (i.e., the maximum admissible risk), as well as subjective considerations.

The second uncertainty measure proposed expresses the uncertainty in relative terms, taking into account the relative magnitude of S_{min} and S_{max}. It is given by the ratio, T_{ratio}, of the return periods corresponding to S_{min} and S_{max} and it can be formally defined as follows:

$$T_{ratio} = \frac{T_D(S_{\min})}{T_D(S_{\max})} \text{ for } (S_{\min} \leq 0 \text{ and } S_{\max} \leq 0) \quad (7a)$$

$$T_{ratio} = \frac{T_W(S_{\max})}{T_W(S_{\min})} \text{ for } (S_{\min} \geq 0 \text{ and } S_{\max} \geq 0) \quad (7b)$$

As for ΔT, T_{ratio} is also not calculated for events characterized by opposite upper and lower limits of the CI. The more T_{ratio} approaches 1, the lower is the uncertainty.

T_{ratio} is particularly suitable for evaluating the reliability of the index estimate. It can, in fact, be assumed that a certain estimate of the standardized index S is unreliable if the corresponding T_{ratio} is greater than a subjective threshold τ. In the case study illustrated in this paper, a $\tau = 3$ was adopted. This relatively low (i.e., severe) threshold was chosen to obtain a significant number of unreliable estimates also under conditions characterized by the lowest overall uncertainty (i.e., for analysis based on very long time series of observations).

Case study

The weather data employed for the case study are those from the Terni station (central Italy), for which daily records of minimum and maximum temperature and rainfall during the 1951–2010 period were available. This station was selected, among others, for the length and completeness of its data series. The climate is typically Mediterranean, with a dry season (mean ET0 > mean P) from April to September and a wet season from October to March, as shown in Figure 1.

The original daily data were used in input to the LARS-WG 5.0 weather generator (Semenov & Barrow 2002) in order to obtain daily time series of different lengths (30, 60, 90, 120, 150, and 180 years), while maintaining the statistical properties of the original data. This was done in order to analyze the effect of the length of the available time series on the uncertainty associated with the standardized indices.

All the daily precipitation and temperature data were aggregated at the monthly time scale. Monthly estimates of reference evapotranspiration (required for SPEI calculation) were then obtained from monthly precipitation and monthly mean minimum and maximum temperature, by applying a modified version of the Hargreaves and Samani equation (Droogers & Allen 2002).

Months with zero rainfall are not very common (1.4%), and in most cases they occur in summer (34% and 21% in August and July, respectively). Zero rainfall values are not present at time scales greater than 2 months. As indicated by some authors (e.g., Wu *et al.* 2007), the presence of zero values is the primary reason for non-normal SPI series.

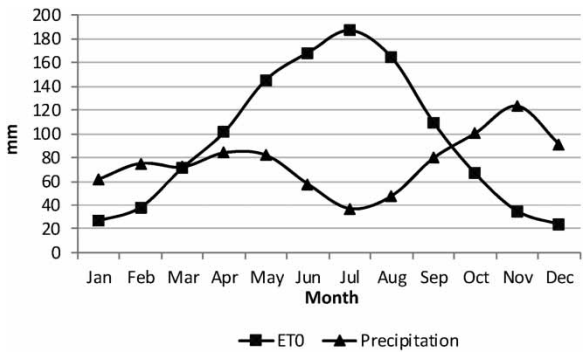

Figure 1 | Mean monthly values of reference evapotranspiration (ET0) and precipitation for Terni (1951–2010).

However, in the authors' opinion, the low occurrence of zero values in the case study considered should not have significant consequences on the results.

Taking into account the number of indices (2), time scales (4), sample sizes (6) and calculation methods (3, of which, 1 nonparametric and 2 parametric with different underlying distributions), 144 different series of standardized indices were considered in the case study.

RESULTS AND DISCUSSION

Goodness-of-fit results

Table 1 summarizes the goodness-of-fit (in terms of mean RMSE) of nonparametric and parametric distributions for the case study, varying the hydrological variable, the time scale, and the sample size. For the sake of simplicity, only the two extreme sample sizes (30 and 180 years) are reported. The fitting was performed for each calendar month, therefore Table 1 shows average RMSE results.

From the results of Table 1, it is evident that the nonparametric approach provides a better goodness-of-fit than parametric approaches. This result is expected since the kernel density function is directly fitted to the sample observations, while in the parametric method, the sample characteristics are 'filtered' by the estimated distribution parameters. Of course, for sample sizes smaller than those here considered, the nonparametric approach could also lead to misleading probabilities estimations.

The comparison between the parametric distributions shows that the PE3 performs better than GAM for precipitation, probably due to the fact that PE3 has one more parameter than GAM. For the climatic balance, the performance of GEV is slightly better than that of GLO.

The 144 series of standardized indices were also analyzed by applying the normality test (Wu *et al.* 2007) for each calendar month. The normality assumption is not rejected for all the series derived from the nonparametric calculation, confirming the good performance of this type of approach. For the series derived from the parametric distributions, the overall percentages of rejection of the normality assumption are: 5.9% for GAM, 1.7% for PE3, 1.4% for GLO, and 1.7% for GEV. These values are consistent with those obtained by similar studies (Stagge *et al.* 2015) and confirm that all the selected distributions provide an adequate description of the reference variables underlying the SPI and SPEI indices. A slightly worse performance is obtained by the adoption of the GAM for SPI. For this distribution, a detailed analysis of the results by time scale (not shown) revealed that most of the rejections are associated with the 1-month time scale. By excluding this time scale, the performance of the GAM is similar to that obtained by the PE3 distribution. This lower performance of the GAM distribution for short accumulation periods is consistent with Stagge *et al.* (2015).

No relationship was found between the results of the normality test and the length of the time series or the calendar month.

Evaluation of the uncertainty on the basis of the bootstrap CI

Some examples of the CI obtained for SPI and SPEI using the nonparametric and parametric approaches (in the latter case assuming different underlying distributions) are provided in Figure 2. For the sake of simplicity these examples only

Table 1 | Average RMSE of different probability distributions for different hydrological variables, time scales and for sample sizes of 30 and 180 years (in brackets)

		Time scale *k* (months)			
Hydrological variable	**Distribution**	**1**	**3**	**6**	**12**
Precipitation (SPI)	Nonparametric (Gaussian kernel)	**6.1 (3.6)**	**10.3 (6.5)**	**14.9 (9.6)**	**20.7 (13.6)**
	gamma	9.1 (5.8)	18.1 (7.6)	20.4 (11.9)	29.9 (13.2)
	Pearson type III	8.2 (3.8)	15.1 (6.9)	18.5 (11.0)	24.6 (13.1)
Climatic balance (SPEI)	Nonparametric (Gaussian kernel)	**6.4 (3.7)**	**10.7 (6.7)**	**14.9 (9.7)**	**21.5 (13.9)**
	GLO	8.4 (6.8)	15.4 (9.9)	18.4 (14.7)	28.3 (17.4)
	GEV	8.1 (4.0)	14.6 (7.1)	18.4 (12.0)	27.4 (14.4)

Bold values indicate the best performing distribution for each combination of variable and time scale. Weather data (1950–2010) of the Terni station.

refer to the 6-month time scale (SPI6 and SPEI6) and to the extreme sample sizes (30 and 180 years) among those considered. Moreover, in order to enhance the readability, only the first 6 years (72 months) of the time series are shown.

Analyzing Figure 2, the following general comments can be made:

(1) The point estimates (full lines) of the standardized indices obtained from different approaches are rather similar, particularly for the larger dataset.
(2) As expected, whatever the calculation approach is, the uncertainty is much lower when the analysis is based on relatively larger datasets.
(3) The uncertainty bounds obtained from different calculation approaches are almost similar for normal conditions and for moderate drought or wet conditions. As the severity of the events increases (e.g., values around the 7th and 61st months of the series), the uncertainty increases and also the differences among different approaches become more evident (particularly for the smaller dataset). This behavior has been already observed in Vergni *et al.* (2015) and it is due to the fact that the extreme values tend to intercept the tails of the bootstrapped probability functions, with a consequent explosion of the uncertainty.

This last result underlines the importance of an adequate sample size to obtain a reliable analysis of extremes. For such type of events, the inconsistency issues deriving from the choice of different suitable parametric distributions, affect, not only in the point estimates, but also in the corresponding measures of uncertainty.

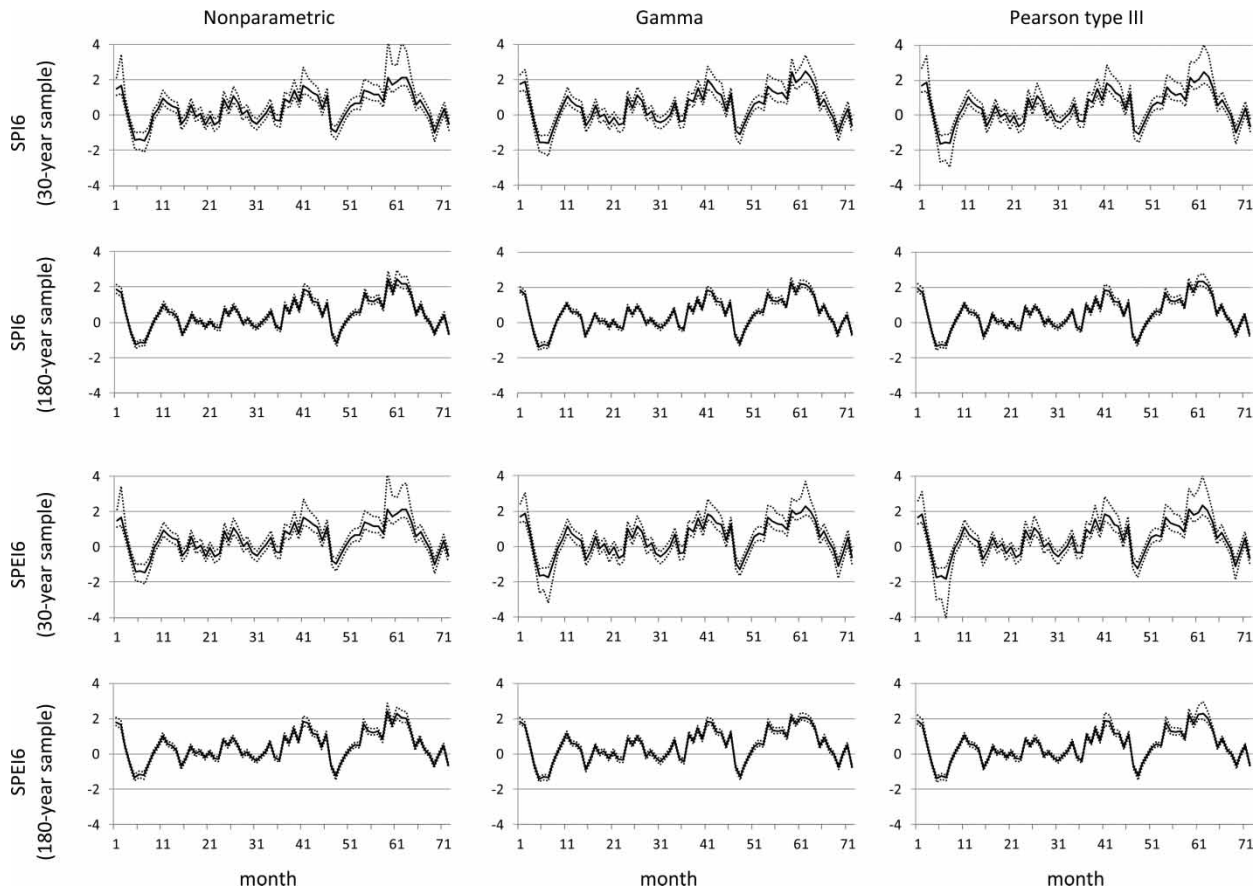

Figure 2 | Exemplificative 6-year time series of SPI6 and SPEI6 (solid lines) computed according to nonparametric and parametric approaches (with different underlying distributions) on the basis of sample sizes 30 and 180 years. Dotted lines represent the bootstrap-based 90% CI. Weather data (1951–2010) of the Terni station.

Uncertainty and reliability analysis based on the proposed measures ΔT and T_{ratio}

The proposed uncertainty measures ΔT and T_{ratio} were calculated for each single value of the 144 series considered in the case study.

Both ΔT and T_{ratio} are calculated on the basis of the CI of standardized indices, of which an illustrative example is provided in Figure 2. As a consequence, also ΔT and T_{ratio} exhibit the already-discussed tendency to increase as the absolute value of the index S increases.

An example of this behavior is given in Figure 3, which shows the scatter plots (S, ΔS), (S, ΔT), and (S, T_{ratio}) for SPI (based on the nonparametric approach and on the fitted GAM and PE3 distributions) and SPEI (based on the nonparametric approach and on the fitted GLO and GEV distributions). The (S, ΔS) plots are only presented as a term of comparison with the two proposed measures ΔT and T_{ratio}. The exemplificative case illustrated in Figure 3 is related to the 6-month time scale and to a 90-year time series; however, similar behaviors could be observed for other characteristics of the raw data. It should be noticed that the y-axes of the plots (S, ΔT) and (S, T_{ratio}) have been limited (for graphical requirements), respectively, to 200 and 100.

Figure 3 also indicates the advantage of both ΔT and T_{ratio} in comparison to ΔS. This last measure, in the range $-1.5 < S < 1.5$, exhibits a more or less constant and irreducible limit of about 0.4 (for all the combinations considered in the figure). The specific value (0.4) is mainly dependent on the sample size (90 years) to which Figure 3 is related. However, it can be observed that in the same range $-1.5 < S < 1.5$, the corresponding ΔT and T_{ratio} are close to 0 and 1, respectively, thus indicating a good reliability of the estimates of the indices. Therefore, a $\Delta S \approx 0.4$ (based on a 90-year time series) can be considered a low-uncertainty condition, but this interpretation has required the quantification of the corresponding ΔT or T_{ratio}. Moreover, apparently negligible increases of ΔS (for $|S| > 2$) can conceal dramatic increases of uncertainty when they are analyzed by the corresponding ΔT or T_{ratio}. This could lead to misleading interpretations of the uncertainty assessment based on ΔS.

These favorable practical aspects of ΔT and T_{ratio} in comparison with ΔS hold in general, irrespective of the location, calculation approach, sample size, distribution, and index type.

In the following sections a detailed analysis of the behavior of the measures ΔT and T_{ratio}, varying the calculation approach, the sample size, the time scale, the types of index, and anomaly, is carried out in relation to the case study considered.

Uncertainty and reliability based on ΔT

The overall results of the uncertainty analysis based on ΔT are presented in Figure 4. This figure shows the median ΔT for SPI and SPEI, varying the time scale, the sample size, the calculation method, and the type of anomaly (drought or wet). The Figure 4 values were obtained by calculating ΔT for each point estimate of the 144 time series of the case study, and then by calculating separately the two medians corresponding to the samples of drought and wet anomalies ($S < 0$ and $S > 0$, respectively).

Figure 4 clearly shows that the length of the time series is one of the most important factors in the regulation of the uncertainty (here measured in terms of ΔT). As expected, the uncertainty decreases as the length of the time series increases. However, this influence is evident up to a length of 90 years. For larger samples, this effect is less evident, and the variability of ΔT due to other factors becomes negligible. The most interesting differences arise from the analysis of ΔT for the smaller sample sizes (30 and 60 years).

First of all, it can be observed that the nonparametric approach usually leads to ΔT smaller and less variable (with the time scale and the index) than those obtained from parametric calculations. Moreover, the ΔT values obtained from the nonparametric approach are systematically lower for the drought than for the wet anomalies. This asymmetry in the CI derives from the fact that the variables considered are usually positively (right) skewed, thus a greater uncertainty is expected on the skew side (i.e., wet anomalies). This aspect is correctly and consistently captured only by the nonparametric approach, that, as also shown in Table 1, provides the best fitting to the data. The parametric distributions (of both SPI and SPEI) provide a correct description of this asymmetric uncertainty only for the 6- and 12-month time scales. Furthermore, the uncertainty of the parametric estimates related to the 30-year sample is much more variable than that related to the

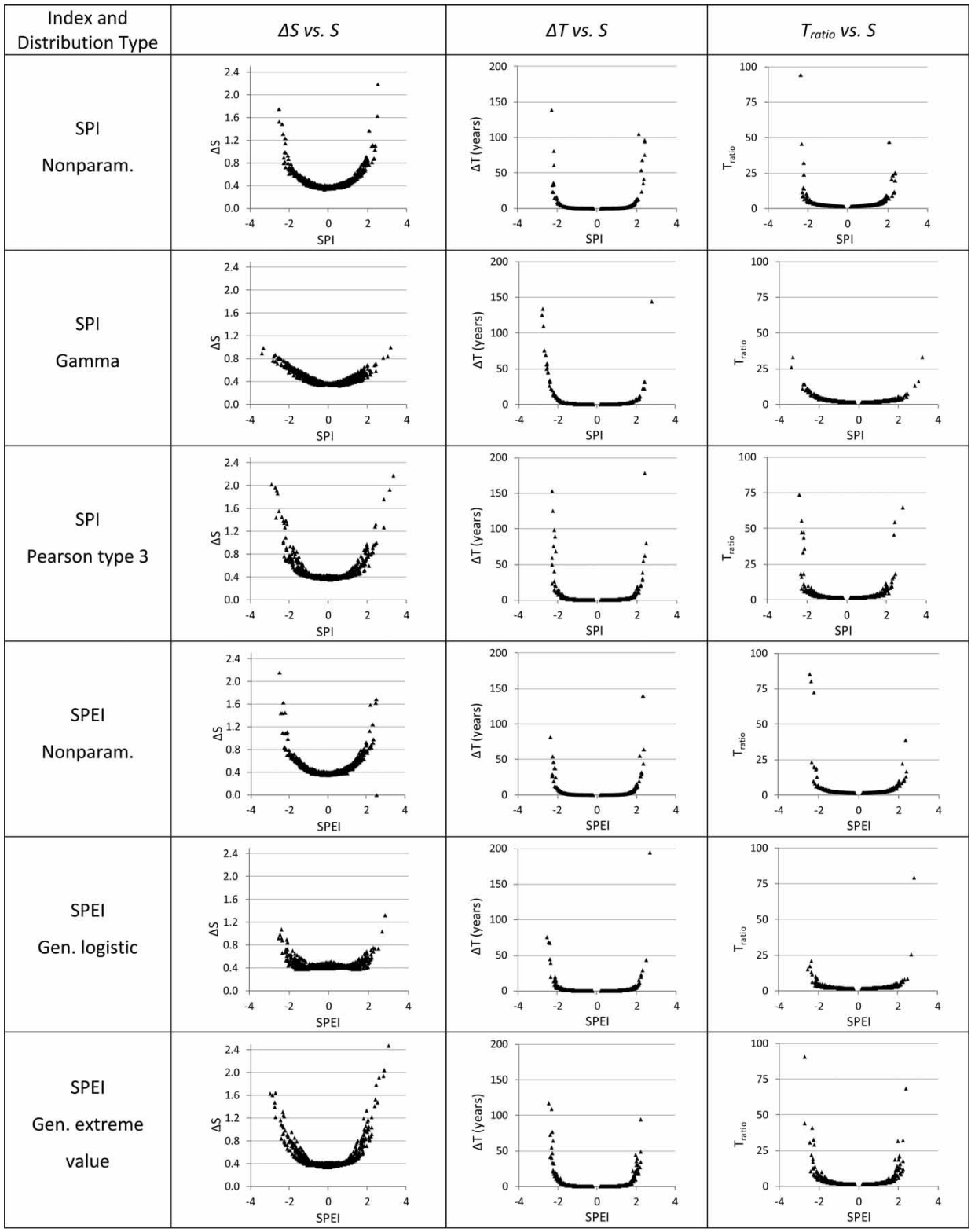

Figure 3 | Plots of the measures of uncertainty ΔS, ΔT, and T_{ratio} versus the values of standardized indices SPI and SPEI estimated by nonparametric and parametric approaches (each index with two different underlying distributions). Example is related to a 6-month time scale and to a 90-year sample size (Terni station, 1951–2010).

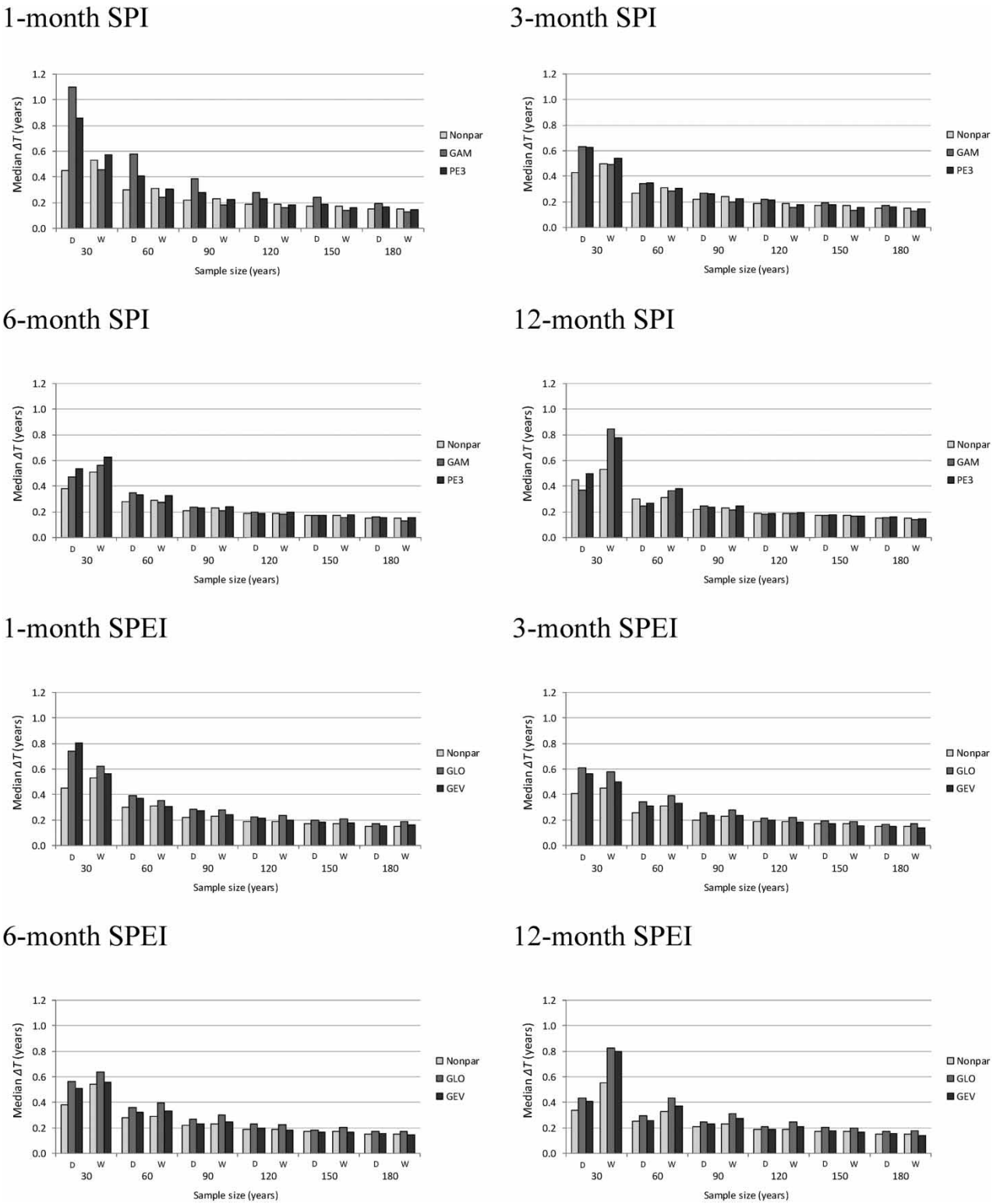

Figure 4 | Median ΔT for SPI and SPEI indices, varying the time scale, the calculation method, the sample size, and the type of anomaly (drought or wet). Nonpar, nonparametric method; PE3, Pearson type III; GAM, gamma; GLO, generalized logistic; GEV, generalized extreme value; D, drought anomalies; W, wet anomalies. Results are related to the Terni station (1951–2010).

nonparametric estimates (this is particularly evident for both the 1-month and 12-month time scales).

As regards the comparison between parametric distributions, for SPI, the ΔT obtained from the PE3 and GAM are usually aligned, apart fron the 1-month time scale, for which the GAM (likely due to the lower goodness-of-fit) presents much higher uncertainty than PE3 for drought conditions. For SPEI, the GEV has, in most cases, slightly lower ΔT values than GLO for both drought and wet anomalies.

The comparative analysis illustrated in Figure 4 was also carried out on the basis of ΔS, instead of ΔT (not shown), with practically identical conclusions. Indeed, as previously explained, the flaw in ΔS does not lie in its capacity to quantify uncertainty, but rather in its practical interpretability.

Uncertainty and reliability based on T_{ratio}

The overall results of the uncertainty analysis based on T_{ratio} are presented in Figure 5. This figure shows the percentages of unreliable point estimates, *UE* (i.e., $T_{ratio} \geq 3$) for SPI and SPEI varying the time scale, the calculation method, the sample size, and the type of anomaly (drought or wet).

As observed for ΔT, the percentage of *UE* also decreases as the length of the time series increases. This decrease is particularly marked when passing from the 30-year to the 60-year lengths, while for longer time series the influence of this factor is less evident (Figure 5). This result clearly demonstrates that the analyses based on only 30 years of observations (which is usually considered the minimum requirement for the calculation of standardized indices) are particularly critical from a reliability viewpoint. On the other hand, in this region of data availability, small increments in the number of observed data can lead to significant improvement of the estimate reliability.

The detailed analysis of the *UE* values obtained for SPI leads to comments similar to those presented in relation to ΔT. For SPEI, the results obtained for the nonparametric approach are aligned with those obtained by considering ΔT. For the parametric distributions, instead, the results are quite different from those obtained by analyzing ΔT: first of all it can be observed that also the parametric distributions often present the expected asymmetry (i.e., less reliability for the wet anomalies). Moreover, in most cases, the GLO has lower percentages of *UE* than GEV (with the sole exception of the wet anomalies for the 30-year time series). In this regard, it can be observed that from an applicative viewpoint, it is preferable to have lower *UE* percentages than lower mean ΔT values. Therefore, the analysis reveals that, for the case study examined, GLO is a better choice than GEV for SPEI calculation. This result is consistent with the recent study by Vicente-Serrano & Beguería (2016), which is in response to the analysis of Stagge *et al.* (2015).

Therefore, it can be observed that the uncertainty measures proposed in this paper (particularly the percentage of *UE*) can be a useful tool also in the selection of the most suitable distribution, which, with other characteristics being equal, should guarantee the lowest percentage of unreliable estimates.

CONCLUSION

In this paper, some bootstrap-based measures of uncertainty and reliability applicable to the standardized indices have been defined. These measures are based on the quantification of the size of the CI of each estimate in terms of years of return period. This allows one to obtain a clear and effective evaluation of the reliability of the index estimate. The application of the measures proposed has been illustrated in relation to the SPI and the SPEI, pointing out the contribution of different factors to the overall uncertainty. In this context, the uncertainty analysis can also be considered as a method, complementary to the goodness-of-fit, for selecting the calculation method most suitable for quantifying a standardized index. Indeed, in this choice, methods able to reduce the percentage of unreliable estimates (other factors being equal) should be preferred. Although the case study is based on a single weather station, some results can be considered valid in general: in particular, nonparametric methods, in the typical regions of data availability, allow one to obtain lower (and more accurate) uncertainty levels than parametric methods.

In the paper, only univariate standardized indices have been considered. The next step might be the extension of this type of uncertainty analysis to bivariate indices. Concerning this, it will be important to adopt a correct resampling technique of the multivariate data (in order to maintain the dependence structure among variables).

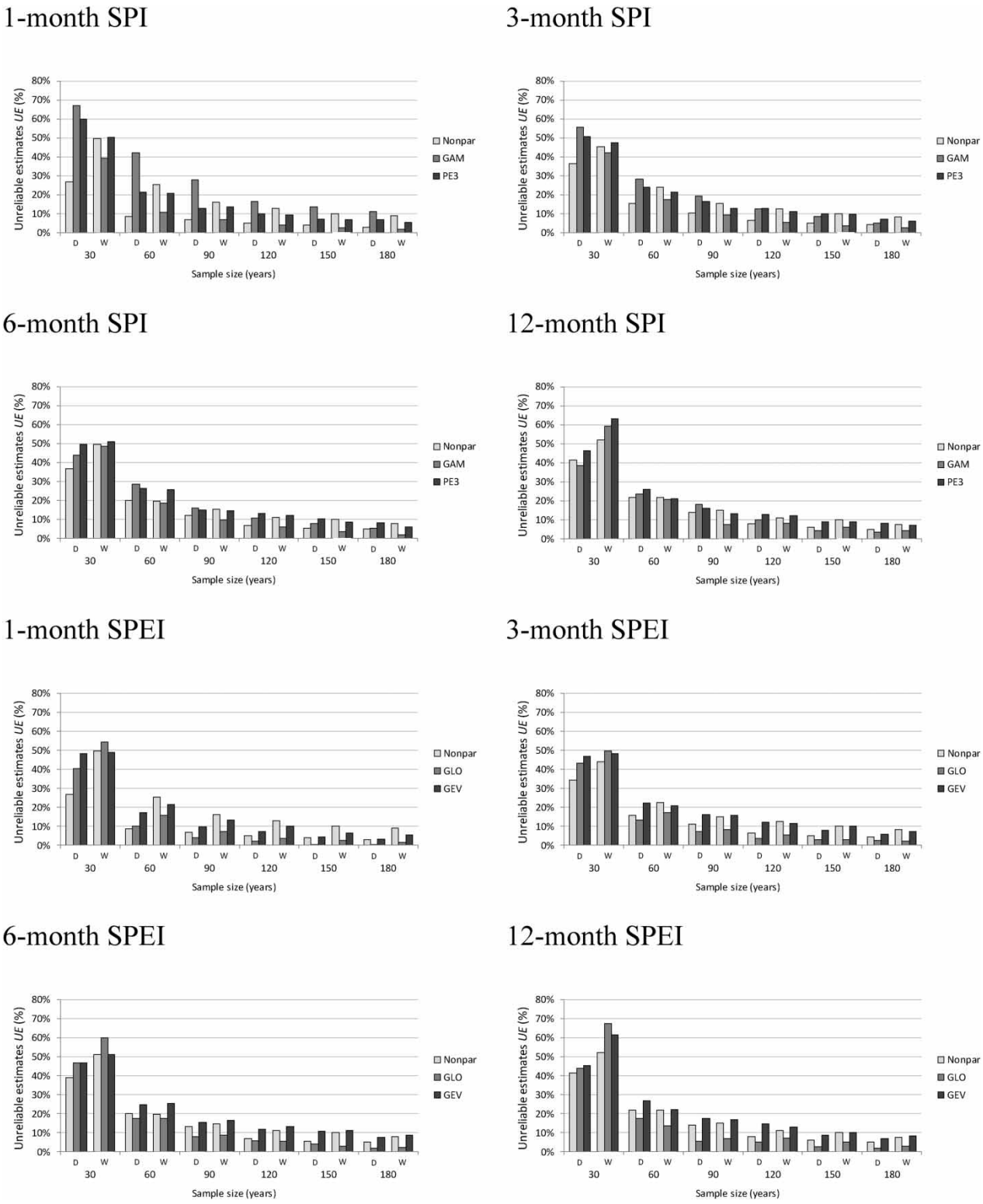

Figure 5 | Average percentage of unreliable estimates *UE* ($T_{ratio} \geq 3$) for SPI and SPEI, varying the time scale, the calculation method, the sample size, and the type of anomaly (drought or wet). Nonpar, nonparametric method; PE3, Pearson type III; GAM, gamma; GLO, generalized logistic; GEV, generalized extreme value; D, drought anomalies; W, wet anomalies. Results are related to the Terni station (1951–2010).

The present computational times of bootstrap technique are reasonable, and therefore this type of uncertainty analysis is advisable as a part of any research or application based on standardized indices.

ACKNOWLEDGEMENTS

This research was financially supported by Fondazione Cassa Risparmio Perugia, project code 2015.0350 021.

REFERENCES

Abramowitz, M. & Stegun, I. A. 1965 *Handbook of Mathematical Function.* Editora, Dover, 1046 pp.

Angelidis, P., Maris, F., Kotsovinos, N. & Hrissanthou, V. 2012 Computation of drought index SPI with alternative distribution functions. *Water Resour. Manage.* **26**, 2453–2473.

Blain, G. C. 2011 Standardized precipitation index based on pearson type III distribution. *Revista Brasileira de Meteorologia* **26**, 167–180.

Droogers, P. & Allen, R. G. 2002 Estimating reference evapotranspiration under inaccurate data conditions. *Irrig. Drain. Syst.* **16**, 33–45.

Efron, B. & Tibshirani, R. J. 1993 *An Introduction to the Bootstrap.* Springer Science + Business Media, Dordrecht.

Farahmand, A. & AghaKouchak, A. 2015 A generalized framework for deriving nonparametric standardized drought indicators. *Adv. Water Resour.* **76**, 140–145.

Guttman, N. B. 1999 Accepting the standardized precipitation index: a calculation algorithm. *J. Am. Water Resour. Ass.* **35**, 311–322.

Hao, Z. & AghaKouchak, A. 2013 A multivariate standardized drought index: a parametric multi-index model. *Adv. Water Resour.* **57**, 12–18.

Hong, X., Guo, S., Zhou, Y. & Xiong, L. 2015 Uncertainties in assessing hydrological drought using streamflow drought index for the upper Yangtze River basin. *Stoch. Environ. Res. Risk. Assess.* **29** (4), 1235–1247.

Hosking, J. R. M. 1990 L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J. R. Stat. Soc. B Stat. Methodol.* **52**, 105–124.

Hu, Y. M., Liang, Z. M., Liu, Y. W., Wang, J., Yao, L. & Ning, Y. 2015 Uncertainty analysis of SPI calculation and drought assessment based on the application of Bootstrap. *Int. J. Climatol.* **35**, 1847–1857.

Kumar, M. N., Murthy, C. S., Sesha Sai, M. V. R. & Roy, P. S. 2009 On the use of Standardized Precipitation Index (SPI) for drought intensity assessment. *Meteorol. Appl.* **16**, 381–389.

Kumar, R., Musuuza, J. L., Van Loon, A. F., Teuling, A. J., Barthel, R., Ten Broek, J., Mai, J., Samaniego, L. & Attinger, S. 2016 Multiscale evaluation of the Standardized Precipitation Index as a groundwater drought indicator. *Hydrol. Earth Syst. Sci.* **20**, 1117–1131.

McKee, T. B. N., Doesken, J. & Kleist, J. 1993 The relationship of drought frequency and duration to time scales. In: *Eight Conference On Applied Climatology*. Amer. Meteor. Soc, Anaheim, CA, pp. 179–184.

Mishra, A. K. & Singh, V. P. 2010 A review of drought concepts. *J. Hydrol.* **391**, 202–216.

Nabaltis, I. & Tsakiris, G. 2009 Assessment of hydrological drought revisited. *Water Resour. Manage.* **23**, 881–897.

Naumann, G., Barbosa, P., Carrao, H., Singleton, A. & Vogt, J. 2012 Monitoring drought conditions and their uncertainties in Africa using TRMM data. *J. Appl. Meteorol. Clim.* **51**, 1867–1874.

Semenov, M. A. & Barrow, E. M. 2002 *LARS-WG – A Stochastic Weather Generator for Use in Climate Impact Studies, User Manual 3.0.* Rothamsted Research, Hertfordshire, 27 pp.

Shukla, S. & Wood, A. 2008 Use of a standardized runoff index for characterizing hydrologic drought. *Geophys. Res. Lett.* **35**, L02405.

Stagge, J. A., Tallaksen, L. M., Gudmundsson, L., Van Loon, A. F. & Stahle, K. 2015 Candidate distributions for climatological drought indices (SPI and spei). *Int. J. Climatol.* **35**, 4027–4040.

Tallaksen, L. M. & van Lanen, H. A. J. 2004 *Hydrological Drought, Processes and Estimation Methods for Streamflow and Groundwater (Development in Water Science).* Elsevier, The Netherlands. p. 579.

Vergni, L., Di Lena, B., Todisco, F. & Mannocchi, F. 2015 Uncertainty in drought monitoring by the Standardized Precipitation Index: the case study of the Abruzzo region (central Italy). *Theor. Appl. Climatol.* DOI: 10.1007/s00704-015-1685-6.

Vicente-Serrano, S. M. & Beguería, S. 2016 Comment on 'Candidate distributions for climatological drought indices (SPI and SPEI)' by James H. Stagge *et al. Int. J. Climatol.* **36**, 2120–2131.

Vicente-Serrano, S. M., Beguería, S. & López-Moreno, J. I. 2010 A multiscalar drought index sensitive to global warming: the Standardized Precipitation Evapotranspiration Index. *J. Clim.* **2**, 1696–1718.

Wand, M. P. & Jones, M. C. 1995 *Kernel Smoothing.* Chapman and Hall, London.

Wu, H., Svoboda, M. D., Hayes, M. J., Wilhite, D. A. & Wen, F. 2007 Appropriate application of the standardised precipitation index in arid locations and dry seasons. *Int. J. Climatol.* **27**, 65–79.

First received 2 March 2016; accepted in revised form 19 July 2016. Available online 30 August 2016