**ORIGINAL PAPER**

# Evaluation of the similarity between drought indices by correlation analysis and Cohen's Kappa test in a Mediterranean area

**L. Vergni[1]** · **F. Todisco[1]** · **B. Di Lena[2]**

## Abstract

In the literature, numerous papers report comparative analyses of drought indices. In these types of studies, the similarity between drought indices is usually evaluated using the Pearson correlation coefficient, $r$, calculated between corresponding severity time series. However, it is well known that the correlation does not describe the strength of agreement between two variables. Two drought indices can exhibit a high degree of correlation but can, at the same time, disagree substantially, for example, if one index is consistently higher than the other. From an operational point of view, two indices can be considered in agreement when they indicate the same severity category for a given period (e.g. moderate drought). In this work, we compared six meteorological drought indices based on both correlation analysis and Cohen's Kappa test. This test is typically used in medical or social sciences to obtain a quantitative assessment of the degree of agreement between different methods or analysts. The indices considered are five timescale-dependent indices, i.e. the Percent of Normal Index, the Deciles Index, the Percentile Index, the Rainfall Anomaly Index, and the Standardised Precipitation Index, computed at the 1-, 3-, and 6-month time-scales, and the Effective Drought Index, a relatively new index, which has a self-defined timescale. The indices were calculated for 15 stations in the Abruzzo region (central Italy) during 1951–2018. We found that the strength of agreement depends on both the criteria of drought severity classification and the different indices' calculation method. The Cohen's Kappa test indicates a prevailing moderate or fair agreement among the indices considered, despite the generally very high correlation between the corresponding severity times series. The results demonstrate that the Cohen's Kappa test is more effective than the correlation analysis in discriminating the actual strength of agreement/disagreement between drought indices.

✉ L. Vergni
  lorenzo.vergni@unipg.it

1    Department of Agricultural, Food and Environmental Sciences, University of Perugia, Borgo XX Giugno 74, 06121 Perugia, Italy

2    Abruzzo Region – Agriculture Directorate- Regional Agro-Meteorological Centre, Scerni, CH, Italy

Published online: 07 May 2021

 Springer

# 1 Introduction

Although it is difficult to give a univocal definition of drought, this can be very generically defined as a natural hazard that arises from a considerable deficiency in precipitation for prolonged periods (Folger 2017). Droughts are usually classified considering the affected domain, for example, distinguishing meteorological, hydrological and agricultural droughts (Whilite et al. 2007). Drought characterisation in a given area is very useful, both for defining early-warning systems (Kogan 2000), and for water resources planning and management (Zargar et al. 2011). For these purposes, drought indices are typically used: these are measures deriving from the elaboration of suitable variables or data such as precipitation, evapotranspiration, or satellite images, which allow us to express the drought characteristics (e.g. severity) in a more effective way than the original information (Zargar et al. 2011). Moreover, the operational use of drought indices is based on specific classification criteria, enabling a qualitative description of the drought severity (e.g. extreme, severe, moderate, etc.) according to the index's value. This way of proceeding has the advantage of allowing a rapid understanding of the type of ongoing event and of making its description independent of the scale of values assumed by a particular index.

The drought indices proposed and used in monitoring the phenomenon in its various implications are very numerous (Keyantash and Dracup 2002; WMO 2006), precisely because they reflect the complexity of the phenomenon and the application contexts' variability. For example, Zargar et al. (2011) mention as many as 74 drought indices in their review. Some indices are very common and universally used and accepted (e.g. the Standardised Precipitation Index, SPI), whereas others are used in a more local context.

This huge availability of drought indices has often led to comparative studies aimed at identifying the index (or group of indices) and the related timescales that are more suitable to describe a specific type of drought in a given area. Usually, the evaluation of the performances is carried out based on the ability of the different indices to correctly describe the characteristics (onset, duration, severity, etc.) of historical drought events or based on their correlation with hydrological or agricultural impacts in a specific region (Myronidis et al. 2018; Blauhut et al., 2015). Examples can be found, amongst others, in Morid et al. (2006) for the Tehran province (Iran), Jain et al. (2015) for India, Salehnia et al. (2017) for Iran, Bayissa et al. (2018) for Ethiopia, Myronidis et al. (2018) for Oregon (USA), and Wable et al. (2019) for a semiarid region of India. In this type of study, a preliminary part of the analysis usually deals with evaluating the similarity among the indices, almost always based on the Pearson correlation coefficient $r$ between the corresponding severity time series.

The correlation coefficient $r$ is a measure of the strength of the linear relationship between two generic variables, and it helps to evaluate the similarity of the temporal dynamics of two indices. However, as Schober et al. (2018) underlined, the correlation does not describe the strength of agreement between two variables. Let us consider, for example, two generic drought indices X1 and X2, based on the same severity classification scheme, such as that of SPI (Table 1). Let us also assume that, in a given period, the two indices follow the dynamics shown in Fig. 1a. The correlation between X1 and X2 in Fig. 1a is perfect ($r=1$), despite the corresponding severities that differ systematically. In this specific case, X1 and X2 indicate different drought categories at each time step considered (i.e. the two indices are not in agreement).

High correlations between drought indices are reported in most of the literature mentioned above, but, according to the previous considerations, this does not necessarily

**Table 1** Severity classification adopted for the indices considered

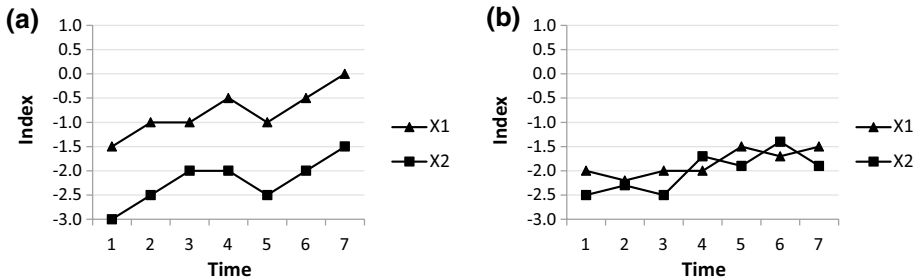| Class, C | PN (%) | DI | PI (%) | SPI | RAI | EDI |
|---|---|---|---|---|---|---|
| Wet | PN > 120 | DI > 7 | PI > 84 | SPI > 1 | RAI > 1 | EDI > 1 |
| Normal | 80 < PN ≤ 120 | 3 < DI ≤ 7 | 16 < DI ≤ 84 | −1 < SPI ≤ 1 | −1 < RAI ≤ 1 | −1 < EDI ≤ 1 |
| Moderately dry | 40 < PN ≤ 80 | 2 < DI ≤ 3 | 7 < PI ≤ 16 | −1.5 < SPI ≤ −1 | −1 < RAI ≤ −2 | −1.5 < EDI ≤ −1 |
| Severely dry | 20 < PN ≤ 40 | 1 < DI ≤ 2 | 3 < PI ≤ 7 | −2 < SPI ≤ −1.5 | −2 < RAI ≤ −3 | −2 < EDI ≤ −1.5 |
| Extreme dry | PN < 20 | DI < 1 | PI < 3 | SPI < −2 | RAI < −3 | EDI < −2 |



**Fig.1** Theoretical ex  amples illustrating the concepts of correlation and agreement between two generic drought indices X1 and X2: **a** X1 and X2 are highly correlated but not in agreement; **b** X1 and X2 are poorly correlated, but substantially in agreement

indicate similarity between the indices. For example, Salehnia et al. (2017), when comparing eight precipitation-based drought indices (SPI, PN (Percent of Normal Index), DI (Deciles Index), EDI (Effective drought index), CZI (China China-Z index), MCZI (Modified CZI), RAI (Rainfall Anomaly Index), and ZSI (Z -score Index)), always found $r \geq 0.94$, with very few differences among indices (except for the comparisons involving EDI, for which the correlations are slightly lower). High correlations ($r \geq 0.74$) were also indicated by Morid et al. 2006 regarding the comparisons among CZI, MCZI, ZSI and SPI. In Wable et al. (2019), the correlation matrix related to SPI, RDI (Reconnaissance Drought Index), PN and SPEI (Standardised Evapotranspiration Precipitation Index) always shows $r \geq 0.79$ for the 6-month and longer timescales. In Jain et al. (2015), average $r$ values greater than 0.88 are indicated for the comparisons among DI, SPI, CZI and ZSI.

A high correlation may depend on the fact that many indices have almost linear relationships in a wide range, although they differ greatly in the extreme values (tails). In this situation, the correlation coefficient is mainly affected by the central (i.e. normal) values and does not adequately highlight the marked differences in the tails (which instead represent the most interesting conditions from an operational viewpoint), thus concealing possible differences in the assessment of severe and extreme drought conditions.

On the other hand, moderate differences between the values of two generic indices and their not synchronous temporal dynamics could lead to low correlation values, even in those cases in which the two indices are in substantial agreement in terms of drought severity classification. In the example of Fig. 1b, the correlation coefficient between X1 and X2 is only 0.52, despite X1 and X2 being in substantial agreement in terms of drought category at each time step.

Since the indices' operational use is mainly done through a qualitative classification of the severities (e.g. Table 1), the similarity among indices should consider the

correspondence between the drought classifications rather than the correlation between the quantitative values of the indices. In practice, from an operational point of view, two drought indices can be considered in agreement when, for a given period, they indicate the same severity classification (e.g. moderate, severe, extreme drought).

This type of evaluation can be effectively obtained by the Cohen's Kappa test (Cohen 1960, 1968), one of the most well-known methods to compare the agreement between two different raters or analysis techniques (Craig 1981). However, this test has been applied in very few cases to compare the similarity between drought indices. Examples can be found in Teweldebirhan Tsige et al. (2019) and in Ezzine et al. (2014), who compared meteorological and agricultural drought indices using the Cohen's Kappa test, amongst other methods.

In this paper, we describe and discuss the advantages of applying the Cohen's Kappa test to assess the similarity between drought indices, using the traditional correlation analysis as a term of comparison. Moreover, both the unweighted and weighted versions of the Cohen's Kappa test are applied. As detailed in the methods section, this will enable us to assess both the strength of agreement between two drought indices (unweighted version) and how different are the severity classes when two indices disagree (weighted version).

To illustrate the method, we compared some well-known meteorological drought indices at 15 stations in the Abruzzo region (central Italy) from 1951 to 2018. We limited the analysis to six meteorological indices, which only require precipitation data: the Percent of Normal Index (PN), the Decile Index (DI), the Percentile Index (PI), the Standardised Precipitation Index (SPI), the Rainfall Anomaly Index (RAI), and the Effective Drought Index (EDI). The indices considered include several different characteristics: five indices, i.e. PN, DI, PI, RAI, and SPI, need the specification of a timescale, while EDI has a self-defined timescale. Moreover, the EDI, unlike the other indices, is based on the concept of effective precipitation, i.e. the precipitation of the previous periods influences the current EDI value with weights that decrease as the temporal distance of those precipitations increases from the current period (Byun and Wilhite 1999). Some indices (DI, PI, SPI) have a precise probabilistic interpretation while others (PN, RAI, EDI) do not. Most indices (DI, SPI, PI, RAI, EDI) are suitable for spatial and temporal comparisons, while PN is not. Finally, the Percentile Index, defined in this paper, is conceptually and quantitatively analogous to the Decile Index, but it is based on a different classification of drought severity.

For simplicity's sake, the calculation of the indices PN, DI, PI, RAI and SPI was carried out only for short-medium timescales (i.e. 1-, 3- and 6-months), which are typically associated with agricultural drought conditions.

## 2 Materials and methods

### 2.1 Study area and data

The region of Abruzzo is located in Central Italy and covers an area of 10,798 km$^2$. The region is predominantly mountainous and hilly and has the highest peaks in the Apennine Mountains, while plains are mainly located in the coastal strip bordering on the Adriatic Sea and in some inland plateaus (Fig. 2). The climate is Mediterranean, with hot, dry summers and mild winter, with rainfall usually being higher in autumn.
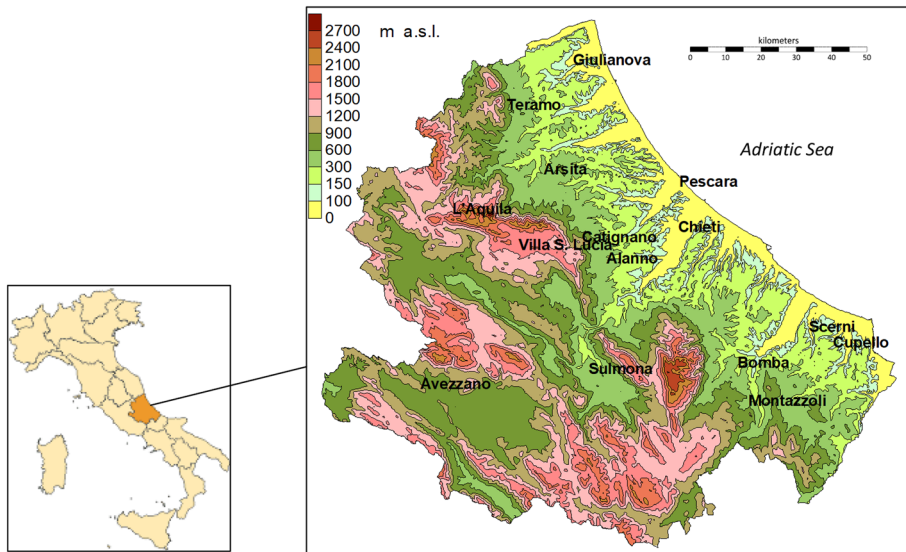
**Fig. 2** Stations considered and orography of the Abruzzo region

The study was carried out based on the rainfall data recorded by the Hydrographic Regional Service from 1951 to 2018[1] at 15 stations, quite uniformly distributed in the region (Fig. 2). These data are part of a more extensive database already used in previous works (e.g. Di Lena et al. 2014; Vergni et al. 2016) and are here updated to the year 2018. In the same references, the details related to the data quality assessment, with particular reference to the homogeneity and the reconstruction of missing data, can also be found.

The mean monthly precipitation at the 15 stations considered is presented in Fig. 3. The mean annual precipitation ranges from 644 mm (Sulmona) to 992 mm (Arsita).

## 2.2 Drought indices

The indices were computed for each station based on monthly precipitation data. For the indices PN, DI, PI, SPI and RAI, requiring the specification of an accumulation period, we only considered the 1-, 3- and 6-month timescales.

The drought severity classification for the different indices was based on a common scheme (Table 1) which creates the distinction between wet, normal, and dry conditions (in turn divided into 3 levels). Wet conditions have not been further subdivided, given that the work is mainly focussed on the degree of agreement of the indices in the classification of drought events. However, the method proposed can be applied to any classification criterion.

Details about the index's calculation, their main characteristics, and the classification criteria are provided in the next paragraphs.

---

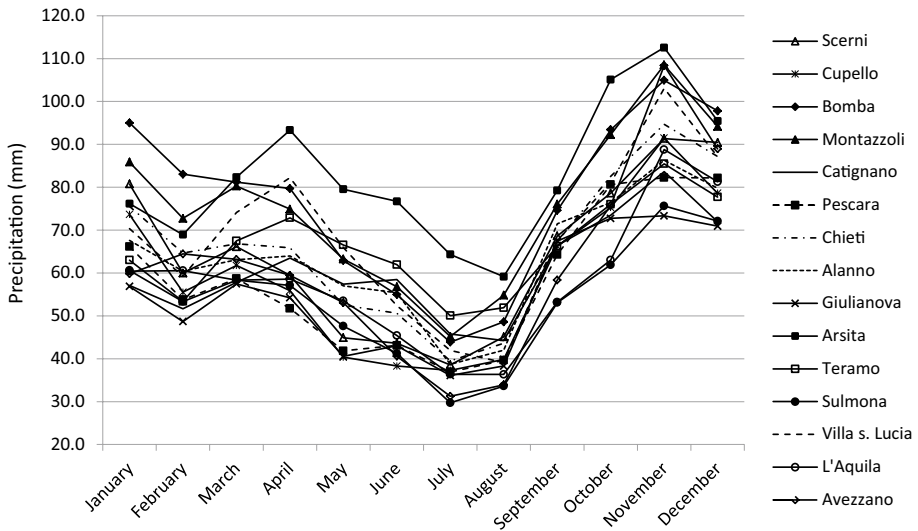[1] For the station of Giulianova the series is 1951–2017.

**Fig. 3** Mean monthly precipitation at the station considered during 1951–2018

### 2.2.1 Percent normal PN

The percent of normal, PN, is calculated by dividing a given observation of cumulative precipitation $X_{i,j}^{k}$, related to the month $j$ ($j=1,\ldots,12$) and timescale $k$, by the normal precipitation of the same period (typically the mean of a data series of at least 30 years). The simplicity is the main strong point of the PN index, and it can be considered quite effective for comparing the anomalies related to a single region or season. PN meets the needs of a general audience since it does not require in-depth statistical knowledge. Disadvantages are mainly associated with the lack of robustness (Zagar et al. 2011): the percent of normal implies a normal distribution where the mean and median are considered the same, but in most cases, precipitation on monthly or seasonal scales does not have a Normal distribution. Moreover, the precipitation probability distribution changes with localities and seasons, making any spatial or temporal comparison based on PN unreliable. The classifications considered for PN (Table 1) were derived from Fluixá-Sanmartín et al. (2018).

### 2.2.2 Deciles Index DI and Percentile Index PI

The Deciles Index, DI, groups the observations of cumulative precipitation $X_{j}^{k}$ in deciles. At this end, the values $X_{i,j}^{k}$ (from a long-term record) are first ranked from lowest to highest, assigning the corresponding cumulative frequency (%) to each observation. The first decile is the precipitation value not exceeded by the lowest 10% of all precipitation values, and the second is between the lowest 10 and 20%, etc. The DI index, developed by Gibbs and Maher (1967), is still a reference point in drought monitoring in some countries, such as Australia. DI is an elementary index, easy to calculate, whose understanding is quite intuitive. Despite this simplicity, it is more robust than PN, as it has a precise statistical meaning and characteristics of spatial and temporal comparability. DI is usually grouped into five classes: if precipitation falls into the lowest 20% (deciles 1 and 2), it is classified

as "much below normal". Deciles 3 and 4 indicate "below normal" precipitation, deciles 5 and 6 indicate "near normal", deciles 7 and 8 "above normal", and deciles 9 and 10 "much above normal". In this work, to increase the comparability among the indices considered, the classification of DI was based on 3 (instead of 2) drought classes (Table 1) as proposed in Morid et al. (2006).

In this paper, the percentile index PI is also proposed. It is computed the same as DI, but PI uses percentiles instead of deciles, allowing more flexibility in the definition of the severity classes. In particular, the class limits used for PI (Table 1) almost exactly correspond to the cumulative probability values (%) of the class limits of SPI (which ultimately corresponds to the standard variable z).

### 2.2.3 Standardised Precipitation Index SPI

The Standardised Precipitation Index (McKee et al. 1993) is one of the most commonly adopted drought indicators and its use in meteorological drought assessment has been recommended by the World Meteorological Organisation (Hayes et al. 2011). A brief description of the calculation steps is provided hereafter, and more details can be found in several papers (e.g. Lloyd-Hughes and Saunders 2002). The first step is the determination of a Probability Density Function (PDF) suitable for the description of the long-term series of observations $X_{i,j}^k$. The fitting is performed for each calendar month to consider the climatic differences due to seasonality. Once this PDF is determined, the cumulative probability of an observed precipitation amount $X_{i,j}^k$ is estimated. Thanks to an equiprobability transformation (Abramowitz and Stegun 1965), the estimated cumulative probability is transformed into a standardised variate representing the SPI value. The Gamma distribution is the commonly adopted PDF (McKee et al. 1993), even if other PDFs could have a better goodness of fit in some contexts (Stagge et al. 2015). Here, the SPI was calculated according to the initially suggested algorithm (specifically, assuming that the cumulated precipitation follows a Gamma distribution).

SPI, as underlined by Guttman (1998) and Lloyd-Hughes and Saunders (2002), has several strong points: (1) suitability to spatial and temporal comparisons; (2) flexibility, as it can be applied at different timescales; (3) reduced calculation complexity in comparison to other indices; (4) ease of application, as it is based only on precipitation. However, some weak points have also been indicated. For example, Wu et al. (2007) recommended cautions in interpreting short-timescale SPI in arid climatic regimes (i.e. series with zero rainfall values) due to the non-normality of SPI. Critical aspects are also associated with the choice of a suitable PDF, which introduces elements of uncertainty in the SPI estimates (Vergni et al. 2017). Moreover, SPI and the other timescale-dependent indices here considered (i.e. PN, DI, PI, and RAI) depend only on the cumulative value of the precipitation in the period of length $k$, regardless of both the distribution of the precipitation in the same period and its temporal distance from the current period. In most contexts (e.g. agricultural, hydrological), this approach can lead to misleading assessments of the current drought conditions and related impacts.

### 2.2.4 Rainfall Anomaly Index RAI

The rainfall anomaly index RAI (van Rooy 1965) is computed by comparing each observation $X_{i,j}^k$ with the mean of the ten highest (for positive anomalies) and the ten lowest (for negative anomalies) precipitation records (for the same month $j$ and timescale $k$). The

calculation starts by arranging the precipitation data $X_{i,j}^k$ in descending order. The ten highest values are averaged to form a threshold $\overline{M}$ for positive anomalies, and the ten lowest values are averaged to form a threshold $\overline{m}$ for negative anomalies. The RAI index corresponding to the $i_{th}$ precipitation value is computed as:

$$RAI = 3\frac{X_{i,j}^k - \overline{X}_j^k}{\overline{M} - \overline{X}_j^k} \quad \text{for } X_{i,j}^k \geq \overline{X}_j^k \tag{1}$$

$$RAI = -3\frac{X_{i,j}^k - \overline{X}_j^k}{\overline{m} - \overline{X}_j^k} \quad \text{for } X_{i,j}^k < \overline{X}_j^k \tag{2}$$

where $X_{i,j}^k$ is the actual precipitation related to the month $j$ and timescale $k$ in the year $i$; and $\overline{X}_j^k$ is the corresponding long-term mean. The RAI classification considered in this paper (Table 1) is derived from Fluixá-Sanmartín et al. (2018), with the only modification being the limits of the normal class, which here also includes the categories defined as "slightly dry" and "slightly wet" in Fluixá-Sanmartín et al. (2018).

## 2.2.5 Effective Drought Index EDI

In its original formulation (Byun and Wilhite 1999), the Effective Drought Index (EDI) was computed on a daily timescale, but the same principles were used later to calculate EDI based on monthly precipitation data (e.g. Smakhtin and Hughes 2007; Pandey et al. 2008; Jain et al. 2015). This last approach is the one considered here for better comparability with the other indices.

The first step is the calculation of EP (Effective Precipitation), defined as a function of the current month's rainfall and weighted rainfall over a defined preceding period. If $P_m$ is the rainfall of the current month and $D$ is the duration of the preceding period, then the $EP$ for the current month is:

$$EP = \sum_{N=1}^{D}\left(\frac{\sum_{m=1}^{N} P_m}{N}\right) \tag{3}$$

For example, for $D=3$, $EP=P_1+(P_1+P_2)/2+(P_1+P_2+P_3)/3$, where $P_1$, $P_2$ and $P_3$ are precipitation amounts during the current month, the previous month and two months before, respectively.

In most application of EDI, $D=365$ days (i.e. 12 months) is assumed since it represents the most common precipitation cycle (Byun and Wilhite 1999; Deo et al. 2017).

Then, the long-term mean of $EP$ (for a specific month) is calculated and named $MEP$. The EDI index is finally calculated as:

$$EDI = \frac{DEP}{St(DEP)} \tag{4}$$

where $DEP$ is the current anomaly $(DEP=EP-MEP)$, and $St(DEP)$ is its standard deviation in the time series.

Unlike SPI, the EDI is not normally distributed and preserves the original time series' skewness (Smakhtin and Hughes 2007).

Among the indices considered, EDI is the one that differs the most in terms of both calculation and physical interpretation. The other indices depend on the simple summation of precipitation in a given period, while EDI is based on the concept of effective precipitation EP (Eq. 3). This implies that more weight is given to the current period's precipitation (day or month) and progressively descending weights to all previous periods. According to Byun and Whilite (1999), the EDI algorithm implicitly considers the progressive decrease of water resources (e.g. soil moisture, water reservoirs) after rainfall due to the runoff and evapotranspiration processes. Several authors (e.g. Pandey et al. 2008; Deo et al. 2017; Jain et al. 2015) demonstrated the advantages of the EDI approach in describing historical drought events. Compared to other indices that refer to fixed timescales, the main advantage of EDI is that it provides continuous and reliable monitoring of the progression of drought (or wet) periods. Moreover, the EDI application fields extend to other contexts, such as flash flood warnings (Deo et al., 2015). This advantage is most appreciable if EDI is calculated with the original daily sliding timescale, but to a certain extent, it also remains at the monthly scale.

In most applications available in the literature (e.g. Salehnia et al. 2017; Deo et al. 2017; Smakhtin and Hughes 2007), the classification proposed for the EDI index is the same as SPI (Table 1).

## 2.3 Methods

The EDI was compared with all the other indices (PN, DI, PI, SPI, and RAI) at their different timescales. However, these last indices were compared with each other, only considering the same timescale.

The similarity between drought indices was first evaluated by analysing the two-by-two scatter plots of corresponding monthly severity time series (1951–2008) and calculating the Pearson correlation coefficient $r$, as is usually done in the literature.

A more effective and innovative assessment of the similarity between the indices was then obtained comparing all possible pairs of indices in terms of the severity category assigned to each month of the time series, and then quantifying their degree of agreement by the Cohen's Kappa test. This test is often applied to evaluate the degree of agreement between two raters, methods, or classifiers, mainly in medical sciences. The test's main principles are here briefly summarised here; more details can be found in Cohen (1960, 1968).

The classification obtained from two classifiers can be represented by an agreement matrix (Cohen 1968), where along the upper-left to lower-right diagonal, there are the cells of perfect agreement (agreement diagonal). An example of an agreement matrix between two generic drought indices with C = 5 severity categories (see Table 1) for climatic conditions is given in Table 2, where $p_{ij}$ represents the generic proportion of classifications assigned to the category $i$ by the index A, and to the category $j$ by the index B (i, j = 1…C). The notations $pi.$ and $p.j$ indicates the row and column proportion totals, respectively.

The Kappa test is based on the following coefficient of agreement $K$, corrected for chance:

$$K = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

where $p_O$ is the total observed frequency of agreement and $p_e$ is the proportion of agreement expected by chance. In practice, $K$ indicates how much the observational frequency

**Table 2** Example of agreement matrix between two generic drought indices A and B

| | Index A | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| $ij$ | Wet | Normal | Moderate drought | Severe drought | Extreme drought | $p_i$ |
| Index B | | | | | | |
| Wet | $p_{11}$ | $p_{12}$ | $p_{13}$ | $p_{14}$ | $p_{15}$ | $p_1$ |
| Normal | $p_{21}$ | $p_{22}$ | $p_{23}$ | $p_{24}$ | $p_{25}$ | $p_2$ |
| Moderate drought | $p_{31}$ | $p_{32}$ | $p_{33}$ | $p_{34}$ | $p_{35}$ | $p_3$ |
| Severe drought | $p_{41}$ | $p_{42}$ | $p_{43}$ | $p_{44}$ | $p_{45}$ | $p_4$ |
| Extreme drought | $p_{51}$ | $p_{52}$ | $p_{53}$ | $p_{54}$ | $p_{55}$ | $p_5$ |
| $p_{\cdot j}$ | $p_{\cdot 1}$ | $p_{\cdot 2}$ | $p_{\cdot 3}$ | $p_{\cdot 4}$ | $p_{\cdot 5}$ | $1$ |

**Table 3** Ranges of the Kappa statistic $K$ and corresponding strength of agreement

| Kappa statistic | Strength of agreement |
| --- | --- |
| <0.00 | Poor |
| 0.00–0.20 | Slight |
| 0.21–0.40 | Fair |
| 0.41–0.60 | Moderate |
| 0.61–0.80 | Substantial |
| 0.81–1.00 | Almost perfect |

of agreement is in excess of the frequency of agreement $p_e$, expected under a random classification. The value of $p_o$ is obtained by summing the observed proportions in the agreement diagonal (i.e. $\sum_{h=1}^{5} p_{hh}$ in Table 2). The value of $p_e$ is obtained by summing the products of the corresponding row and column proportions in the agreement diagonal (i.e. $\sum_{h=1}^{5} (p_{\cdot h} \cdot p_{h \cdot})$ in Table 2).

The statistical test associated with $K$ tests the null hypothesis that the agreement's extent is the same as random ($K=0$). The relative strength of agreement associated with kappa statistics is sometimes evaluated referring to a qualitative description of the strength of agreement like that given in Table 3 (Landis and Koch 1977).

The $K$ statistic is based only on the proportions in the agreement diagonal, and therefore the result is not dependent on the proportions in the disagreement cells. This can sometimes lead to misleading information because some disagreements in assignments are of greater gravity than others. Cohen (1968) proposed a generalisation of the previous statistics (weighted Kappa, $K_w$), which is able to overcome the previous shortcoming:

$$K_w = 1 - \frac{\sum v_{ij} p_{oij}}{\sum v_{ij} p_{eij}} \qquad (6)$$

where $v_{ij}$ ($v_{ij} > 0$) is the disagreement weight assigned to the cell $ij$, $p_{oij}$ is the proportion of the joint judgment observed in the $ij$ cell, and $p_{eij}$ is the proportion in the same cell expected by chance (calculated as the product of the corresponding row and column proportions).

The weights are assigned according to a subjective but rational criterion. The weights quantitatively influence the value of $K_w$ and its statistical significance. Consequently, even if $K_w$ allows obtaining a more exhaustive evaluation of the agreement degree, it also

introduces an element of subjectivity due to the selected weights. In this paper, the weights were assigned according to the pattern named "squared", in which the weights are a function of the squared distance from the agreement diagonal, as illustrated in Table 4 (Cohen 1968; Gamer et al. 2019).

As for $K$, Cohen (1968) proposed a statistical procedure to test the null hypothesis that the extent of agreement is the same as random ($K_w = 0$). Instead, due to the weight dependence, a qualitative classification similar to that of $K$ (Table 3) is not available for $K_w$.

$K_w$ can be either larger or smaller than $K$. This depends on how the different weights are combined with the disagreement proportions. For example, a $K_w$ greater than $K$ indicates that, in general, the two classifiers disagree much less than chance expectation where it counts very much (high weight) and disagree at about a chance level where it counts little (low weight).

Moreover, it should be noted that the use of a constant generic weight for all the disagreement cells ($i \neq j$) leads to $K_w = K$, and for this reason, Eq. (6) is considered a generalisation of (5).

In the paper, the evaluation of the classification agreement among indices was obtained on the basis of both $K$ and $K_w$, whose values and statistical significance were determined by using the procedures available in the R package "irr" (Gamer et al. 2019).

# 3 Results

In the presentation and discussion of the results, the indices that require a timescale specification will be indicated with the notation $Xk$, where $X$ is the generic acronym of the index and $k$ is the timescale (e.g. SPI1 indicates the SPI computed at the 1-month time scale). When the timescale is not specified, it means that the result is valid for all the timescales.

## 3.1 Scatter plots and correlation

An overall description of the relationships among indices is given in Fig. 4 which shows all the possible two-by-two scatterplots of corresponding severities. For each scatterplot of Fig. 4, the value of the Pearson correlation coefficient $r$ is also shown.

The scatterplots were obtained considering the data of all the stations together. The same analysis (not shown), carried out for each station, led to very similar results, and therefore, all stations' data were merged together for simplicity's sake.

In Fig. 4, the scatter plots are subdivided in three groups related to the timescales adopted to calculate the indices DI, PI, PN, SPI, and RAI. When expressed as the empirical cumulative frequency in percentage (instead of decile or percentile values), the DI and PI indices are perfectly equivalent. Therefore, the scatterplots DI-PI are not shown in Fig. 4, as they would be in a 1: 1 line with a correlation coefficient $r = 1$.

The scatterplots' analysis reveals that all the pairs of indices exhibit a more or less evident positive correlation at all the timescales, with $r$ ranging from 0.63 to 0.99.

The highest scatterings are observed for the comparisons involving the EDI index. This result was expected, considering that EDI has a very different calculation algorithm compared to the other indices. The correlations between the EDI and the other indices computed at the 1-month scale are the lowest (average $r \approx 0.65$), but when the same indices are calculated at the 3- and 6-month scales, the correlations with EDI rise well beyond 0.8, although the points scattering remains relevant.

Table 4 Observed frequencies of agreement $p_{oij}$ between DI and RAI at a 3-month timescale (DI3 and RAI3, respectively)

|  |  | RAI3 classification |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  | Wet | Normal | Moderate Drought | Severe Drought | Extreme drought | $p_i$ |
| **DI3** classification | Wet | **0.2495** (0) | 0.0617 (1) | 0.0001 (4) | 0.0001 (9) | 0.0000 (16) | 0.3113 |
|  | Normal | 0.0026 (1) | **0.3137** (0) | 0.0780 (1) | 0.0005 (4) | 0.0001 (9) | 0.3949 |
|  | Moderate Drought | 0.0002 (4) | 0.0009 (1) | **0.0877** (0) | 0.0135 (1) | 0.0002 (4) | 0.1026 |
|  | Severe Drought | 0.0000 (9) | 0.0002 (4) | 0.0124 (1) | **0.0898** (0) | 0.0003 (1) | 0.1027 |
|  | Extreme Drought | 0.0000 (16) | 0.0002 (9) | 0.0001 (4) | 0.0262 (1) | **0.0620** (0) | 0.0885 |
|  | $p_j$ | 0.2524 | 0.3767 | 0.1784 | 0.1300 | 0.0626 | 1.0000 |

Values in parentheses are the disagreement weights $v_{ij}$ assigned according to a "squared distance" pattern (Cohen 1968). Values in bold indicate the agreement diagonal. The total frequency of agreement is $p_O = 0.8027$, the proportion of agreement expected by chance, $p_e$, is 0.2645. Overall data from 15 stations during 1951–2018
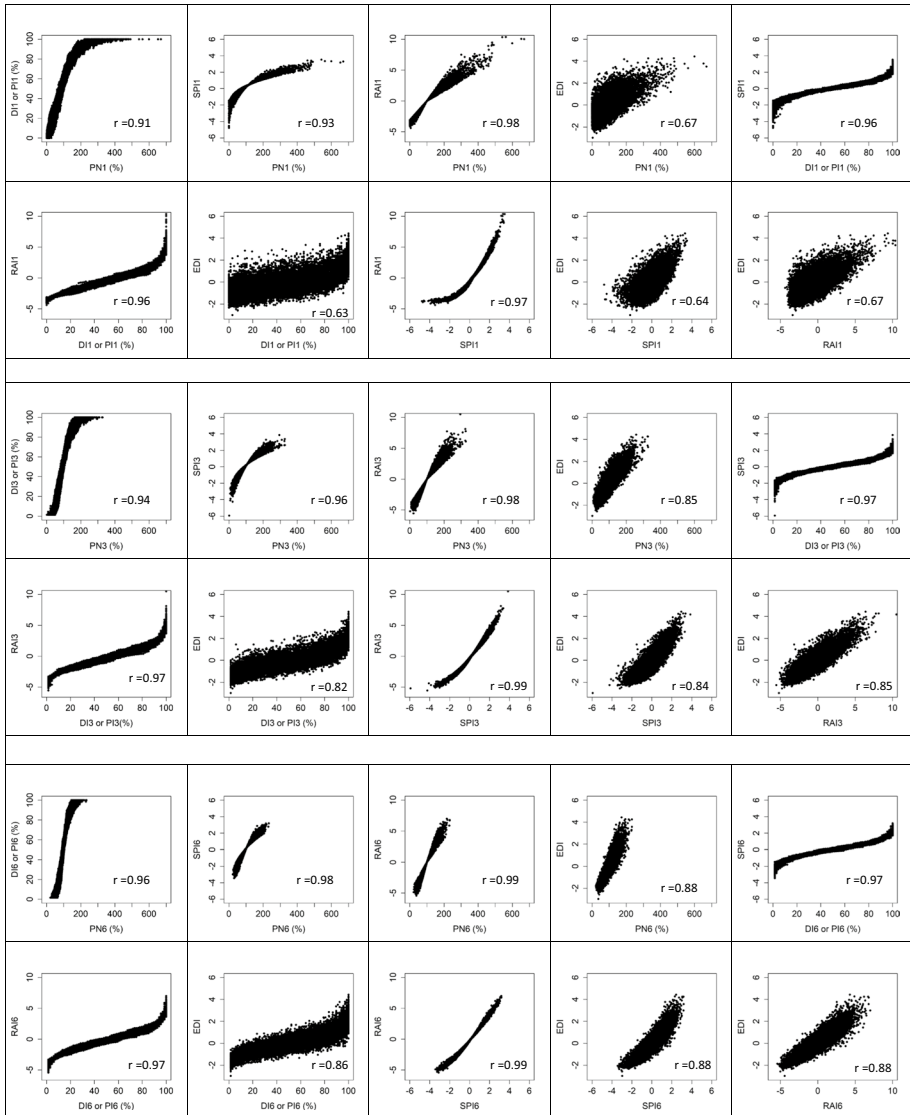
**Fig. 4** Scatter plots and correlation coefficients of the pairs of indices considered in the comparative analysis

Apart from the comparisons involving EDI, for all the other comparisons the correlation coefficients are > 0.91 and tend to increase slightly with the timescale. Moreover, all the comparisons are characterised by a linear relationship in a more or less wide range of values. In most cases, this linearity is lost when approaching the extreme lower (e.g. PN-SPI, DI-SPI, SPI-RAI) or higher values (e.g. PN-DI, DI-SPI, DI-RAI). Another aspect worthy of note is the sharp reduction in the range of values assumed by the PN index, passing from the 1-month to the 6-month timescale. In fact, as the accumulation scale increases, the probability of observing exceptionally different values from the mean is also reduced. This

contributes to a certain increment of the correlation coefficients between PN and the other indices with the timescale. However, this also denotes the lack of robustness of PN, which is, therefore, not very reliable for comparative analysis involving different timescales.

## 3.2 Cohen's Kappa test

As in the case of the correlation analysis, the indices were compared considering all the stations jointly. For each pair of indices, a comparative analysis of the severity classes assigned to each month of the time series was carried out, evaluating both the unweighted ($K$) and weighted ($K_w$) agreement statistics.

### 3.2.1 Detailed examples of agreement tables between drought indices

To understand the Kappa test's functioning, two detailed examples are first given in Tables 4 and 5, which show, respectively, the agreement matrices for the comparisons DI3-RAI3 and SPI3-RAI3, considering the data of all the stations. For Table 4, the sum of the agreement diagonal frequencies is $p_O = 0.8027$, while the proportion of agreement expected by chance, $p_e$, is 0.2645. For Table 5, $p_O = 0.5693$ and $p_e$ is 0.3219. Therefore, according to Eq. (5), the $K$ statistic is 0.73 and 0.36 for the DI3-RAI3 and SPI3-RAI3 comparisons, respectively. Both $K = 0.73$ and $K = 0.36$ are significant at the 1% significance level, but, according to Table 3, the agreement between DI3 and RAI3 is substantial, while the agreement between SPI3 and RAI3 is only fair. It should be noted that the correlation coefficients (Fig. 4) for the two previous comparisons are very high in both cases (0.97 and 0.99, respectively), which confirms that the comparisons between drought indices based on this coefficient are not very informative.

The weighted Kappa statistic $K_w$, calculated by Eq. (6) and adopting the weights indicated in parentheses in Tables 4 and 5, leads to $K_w = 0.93$ and $K_w = 0.72$ for DI3-RAI3 and SPI3-RAI3, respectively. As explained in Sect. 2.3, the $K_w$ statistic focusses on what happens in the disagreement cells. In this case, $K_w$ provides information aligned with that obtained by $K$. The fact that $K_w$ is higher than the corresponding $K$ indicates that, for both comparisons, the indices disagree much less than chance expectation where it counts very much (high weight) and disagree at about a chance level where it counts little (low weight).

In the example of Table 4, a substantial agreement ($K = 0.73$) is found between DI3 and RAI3. DI is an index having a precise statistical meaning as it is based on deciles. It is no coincidence that the line totals of Table 4 (i.e. the DI experimental frequencies for the different categories) are very close to the theoretical probabilities expected for the DI's categories (Table 1). Even if RAI does not have a precise statistical meaning, the similarity between the column and lines totals of Table 4 tells us that RAI tends to classify drought severities with logic not too far from that of DI. However, the classification operated by DI cannot be considered very effective since all the drought categories (moderate, severe, extreme) are based on the same decile interval and have the same expected occurrence probabilities of 10%. In contrast, one should expect decreasing probabilities passing from moderate to severe drought categories. In this sense, RAI provides a more reliable classification since the frequencies associated with moderate, severe, and extreme droughts decrease (18%, 13% and 6%, respectively). These decreasing frequencies are also observed for SPI (Table 5). However, the agreement between RAI3 and SPI3 is limited due to very relevant differences in the frequencies expected for the same drought categories (Table 5).

**Table 5** Observed frequencies of agreement $p_{oij}$ between SPI and RAI at a 3-month timescale (SPI3 and RAI3, respectively)

| | | RAI3 classification | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| | | Wet | Normal | Moderate Drought | Severe Drought | Extreme drought | $p_i$ |
| **SPI3** classification | Wet | **0.1609** (0) | 0.0001 (1) | 0.0002 (4) | 0.0000 (9) | 0.0000 (16) | 0.1612 |
| | Normal | 0.0914 (1) | **0.3766** (0) | 0.1779 (1) | 0.0417 (4) | 0.0000 (9) | 0.6876 |
| | Moderate Drought | 0.0000 (4) | 0.0000 (1) | **0.0003** (0) | 0.0807 (1) | 0.0056 (4) | 0.0866 |
| | Severe Drought | 0.0000(9) | 0.0000 (4) | 0.0000 (1) | **0.0076** (0) | 0.0331 (1) | 0.0407 |
| | Extreme Dr ought | 0.0000 (16) | 0.0000 (9) | 0.0000 (4) | 0.0000 (1) | **0.0239** (0) | 0.0239 |
| | $p_{\cdot j}$ | 0.2524 | 0.3767 | 0.1784 | 0.1300 | 0.0626 | 1.0000 |

Values in parentheses are the disagreement weights $v_{ij}$ assigned according to a "squared distance" pattern (Cohen 1968). Values in bold indicate the agreement diagonal. The total frequency of agreement is $p_O$ = 0.5693, the proportion of agreement expected by chance, $p_{e_i}$ is 0.3219. Overall data from 15 stations during 1951−2018

**Table 6** Values of the unweighted Cohen's statistic $K$ (Eq. 5) related to the two by two comparisons among indices at different time scales

| | PN1 | DI1 | PI1 | SPI1 | RAI1 | EDI |
|---|---|---|---|---|---|---|
| PN1 | 1 | | | | | |
| DI1 | **0.65** | 1 | | | | |
| PI1 | 0.27 | 0.37 | 1 | | | |
| SPI1 | 0.27 | 0.37 | **0.85** | 1 | | |
| RAI1 | **0.74** | **0.66** | 0.37 | 0.37 | 1 | |
| EDI | 0.16 | 0.20 | 0.27 | 0.27 | 0.19 | 1 |
| | PN3 | DI3 | PI3 | SPI3 | RAI3 | EDI |
| PN3 | 1 | | | | | |
| DI3 | 0.60 | 1 | | | | |
| PI3 | 0.47 | 0.35 | 1 | | | |
| SPI3 | 0.46 | 0.37 | **0.85** | 1 | | |
| RAI3 | **0.66** | **0.73** | 0.35 | 0.36 | 1 | |
| EDI | 0.36 | 0.31 | 0.47 | 0.47 | 0.31 | 1 |
| | PN6 | DI6 | PI6 | SPI6 | RAI6 | EDI |
| PN6 | 1 | | | | | |
| DI6 | 0.53 | 1 | | | | |
| PI6 | **0.61** | 0.35 | 1 | | | |
| SPI6 | **0.61** | 0.37 | **0.85** | 1 | | |
| RAI6 | 0.55 | **0.78** | 0.36 | 0.38 | 1 | |
| EDI | 0.49 | 0.32 | 0.54 | 0.54 | 0.32 | 1 |

Values in bold indicate substantial or almost perfect agreements (according to Table 3)

In the next section, all the other comparisons will be synthetically described in terms of $K$ and $K_w$.

### 3.2.2 Synthetic results

The values of the unweighted (Eq. 5) and weighted Kappa (Eq. 6) statistics for all the two-by-two comparisons considered are shown in Tables 6 and 7, respectively. All the values are significant ($\alpha = 1\%$), denoting a degree of agreement among indices higher than that expected by chance. On the other hand, the strength of agreement varies considerably with both the indices and the timescale. Considering the unweighted $K$ (Table 6), it can be observed that a substantial or even perfect agreement (bold values) at all the timescales considered is only present for the pairs DI-RAI and PI-SPI. A substantial agreement is also present for the pairs PN1-RAI1, PN3-RAI3, PN1-DI1, PN6-SPI6, and PN6-PI6. For all the other comparisons, the agreement is moderate to fair, or even slight, such as for the comparisons PN1-EDI and RAI1-EDI. The worst agreement at all the timescales can be found for the pairs DI-EDI and RAI-EDI.

In general, the analysis of the weighted Kappa statistic $K_w$ (Table 7) does not carry relevant additional information with respect to $K$, indicating that even by applying weights to the disagreement cells, the assessment of the strength of agreement/disagreement does not

**Table 7** Values of the weighted Cohen's statistic, $K_w$ (Eq. 6) related to all possible two by two comparisons among indices at different time scales

|      | PN1 | DI1 | PI1 | SPI1 | RAI1 | EDI |
|------|-----|-----|-----|------|------|-----|
| PN1  | 1   |     |     |      |      |     |
| DI1  | 0.92 | 1  |     |      |      |     |
| PI1  | 0.69 | 0.73 | 1 |      |      |     |
| SPI1 | 0.69 | 0.73 | 0.94 | 1  |      |     |
| RAI1 | 0.93 | 0.92 | 0.71 | 0.71 | 1 |     |
| EDI  | 0.38 | 0.38 | 0.42 | 0.42 | 0.38 | 1 |
|      | PN3 | DI3 | PI3 | SPI3 | RAI3 | EDI |
| PN3  | 1   |     |     |      |      |     |
| DI3  | 0.81 | 1  |     |      |      |     |
| PI3  | 0.73 | 0.70 | 1 |      |      |     |
| SPI3 | 0.73 | 0.72 | 0.94 | 1  |      |     |
| RAI3 | 0.86 | 0.93 | 0.71 | 0.72 | 1 |     |
| EDI  | 0.62 | 0.54 | 0.66 | 0.67 | 0.55 | 1 |
|      | PN6 | DI6 | PI6 | SPI6 | RAI6 | EDI |
| PN6  | 1   |     |     |      |      |     |
| DI6  | 0.72 | 1  |     |      |      |     |
| PI6  | 0.75 | 0.7 | 1 |      |      |     |
| SPI6 | 0.75 | 0.73 | 0.94 | 1  |      |     |
| RAI6 | 0.75 | 0.95 | 0.72 | 0.75 | 1 |     |
| EDI  | 0.68 | 0.59 | 0.74 | 0.75 | 0.61 | 1 |

Weights are assigned according to a squared distance pattern (as detailed in Tables 4 and 5)

change significantly. Really, with more detailed analysis, it is possible to describe some interesting cases that allow us to appreciate the added value of $K_w$ with respect to $K$. An example is represented by the results obtained for the pairs PN3-EDI and DI3-SPI3. As shown in Tables 6 and 7, these pairs have very similar $K$ (0.36 and 0.37, respectively), but they differ more when the $K_w$ is considered (0.62 and 0.72, respectively). This indicates that, despite a very similar behaviour in the agreement diagonal (quantified by $K$), the two pairs have quite different behaviour in the disagreement cells (quantified by $K_w$). In particular, for the pair PN3-EDI, the proportions of events classified differently by the two indices (and probably with weights higher than 1) are higher than those obtained for the pair DI3-SPI3.

# 4 Discussion

It is observed that almost all the indices considered have a high degree of positive correlation between the corresponding values of monthly severity (Fig. 4). The only exception is represented by the comparisons between the EDI and the other indices computed at the 1-month time scales. At any rate, an analysis of Fig. 4 and the corresponding correlation

coefficients could lead to believing that, in most cases, the indices considered provide similar information about drought severity.

However, the indices are much less in agreement than they appear by a correlation analysis when they are compared in terms of drought severity classification (Tables 6, 7). To analyse in more detail the different information deriving from the correlation analysis and the Cohen's Kappa test, in Fig. 5 the correlation coefficients $r$ are plotted versus the corresponding unweighted $K$ statistics.

From Fig. 5, it is evident that several pairs of indices are characterised by a very similar correlation but a very different strength of agreement. For example, considering Fig. 5b, in correspondence of the correlation range $0.9 < r < 1$, there are pairs of drought indices with agreement fair (PI3-RAI3, DI3-PI3, SPI3-RAI3, DI3-PI3), moderate (PN3-SPI3, PN3-PI3, PN3-DI3), substantial (PN3-RAI3, DI3-RAI3), and almost perfect (PI3-SPI3).

Similar considerations can also be made for the comparisons concerning the EDI. Considering Fig. 5b again as an example, we find pairs in fair (RAI3-EDI, DI3-EDI, PN3-EDI) and moderate agreement (PI3-EDI and SPI3-EDI) in correspondence of very similar correlations ($0.82 < r < 0.85$). Even, the pair RAI3-EDI has a correlation higher than SPI3-EDI, but, based on the $K$ value, the similarity between SPI3 and EDI is higher than that between RAI3 and EDI.

In general, there are two main reasons why two drought indices may disagree on the severity category assigned to a given month of the time series. The first reason, of course, lies in the severity classification criteria (Table 1). The other is related to actual differences in the index calculation methods, despite that the reference variable is always represented by precipitation.

An enlightening example of the first reason is the result obtained for the comparisons DI-SPI and PI-SPI. Despite DI and PI deriving from the same quantitative values (empirical cumulative frequencies of precipitation in a given period), they differ for the severity classification criterion (Table 1). In particular, the PI classes have been defined in this paper based on a probabilistic correspondence with the SPI classification. Thus, SPI and PI's agreement is almost perfect ($K = 0.85$ for all the time scales), while between SPI and DI it is only fair ($K = 0.37$ for all the time scales). This also indicates that for central Italy's climatic conditions, a classification of drought events based on the PI index (under the further hypothesis of sufficiently long series) could be effective and analogous with that obtained from SPI. The advantage in the use of PI instead of SPI would be the greater ease of calculation and, of course, the ease of understanding for a wider audience.

Another example of the effect of the severity classification criterion can be found by analysing the similarity between RAI and SPI. The poor agreement between RAI and SPI (Tables 5, 6, 7, and Fig. 5) is determined by the mismatch in the classification criteria because the two indices show an excellent correlation at all the timescales (Fig. 4). In fact, by changing the class limits adopted for RAI (Table 1) to 1.5, -2.4, -3, and -3.3 (instead of 1, -1, -2, and -3), there could be a relevant increment of the Kappa statistic (0.85, 0.86, and 0.82, for the 1-, 3-, and 6-month timescales, respectively), which indicates an almost perfect agreement between the two indices.

The other reason for the presence of a limited agreement resides in the different algorithms adopted for calculating the index. The most striking example of this situation is represented by the comparisons involving the EDI index. As explained in Sect. 2.2.5, the EDI is based on the precipitation of the previous periods cumulated using weights that decrease as the temporal distance from the current period increases (Byun and Whilite, 1999). This specific calculation approach leads to a general moderate agreement between the EDI and the other indices (Tables 6, 7, and Fig. 5). The fact that EDI's agreement with

**Fig. 5** Comparison between the correlation coefficient and the Cohen'K for all the pairs of drought indices considered. The pairs have been divided into three graphs, relating to the 3 time-scales considered for the indices PN, DI, PI, SPI and RAI: **a** 1 month; **b** 3 months; **c** 6 months. The classification of the strength of agreement is based on Table 1
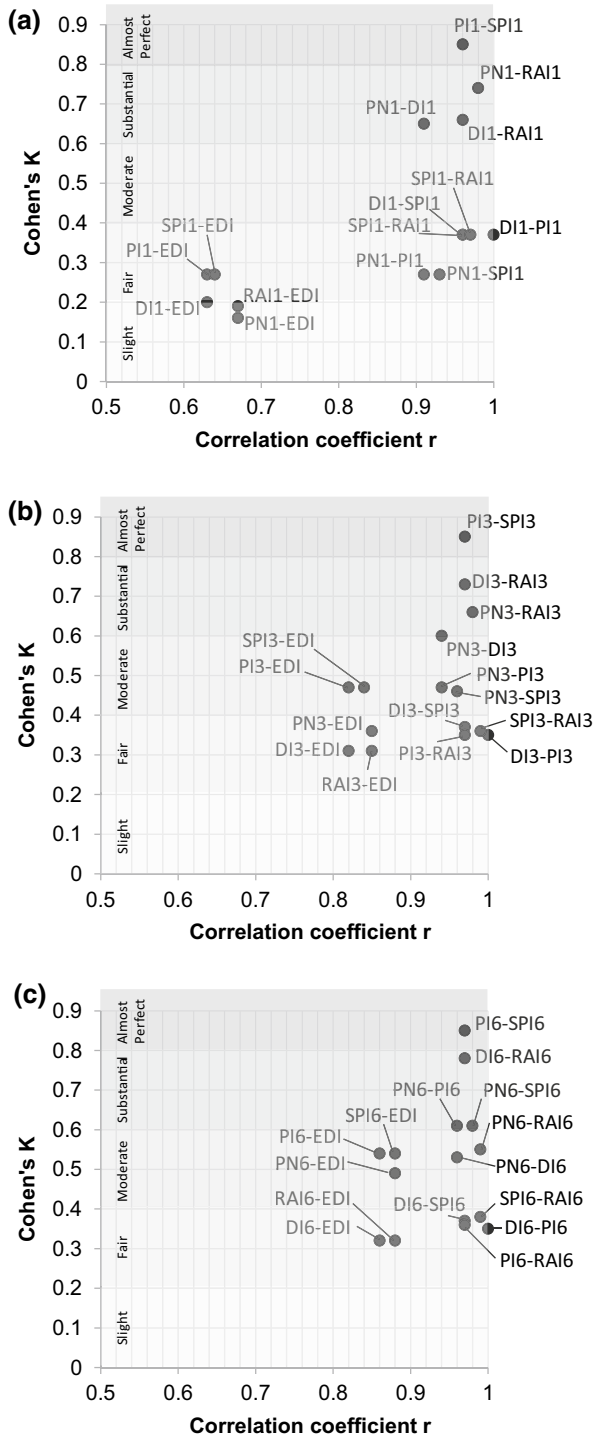
the other indices increases when these are calculated at longer timescales (e.g. 6 months) is the normal consequence of EDI accounting for precipitation over a period prior to 365 days. Therefore, even if the calculation method remains conceptually very different, it is certainly expected that the similarity between the EDI and the timescale-dependent indices increases when the latter are calculated on medium-long timescales. This was also observed by Jain et al. (2015), which, based on the correlation coefficient, found a maximum similarity between EDI and the timescale-dependent indices (e.g. SPI) when these were calculated on a 9-month scale. However, even for EDI, one can appreciate the more reliable assessment of similarity provided by Cohen's Kappa test with respect to the correlation coefficient (Figs. 4 and 5). In particular, the correlation of EDI with the other indices at the 6-month timescale (Fig. 5b) is very high (average r ≈ 0.88), but this conceals an actual agreement between drought severity classification that is only fair or moderate. Therefore, the comparison of drought indices, based on Cohen's Kappa test, appears particularly useful in highlighting the specificities of some indices, as in the case of EDI.

The SPI provides another example of the calculation algorithm's effect on the strength of agreement between drought indices. The SPI is the only index among those considered which requires a preliminary fitting of the precipitation data to a theoretical probability distribution (i.e. the Gamma distribution). The most direct effect of this aspect can be seen in the agreement with PI, which is very high ($K = 0.85$) but not perfect, despite the imposed correspondence among the severity classification criteria. Indeed, PI is based on the implicit assumption that precipitation is normally distributed. In the region considered (Mediterranean climate), the consequences of these two different hypotheses, regarding the probability distribution of the precipitation, appear moderated, but in other climatic conditions (e.g. drier), there could more marked effects on the agreement. For example, a detailed analysis of the variability of $K$ with the month and the timescale indicated that the agreement between SPI and the other indices (PI included) tends to be lower in the spring–summer period (i.e. dry period), particularly at the 1-month timescale. This can be likely attributed to the fact that the precipitation of these months is more positively skewed than that of the winter–spring period (due to the relevant presence of low or even zero values). In these circumstances, the goodness of fit of the Gamma distribution is much better than that of normal distribution, and larger differences among the indices computed under the two underlying distributions are expected.

No significant differences in the strength of agreement were detected among the 15 stations. This is certainly due to the relatively small extent of the region considered, and thus to the limited climatic differences among stations (Figs. 2 and 3). In other climatic conditions, the strength of agreement among indices could significantly differ from those reported in this paper.

## 5 Conclusion

In this paper, the agreement in the drought severity classification obtained from some popular meteorological drought indices (Percentage of Normal PN, Deciles Index DI, Percentile Index PI, Rainfall Anomaly Index RAI, Standardised Precipitation Index SPI, and Effective Drought Index EDI) has been evaluated for a typical Mediterranean region, based on a 68-year time series. The main objective was to illustrate the advantage of a comparison technique based on Cohen's Kappa test (Cohen 1960, 1968), with respect to the commonly adopted Pearson correlation coefficient.

The Cohen's Kappa test proved to be more effective and flexible than the correlation analysis in evaluating the actual similarity between drought indices. In particular, the Cohen's Kappa test indicates a prevailing moderate or fair agreement between the indices considered, despite the generally very high correlation between the corresponding severity times series.

The first reason for the lack of agreement resides in the severity classification criteria. New classification criteria have been proposed for DI (defining a new index based on percentiles, PI), which showed an almost perfect agreement with SPI. It was also shown that a revision of the severity classification criteria usually adopted for RAI could lead to an almost perfect agreement with the SPI.

The strength of agreement between the indices is also affected by the index calculation method. For this aspect, EDI is the index that is less in agreement with the other indices, due to the specificity of its calculation algorithm based on the concept of effective precipitation. This difference does not emerge clearly if the EDI is compared with the other indices on the basis of a traditional correlation analysis.

For the other indices (PN, PI, DI, SPI, RAI), the differences due to the calculation have a limited influence on the agreement between indices, probably because the reference variable is always the same (i.e. cumulative precipitation in a given timescale). However, in different climatic conditions (i.e. drier than those examined in the present paper), there could be larger differences between indices, particularly between those implying different underlying probability distribution of cumulative precipitation. For example, by analysing the intra-annual variability of the agreement, it was found that it tends to be lower in the summer months, which, in the region considered, have the lowest precipitation amounts and sometimes include zero values at the shortest timescale (1 month).

The procedure illustrated in the paper appears exportable to any other area in order to obtain an effective evaluation of the strength of agreement among any set of drought indices.

**Declarations**

permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

# References

Abramowitz M, Stegun IA (1965) Handbook of mathematical function. Dover Publications

Blauhut V, Stahl K, Gudmundsson L (2015) Towards pan-European drought risk maps: quantifying the link between drought indices and reported drought impacts. Environ Res Lett 10(1):10. https://doi.org/10.1088/1748-9326/10/1/014008

Byun HR, Wilhite DA (1999) Objective quantification of drought severity and duration. J Clim 12:2747–2756

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:7–46

Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70(4):213–220

Craig RT (1981) Generalization of Scott's index of intercode agreement. Public Opinion Quart 45:260–264

Deo RC, Byun HR, Adamowski JF, Kim DW (2015) A real-time flood monitoring index based on daily effective precipitation and its application to Brisbane and Lockyer Valley flood events. Water Resour Manag 29:1–19

Deo RC, Byun HR, Adamowski JF, Begum K (2017) Application of effective drought index for quantification of meteorological drought events: a case study in Australia. Theor Appl Climatol 128:359–379

Di Lena B, Vergni L, Antenucci F, Todisco F, Mannocchi F (2014) Analysis of drought in the region of Abruzzo (Central Italy) by the Standardized Precipitation Index. Theor Appl Climatol 115(1–2):41–52

Ezzine H, Bouziane A, Ouazar D (2014) Seasonal comparisons of meteorological and agricultural drought indices in Morocco using open short time-series data. Int J Appl Earth Obs Geoinf 26:36–48

Fluixá-Sanmartín J, Pan D, Fischer L, Orlowsky B, García-Hernández J, Jordan F, Haemmig C, Zhang F, Xu J (2018) Searching for the optimal drought index and timescale combination to detect drought: a case study from the lower Jinsha River basin. China, Hydrol Earth Syst Sci 22:889–910

Folger P (2017) Drought in the United States: Causes and Current understanding. In: CRS Report R43407

Gamer M, Fellows J, Lemon I, Singh P (2019) Package "irr". Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1 https://CRAN.R-project.org/package=irr

Gibbs WJ, Maher JV, (1967) Rainfall deciles as drought indicators. Bureau of Meteorology Bulletin No. 48. Commonwealth of Australia, Melbourne

Guttman NB (1998) Comparing the Palmer drought index and the Standardized Precipitation Index. J Am Water Resour Assoc 34:113–121

Jain VK, Pandey RP, Jain MK, Byun H (2015) Comparison of drought indices for appraisal of drought characteristics in the Ken River Basin. Weather Clim Extremes 8:1–11

Keyantash JA, Dracup JA (2002) The quantification of drought: an evaluation of drought indices. Bull Am Meteorol Soc 83(8):1167–1180

Kogan FN (2000) Contribution of remote sensing to drought early warning. In: Wilhite DA, Sivakumar MVK, Wood DA (eds) Early warning systems for drought preparedness and drought management, Proceedings of an expert group meeting held on warning systems for drought preparedness and drought management. Lisbon, Portugal

Hayes MJ (2006) Drought indices. Van Nostrand's scientific encyclopedia. Wiley, London. https://doi.org/10.1002/0471743984.vse8593

Hayes M, Svoboda M, Wall N, Widhalm M (2011) The Lincoln declaration on drought indices: universal meteorological drought index recommended. Bull Am Meteorol Soc 92:485–488

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Lloyd-Hughes B, Saunders MA (2002) A drought climatology for Europe. Int J Climatol 22:1571–1592

McKee TB, Doesken NJ, Kliest J (1993) The relationship of drought frequency and duration time scales. In: Proceedings of the 8th international conference on applied climatol. Boston, MA, pp 179–184

Morid S, Smakhtin V, Moghaddasi M (2006) Comparison of seven meteorological indices for drought monitoring in Iran. Int J Climatol 26(7):971–985

Pandey RP, Dash BB, Mishra SK, Singh R (2008) Study of indices for drought characterization in KBK districts in Orissa (India). Hydrol Process 22:1895–1907

Salehnia N, Alizadeh A, Sanaeinejad H, Bannayan M, Zarrin A, Hoogenboom G (2017) Estimation of meteorological drought indices based on AgMERRA precipitation data and station station-observed precipitation data. J Arid Land 6:797–809

Schober P, Boer C, Schwarte LA (2018) Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg 126(5):1763–1768

Smakhtin VU, Hughes DA (2007) Automated estimation and analyses of meteorological drought characteristics from monthly rainfall data. Environ Model Softw 22:880–890

Stagge JH, Tallaksen LM, Gudmundsson L, Van Loon AF, Stahl K (2015) Candidate distributions for climatological drought indices (SPI and SPEI). Int J Climatol 35:4027–4040

Teweldebirhan Tsige D, Uddameri V, Forghanparast F, Hernandez EA, Ekwaro-Osire S (2019) Comparison of meteorological and agriculture-related drought indicators across Ethiopia. Water 11(11):2218

van Rooy MP (1965) A rainfall anomaly index independent of time and space. Notos 14:43–48

Vergni L, Di Lena B, Chiaudani A (2016) Statistical characterization of winter precipitation in the Abruzzo region (Italy) in relation to the North Atlantic Oscillation (NAO). Atmos Res 178–179:279–290

Vergni L, Di Lena B, Todisco F, Mannocchi F (2017) Uncertainty in drought monitoring by the Standardized Precipitation Index: the case study of the Abruzzo region (central Italy). Theor Appl Climatol 128(1–2):13–26

Wable PS, Jha MK, Shekhar A (2019) Comparison of Ddrought Indices in a semi-arid river basin of India. Water Resour Manag 33:75–102

WMO (2006) Drought monitoring and early warning: Concepts, progress and future challenges. WMO. Available from: http://www.wamis.org/agm/pubs/brochures/WMO1006e.pdf

Wilhite A, Svoboda D, Hayes J (2007) Understanding the complex impacts of drought: a key to enhancing drought mitigation and preparedness. Water Resour Manag 21:763–774

Wu H, Svoboda MD, Hayes MJ, Wilhite DA, Wen F (2007) Appropriate application of the standardized precipitation index in arid locations and dry seasons. Int J Climatol 27:65–79

Zargar A, Sadiq R, Naser B, Khan FI (2011) A review of drought indices. Environ Rev 19:333–349