

Selection of appropriate time scale with Boruta algorithm for regional drought monitoring using multi-scalar drought index

By ZULFIQAR ALI^{1,2}, IJAZ HUSSAIN^{1*}, MUHAMMAD FAISAL^{3,4}, ELSAYED ELSHERBINI ELASHKAR⁵, SHOWKAT GANI⁶, and MUHAMMAD AHMED SHEHZAD⁷, ¹*Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan;* ²*Johann Bernoulli Institute (JBI) Rijksuniversiteit Groningen, Groningen, The Netherlands;* ³*Faculty of Health Studies, University of Bradford, Bradford, UK;* ⁴*Bradford Teaching Hospitals NHS Foundation Trust, Bradford Institute for Health Research, Bradford, UK;* ⁵*Arriyadh Community College King Saud University, Riyadh, Saudi Arabia;* ⁶*College of Business Administration, King Saud University Muzahimiyah, Al Muzahimiyah, Saudi Arabia;* ⁷*Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan*

(Manuscript received 26 July 2018; in final form 3 April 2019)

ABSTRACT

Spatial distribution of drought plays key role specifically in hydrological research. Drought is a complex stochastic natural hazard caused by prolonging shortage of rainfall and available water resources. The multi-scalar drought indices (based on probability distribution) are commonly used for making effective drought mitigation policies. In addition, the multi-scalar drought indices have great extent of the inaccurate determination of drought classes due to its probabilistic nature. However, the interpretation and applicability of various time scales are cumbersome for multi-scalar drought in various meteorological stations at a particular region. In this regards, accurate estimation and continuous monitoring of future drought at regional level requires a more appropriate and important time scale with respect to regions under study. In this study, we aimed to investigate the appropriate time scale for multi-scalar drought indices by using geo-reference points of meteorological stations. We used Boruta algorithm with two random forest adapters of machine learning algorithms for regionalized optimization of drought monitoring time scale. We explored the appropriate time scale for the Standardized Precipitation Temperature Index (SPTI) of 52 meteorological stations that are located across Pakistan. Results show that the significant importance of SPTI-1 (1-month time scale) for further spatial and temporal studies. That is, being high ranked and prominence, SPTI-1 has the ability to capture the characteristics of all other time scales that are in some circumstances applicable for drought characterization and classification.

Keywords: multi-scalar drought index, drought monitoring, random forest

1. Introduction

Spatial distribution of drought plays key role specifically in hydrological research. Drought is a complex natural hazard which can severely affect on a wide range of land, agricultural and socioeconomic infrastructure. Drought can be defined as ‘an extended period of deficient rainfall relative to the multi-year mean for a region’ (Schneider et al., 2011). Probably, all the climatic zones are suffering from drought, however, its characteristics may vary from zone to zone. Several factors are involved in drought

occurrence such as, high wind, low relative humidity, temperature, characteristics and duration of rain, intensity and onset (Wilhite, 1994). In addition, changed climatic conditions (Dawadi and Ahmad, 2013) and contaminated water are also causing water scarcity (Aswathanarayana, 2001). However, demand of water has been increased since past two decades because of enhanced agricultural and industrial activities. Therefore, it is necessary for drought management to define drought conditions for the region under consideration (Wilhite, 2000).

*Corresponding author. e-mail: ijaz@qau.edu.pk

There are several tools for the assessment of drought, but drought indices are the most commonly used because of its probabilistic structure and universal acceptability around the globe. Drought indices may help in developing effective drought mitigation policies (Kogan, 2000; Hayes et al., 2004). Moreover, for the precise and accurate monitoring and quantification of drought, a single climatic variable for drought indices is insufficient (Vicente-Serrano et al., 2010), as multiple factors are involved in drought phenomena such as, precipitation, humidity, wind speed, temperature, and heat waves (Wilhite, 1994; Sheffield et al., 2012; Hao and Singh, 2015; Vicente-Serrano et al., 2016). In preliminary stage of the evaluation of drought indices, Palmer et al. (1965) introduced Palmer Drought Severity Index (PDSI) for the characterization and quantification of drought based on precipitation, temperature, latitude and available water capacity. In recent research, it was found that PDSI has the lack of multi-scaling drought study and ability to compare regions (Wang et al., 2017). Various studies have been conducted to assess the suitability of different drought indices in several applications (Guttman, 1998; Keyantash and Dracup, 2002; Musuuza et al., 2016; Tigkas et al., 2016; Adnan et al., 2018). A comprehensive review on the list of available drought indices corresponding with their data requirement is available (Svoboda et al., 2016).

Parallel to including multiple drought indices in drought monitoring protocols, the role of non-parametric methods have significant impact to explore the information about extreme drought classes (Ghamghami et al., 2017). In previous research, numerous authors used probability distributions to obtain standardized drought values. Ntale and Gan (2003) used Pearson type III (P3) distribution and probability plotting method to improve Standardized Precipitation Index (SPI) values. Blain and Meschiatti (2015) provided an alternate procedure to quantify the risk of drought by introducing generalized normal distribution in SPI model. Recently, Stagge et al. (2015) proposed multiple distribution for various drought indicators such as SPI and Standardized Precipitation Evapotranspiration Index (SPEI). One of the major advantage of multi-scalar drought indices is to capture drought classification states at various time scales that are essential for both assessing drought in relation to different hydrological systems, and differentiating among different drought types. However, in probabilistic models of each multi-scalar drought indices, uncertainty in accurate and precise estimation of drought classes always exists (Parker, 2014). Further, the selection of probability distributions in each index is purely subjective in nature. Moreover, it is difficult to include all the time scales in drought monitoring protocol for efficient and accurate

drought mitigation policies (Hao and AghaKouchak, 2013; Hao et al., 2017).

A non-parametric approach – a graphical method that is based on probability plotting position formula is suggested as a substitute of probability distribution. The fact of using graphical method is to capture the uncertainty of extreme events and hence to reduce errors in accurate and precise estimation of drought indices. Initially, Farahmand and AghaKouchak (2015) gave an idea of non-parametric for drought monitoring. They used Gintorten (Gintorten, 1963) probability position formula as an alternative of Gamma distribution for obtaining standardized drought index. Several other studies used non-parametric approach for drought monitoring (Ghamghami et al., 2017).

In previous research, various studies used different multivariate techniques to address the multi-scaling issue. Bazrafshan et al. (2014) used principle component analysis (PCA) technique to overcome the issue in the selection of appropriate time scale by introducing Multivariate Standardized Precipitation Index (MSPI). Several studies used different time scales to monitor both spatial and temporal behaviour of drought for various climatic regions (Mishra and Desai, 2005; Patel et al., 2007; Fiorillo and Guadagno, 2010). Dogan et al. (2012) compared different drought indices using precipitation data with the different time scales and indicated that while comparing drought indices, appropriateness of important time scale has great extent in drought monitoring context. One can quantify a big picture of drought for any region under study without requiring any complex environmental data because of multi-scaling characteristic. Guttman (1998) compared different drought indices using precipitation data with different time scales and highlighted that the appropriate selection of time scale has great importance in drought monitoring. Regardless of temporal time scale, classification of affected system due to drought severity for any specified region, spatial scale also contributes significantly (Tsakiris et al., 2013).

In this study, we aimed to search more relevant and important time scales for Standardized Precipitation Temperature Index (SPTI) (Ali et al., 2017) at regional level. We proposed Boruta (Kursa and Rudnicki, 2010) algorithm from the list of machine learning methods for the analysis of 52 meteorological stations of Pakistan. We hypothesises that ranking important time scale using a feature selection approach will lead to overcome the computational difficulties in existing multi-scalar drought indices for future drought monitoring at regional level. Consequently, efficient use of selected drought time scale is possible to inference spatially referenced meteorological stations.

Table 1. Drought classification criteria of SPTI.

Range of SPTI values	Drought classes
≥ 2	Extremely wet (EW)
1.5 to 1.99	Severely wet (SW)
1 to 1.49	Moderate wet (MW)
0.99 to -0.99	Near normal (NN)
-1 to -1.49	Moderate drought (MD)
-1.5 to -1.99	Severe drought (SD)
≤ -2	Extreme drought (ED)

2. Methodology

2.1. SPTI – the multi-scalar drought index

There are several procedures to report drought severity using multi-scalar drought index. McKee et al. (1993) developed an SPI drought index, which is based on long term precipitation record to quantify the precipitation scarcity. One of the major advantage of SPI index is that it can be used to monitor drought at various time scales. Estimation of quantitative values of SPI can be made by normalizing appropriate probability distributions of the observed monthly cumulative precipitation time series. Consequently, negative and positive SPI values designate less than or greater than median precipitation, respectively.

Parallel to McKee et al. (1993) and Vicente-Serrano et al. (2010) developed a new multi-scalar drought index: the Standardized Precipitation Evapotranspiration Index (SPEI). In SPEI, water balance model based on the difference between precipitation and potential evapotranspiration (PET) is used with the same mathematical procedure of SPI. One major advantage of SPEI over SPI is to include the effect of evaporation to characterize the regions under study.

Following the same methodology of SPI and SPEI, Ali et al. (2017) developed a new multi-scalar drought index; SPTI, to capture drought characterization in both cold and hot climate regions. However, among the list of available drought indices, appropriate selection of particular depends on the availability of data and the nature of climatic zone.

In this paper, we use multi-scalar drought index – SPTI (Ali et al., 2017) for our case study catchment because of three reasons: (1) the climate of the selected regions has both cold and hot weather, (2) SPTI gives accurate results for regions having low temperature, and (3) SPTI works without any mathematical conflict.

A brief explanation of the stepwise estimation procedure of SPTI is as follows. In the first step, De Marton Aridity Index (DAI) is computed using monthly total precipitation and average monthly temperature for each selected station by the following equation.

$$DAI_i = \frac{P_i}{10 + T_i} \quad (1)$$

Where P_i is the monthly total precipitation and T_i is the mean monthly temperature.

Second step consists on the candidacy of the appropriate probability distribution for DAI_i series of each station. In this study, 32 most commonly used probability distributions were used to observe the most appropriate probability distribution. The list of these distributions is available in *propagate* (Spiess, 2014) package in R package.

In third step, fitted distributions are selected for each station's time series data of DAI_i that have minimum value of Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). For each station, the Cumulative Distribution Function (CDF) of fitted distribution is standardized by using the following method:

$$H(x) = q + (1-q)F(x) \quad (2)$$

In Equation 2, a little modification in the CDF is made to adjust the effect of undefined values in DAI. For example, in case of gamma distribution q is the probability of zero value in the time series data DAI for each station. If m be the number of zeros present in DAI_i time series data, then q is estimated by m/n , where n is the total number of observation in the DAI_i time series. Here, instead of the abovestated method, other well-known methods of probability plotting position such as Cunnane (1978) can be adopted for adjusting probability of undefined values in the CDF.

In last step, the standardized values of SPTI index are estimated by the following approximation procedure which is available in Abramowitz and Stegun (1965).

$$SPTI = - \left(P + \frac{c_0 + c_1 P + c_2 P^2}{1 + d_0 P + d_1 P^2 + d_2 P^3} \right) \quad (3)$$

for

$$P = \sqrt{\ln \left[\frac{1}{\{H(x)\}^2} \right]}$$

when

$$0 < H(x) \leq 0.5 \quad (4)$$

$$SPTI = + \left(P - \frac{c_0 + c_1 P + c_2 P^2}{1 + d_0 P + d_1 P^2 + d_2 P^3} \right) \quad (5)$$

and for,

$$P = \sqrt{\ln \left[\frac{1}{\{1 - H(x)\}^2} \right]}$$

when

$$0.5 < H(x) \leq 1 \quad (6)$$

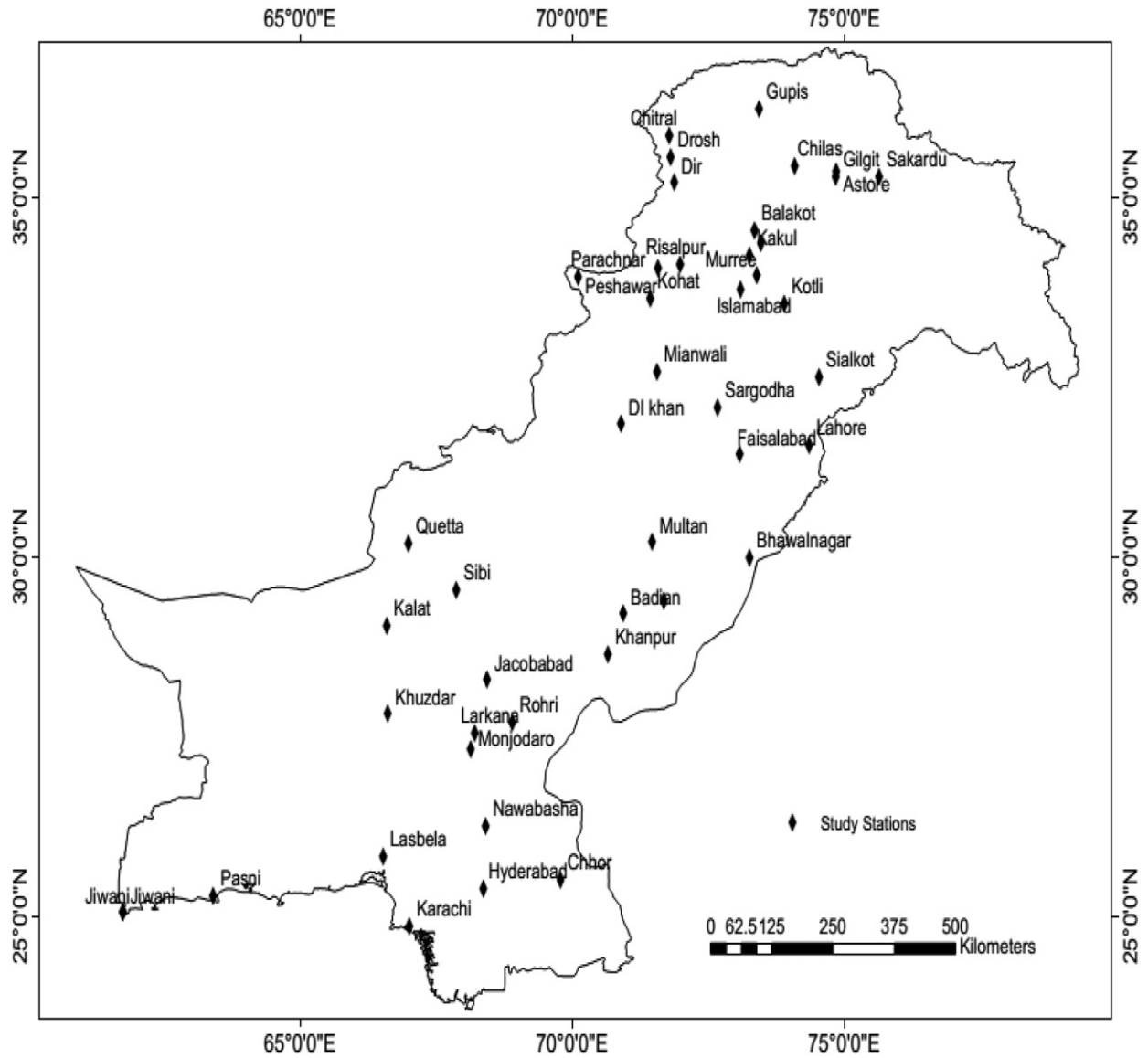


Fig. 1. Locations of stations in Pakistan.

Table 2. CDF of selected distribution for various time scale of SPTI at Chirat station.

Time scale	Distribution	CDF	AIC
DAI-1	3P Weibul distribution	$F(x) = \begin{cases} 1 - e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$	508.519
DAI-3	Generalized extreme value distribution	$F(x) = e^{-t(x)}, \text{ where } t(x) = \begin{cases} (1 + \xi(\frac{x-u}{\sigma}))^{\frac{1}{\xi}} & \text{iff } \xi \neq 0 \\ e^{-(x-u)/\sigma} & \text{iff } \xi = 0 \end{cases}$	542.364
DAI-6	Generalized trapezoida distribution	Complex: see van Dorp and Kotz (2003)	691.105
DAI-9	Gumbel distribution	$F(x) = e^{-e^{-(x-u)/\beta}}$	827.091
DAI-12	Generalized trapezoidal distribution	Complex: see van Dorp and Kotz (2003)	997.765
DAI-24	Johnson's SU-distribution	$F(x) = \Phi(\gamma + \delta \sinh^{-1}(\frac{x-\zeta}{\lambda}))$	777.716
DAI-48	Laplace distribution	$F(x) = \begin{cases} \frac{1}{2} \exp(-\frac{x-u}{\sigma}) & \text{if } x \geq u \\ 1 - \frac{1}{2} \exp(-\frac{x-u}{\sigma}) & \text{if } x < u \end{cases}$	1058.08

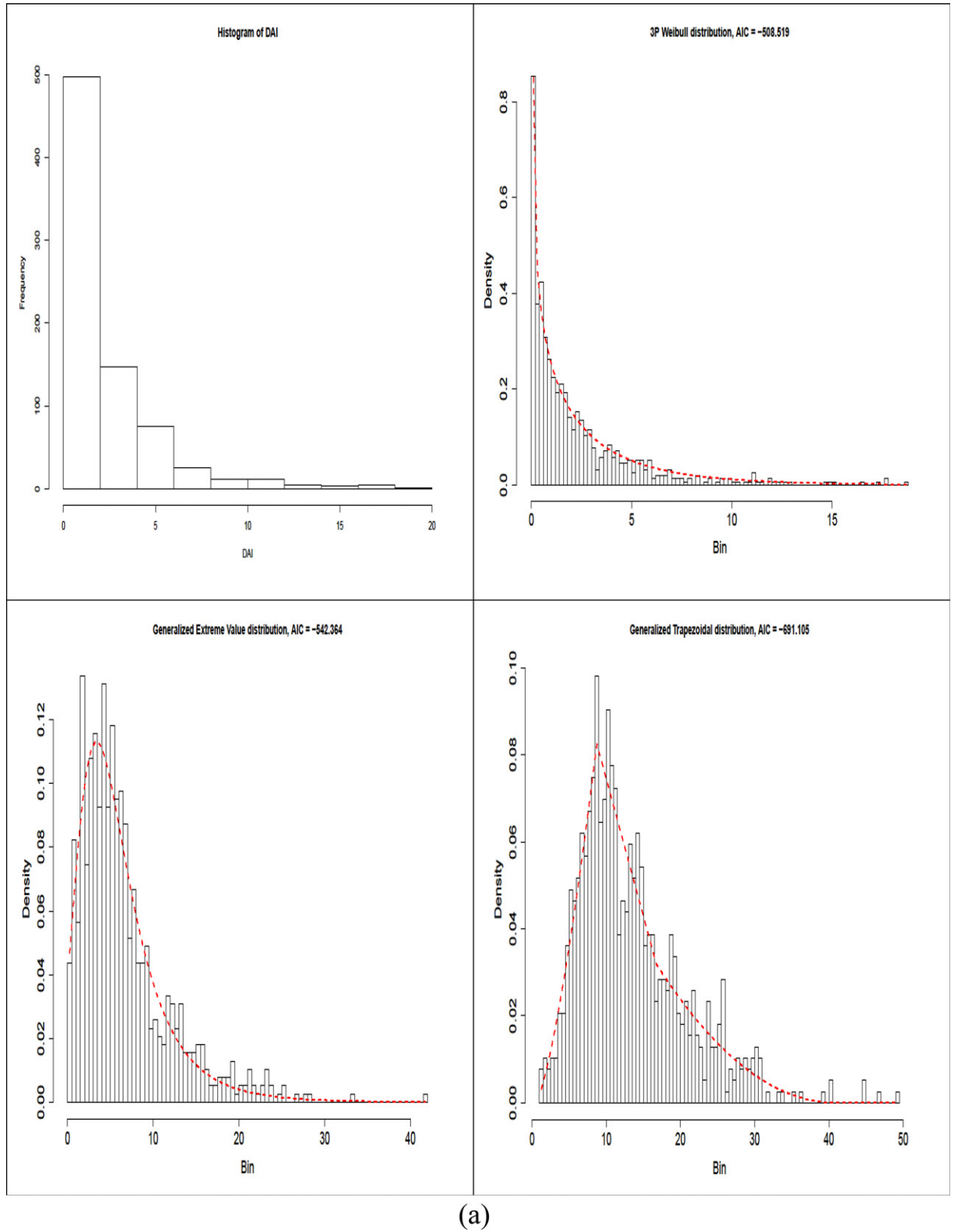


Fig. 2. (a) Theoretical vs. empirical histograms of selected distribution of various time scale of SPTI at Chirat station. (b) Theoretical vs. empirical histograms of selected distribution of various time scale of SPTI at Chirat station.

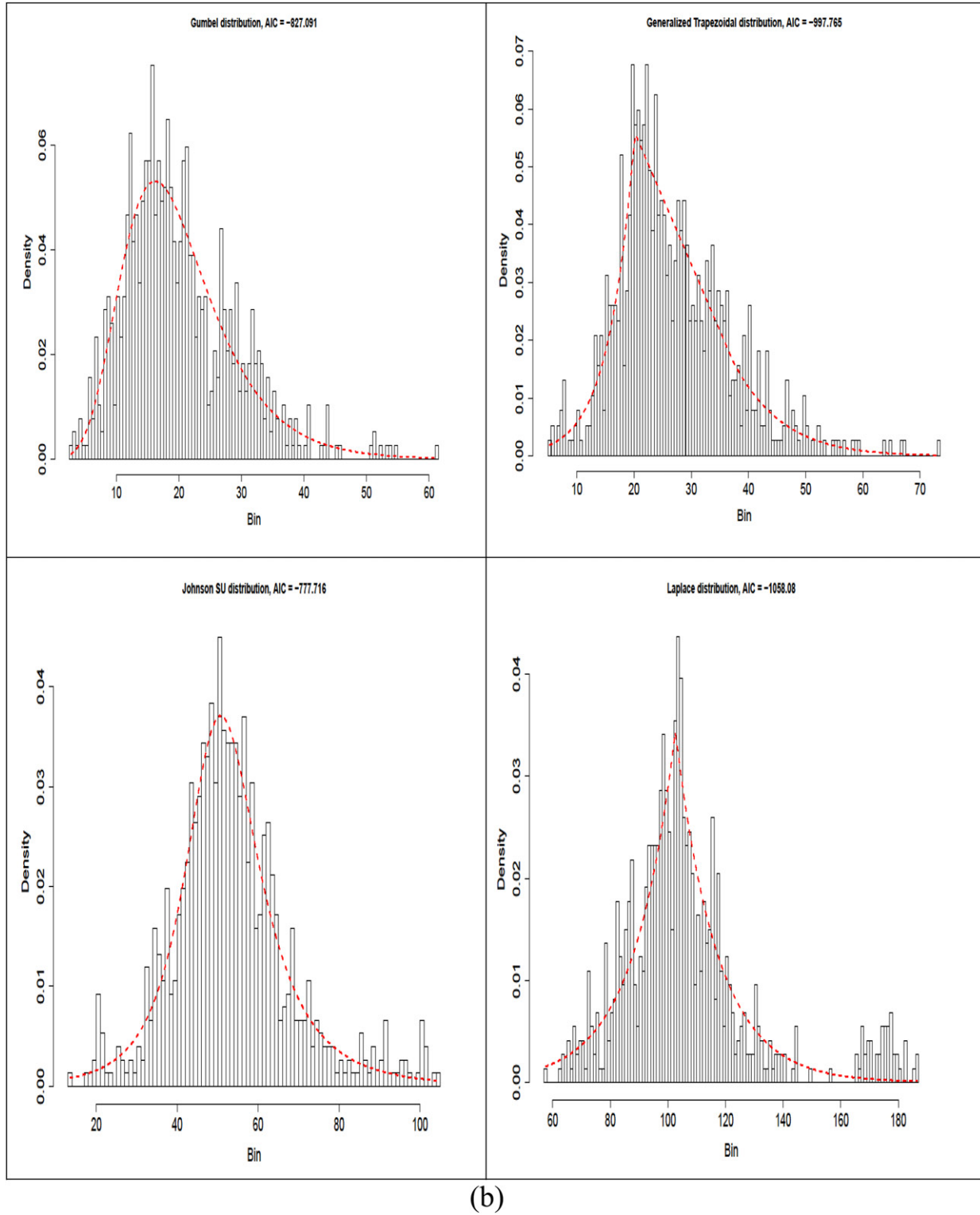


Fig. 2. Continued

where $c_0 = 2.515517$, $c_1 = 0.802853$, $c_2 = 0.010328$, $d_0 = 1.432788$, $d_1 = 0.985269$ and $d_2 = 0.001308$ are constant factors (Abramowitz and Stegun, 1965).

The resulting standardized values can be classified by using the ranges of SPTI values corresponding to their drought categories as given in Table 1.

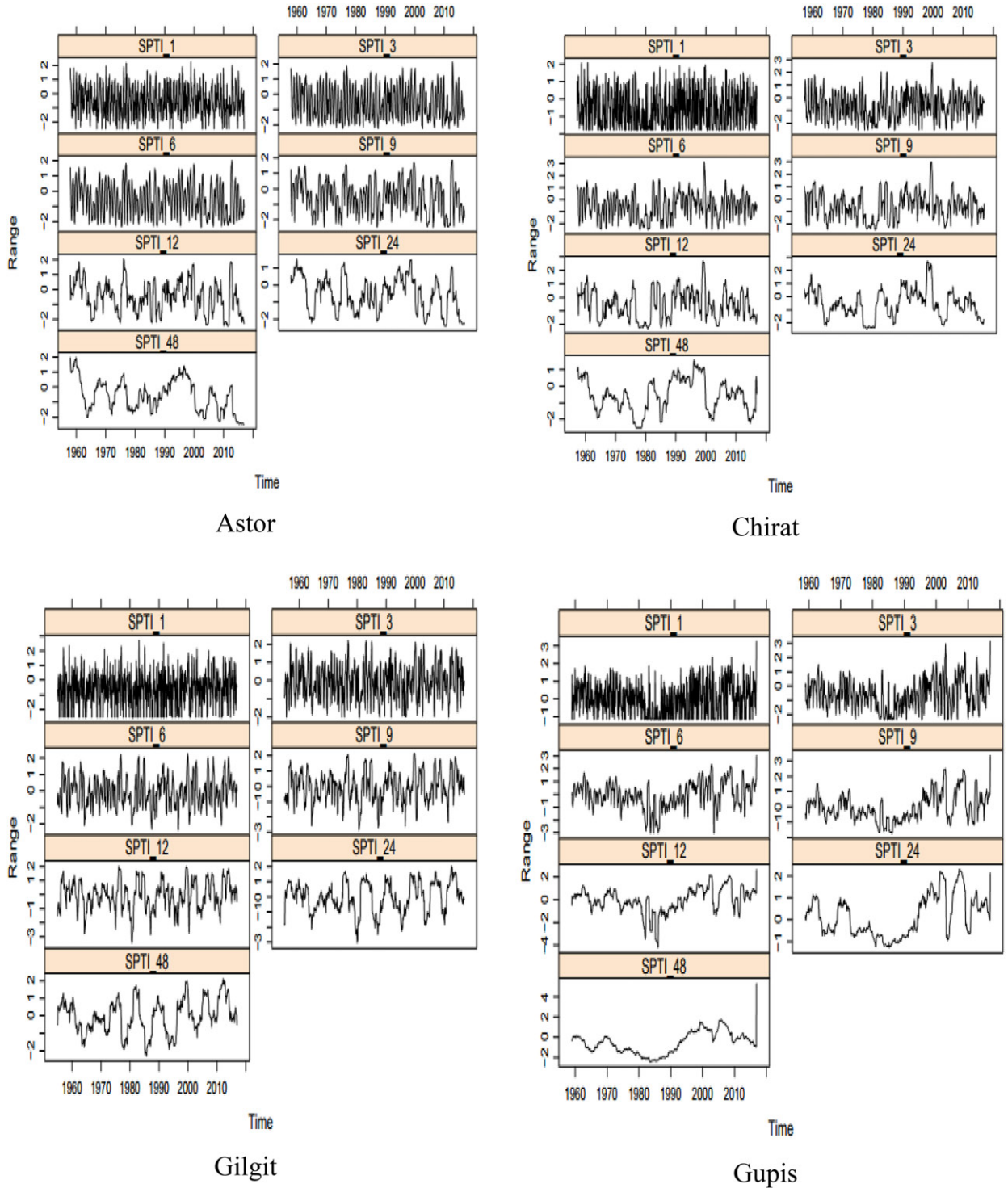


Fig. 3. Temporal plot of SPTI value at various Time-Scales.

Moreover, the estimation of SPTI at various time scales such as 3, 6 and 9 can be made by just taking moving total of DAI_i series at 3, 6, and 9 months, respectively (Vicente-Serrano, 2006).

The selection of appropriate probability distribution and the calculation procedure of multi-scalar drought index-SPTI at 1, 3, 6, 9, 12, 24 and 48 time scale are estimated by following the guidelines provided in Stagge et al. (2015).

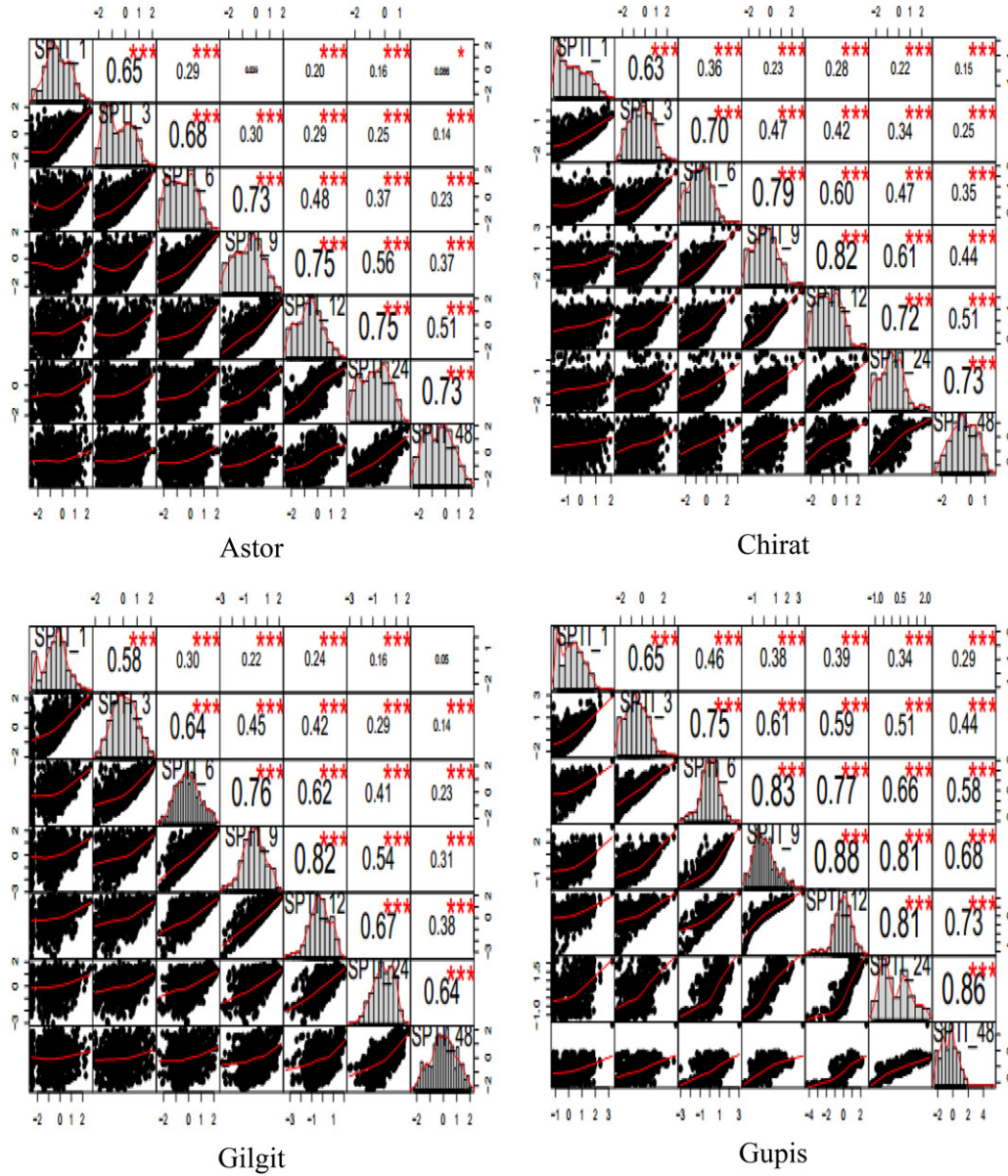


Fig. 4. Correlation among various SPTI Time-Scales.

2.2. Boruta algorithm and relative importance of time scales

To rank and obtain most important, unbiased, and stable time scales of SPTI drought, a wrapper algorithm boruta (Kursa and Rudnicki, 2010) is employed using R statistical software. Boruta uses random forest classifier which offers two main adapters: one is permutation importance (raw permutation adapter and normalized permutation adapter) and the other is Gini importance. Gini importance measures the importance of each feature as the sum over the number of splits that include the feature.

Computational strategies of these measures are done using *randomforest* (Liaw and Wiener, 2002) R package. Additional technical details on the random forest based variable importance measures can be found in Archer and Kimes (2008).

Among several machine learning algorithms, the choice of random forest method depends on its two basic reasons: (1) a random forest is the most popular and prevailing ensemble machine learning algorithm and (2) there is no need of manual tuning of its parameter and calculations. We use *Boruta* R package, however several other R

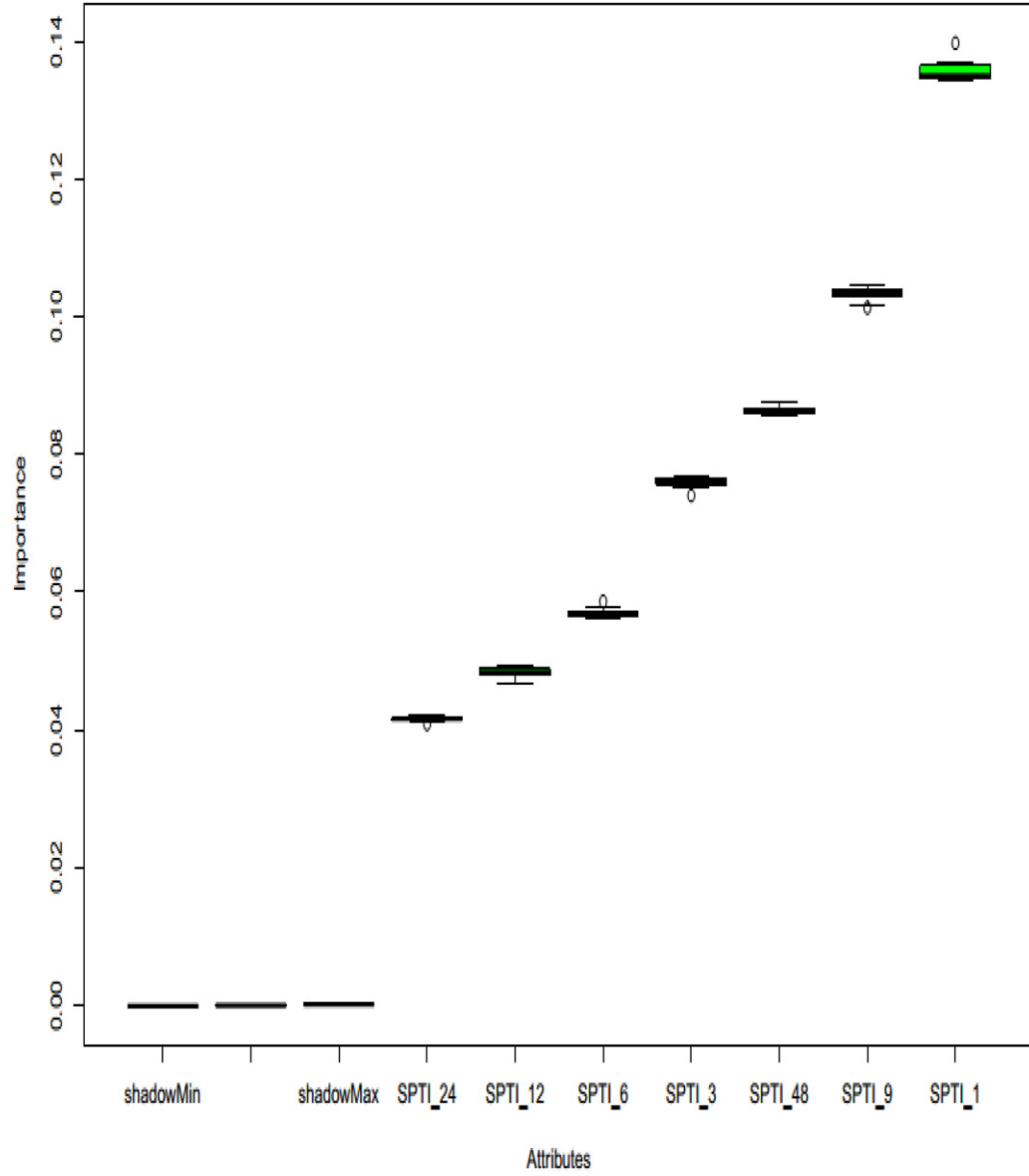


Fig. 5. Boruta using normal permutation adapter.

packages and software are available (Wright and Ziegler, 2015).

Random forests include a combination of three predictors, where each tree produces response vectors when it is used as a predictor. In order to estimate error rate and prediction of data, the random forest classifier produces several classifications trees (Liaw and Wiener, 2002). In Boruta wrapper algorithm with the random classifier, randomness is added to the system and results are obtained from the collective randomized samples by reducing the disingenuous influence of the random variation and correlation. This algorithm calculates weights

associated with each feature by iteratively learning the random forest classifier. In our proposed model of Boruta algorithm, spatial reference points of meteorological stations are used as response class, while temporal data on SPTI at different time scales are considered as the predictor variables.

In this research, the steps involved in Boruta algorithm with the random forest as a classifier for the selection of unbiased and stable scale are as follows (Kursa and Rudnicki, 2010).

- In the first step, for the extension of the range of information system, additional copies of all the

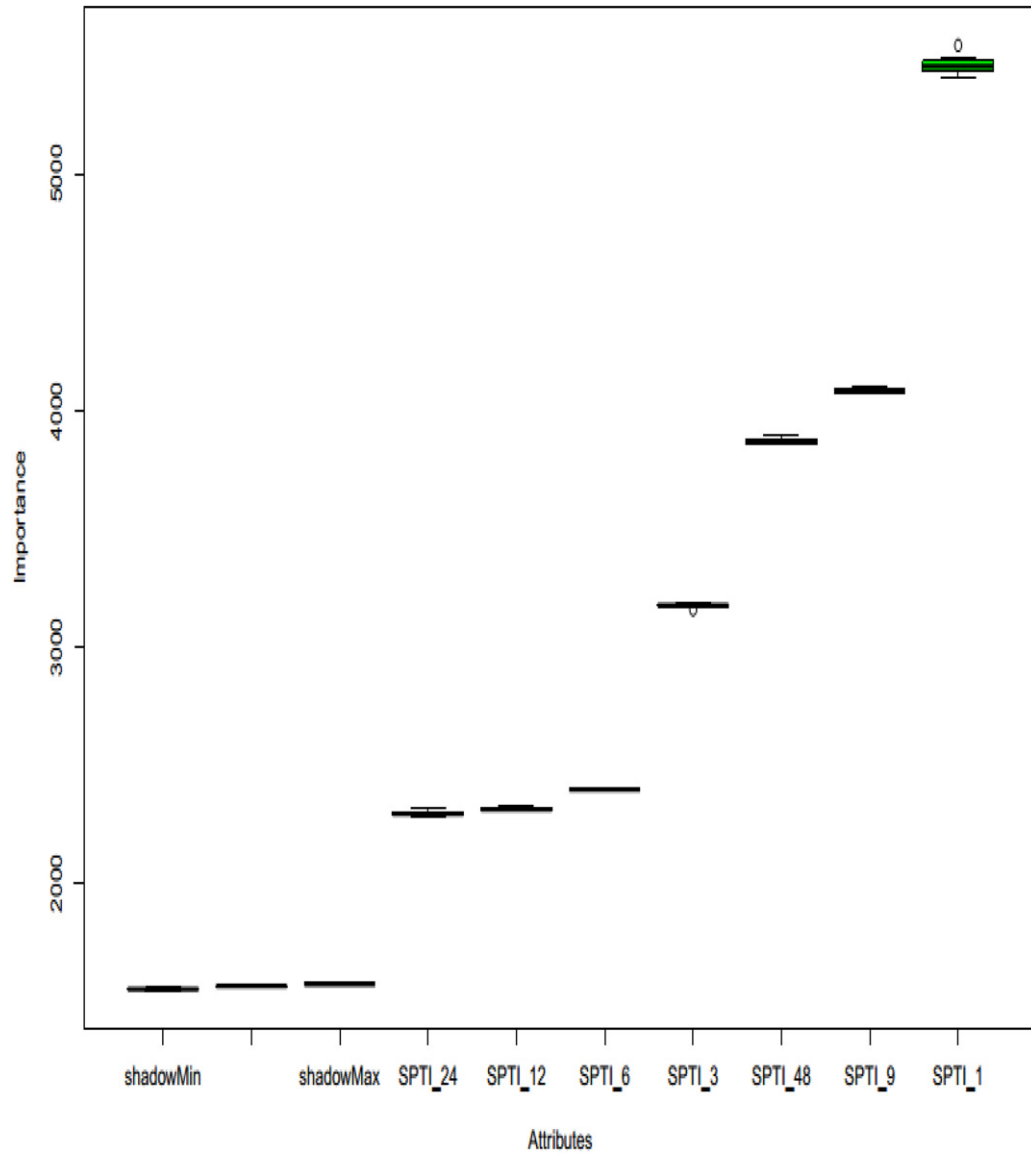


Fig. 6. Boruta using raw permutation adapter.

variables (time series data on each time scales) are added in the system with a minimum of five shadows.

- In the second step, all the added variables are shuffled to eliminate correlation among each time scale from spatially referenced meteorological stations.
- In the third step, Z-score statistics are obtained by implementing random forest classifier on the entire information system. Where Z-score is computed by dividing the average loss by its standard deviation which can further be used as an importance measure.
- In the fourth step, a hit is assigned to each time scale by comparing it with those time scales, which have a

better score than the maximum value of the Z-score among the shadow attributes.

- In the fifth step, a two-sided test is performed for each attribute (variable) with indecisive significance by shadow attributes.
- In the sixth step, permanently removed those original time scales that are significantly unimportant attribute than shadow attribute. In this step, shadow attributes are playing a role as unimportant attributes.
- In the seventh step, attributes which have significant importance over the shadow attributes are then confirmed. In this step, shadow attributes are playing a role as an important attributes.

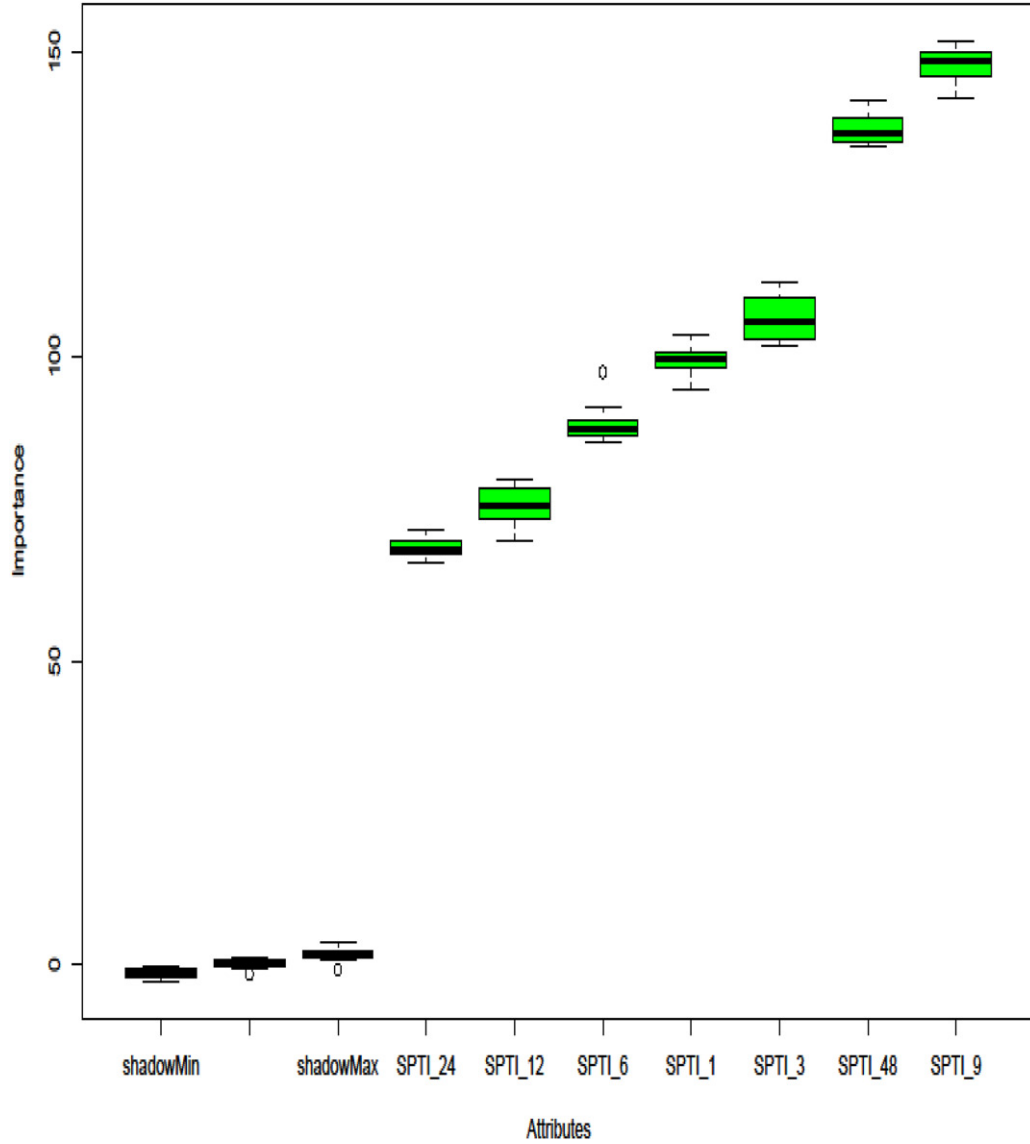


Fig. 7. Boruta using Gini importance measure.

Table 3. Attribute statistics using normal permutation adapter.

Time scale	meanImp	medianImp	minImp	maxImp	normHits	Decision
SPTI-1	0.943056	0.9375	0.931944	0.970833	1	Confirmed
SPTI-3	0.527083	0.528472	0.514583	0.531944	1	Confirmed
SPTI-6	0.396528	0.396528	0.390972	0.407639	1	Confirmed
SPTI-9	0.717361	0.718056	0.703472	0.727083	1	Confirmed
SPTI-12	0.336111	0.336806	0.325	0.342361	1	Confirmed
SPTI-24	0.288889	0.289583	0.284028	0.292361	1	Confirmed
SPTI-48	0.6	0.6	0.595139	0.607639	1	Confirmed

- In the eight step, all the shadow attributes are removed from the system.

Finally, all the process is repeated for large number times until each time scale in the system receives optimal

Table 4. Attribute statistics using raw permutation adapter.

Time scale	meanImp	medianImp	minImp	maxImp	normHits	Decision
SPTI-1	5462.07	5455.70	5410.07	5545.37	1	Confirmed
SPTI-3	3174.62	3174.60	3157.79	3189.49	1	Confirmed
SPTI-6	2396.07	2395.72	2392.16	2400.24	1	Confirmed
SPTI-9	4081.73	4083.40	4068.50	4096.21	1	Confirmed
SPTI-12	2311.93	2312.25	2304.67	2322.54	1	Confirmed
SPTI-24	2295.08	2293.00	2282.34	2313.87	1	Confirmed
SPTI-48	3868.35	3865.60	3853.09	3890.85	1	Confirmed

Table 5. Attribute statistics using Gini importance measure.

Time scale	meanImp	medianImp	minImp	maxImp	normHits	Decision
SPTI-1	99.40	99.54	94.67	103.46	1	Confirmed
SPTI-3	106.19	105.55	101.87	112.21	1	Confirmed
SPTI-6	88.86	88.02	85.71	97.41	1	Confirmed
SPTI-9	147.96	148.41	142.56	151.70	1	Confirmed
SPTI-12	75.71	75.54	69.81	73.64	1	Confirmed
SPTI-24	68.43	68.30	65.94	71.52	1	Confirmed
SPTI-48	137.37	136.75	134.54	142.15	1	Confirmed

weights. Hence, the time scales which are statistically significant (i.e. have the higher value of Z-score than shadow attributes) are selected as the most important and unbiased time scale. However, those time scales having a small value of Z-score than shadow attributes deemed as irrelevant, and are removed from the system.

Before performing Boruta algorithm to rank and search important time scale for socio-geography of Pakistan, we first infer the behaviour of SPTI as a moving average process. The Pearson correlation is used to observe the association among time scales for case study catchment stations.

3. Case study

3.1. Data description and study area

We use time series data of monthly total precipitation and monthly mean temperature of 52 meteorological stations that are located across Pakistan. Pakistan is located in South Asia that is the junction of Central Asia and Middle East, which gives its location great significance. According to the census 2016, its total population is 207,774,520. Where the majority of people belongs to agriculture sectors either directly or indirectly (Rehman et al., 2015). Its varying regional temperature and precipitation amount in various season indulge the traditional hazard monitoring structure and, require some complex indicators and research frameworks to make exhilarated and effective drought mitigation policies. From past few years, the country suffered severe bad effects which

include loss of properties, agricultural lands, prevalence of disease and human deaths, in its historical hazards profile (Spate and Learmonth, 2017).

In this study, we explore and infer the historical profile of extreme drought events at 52 meteorological stations covering triple of low, moderate and high concentrated climatic variables. Figure 1 shows the selected meteorological stations, and further details about selected meteorological stations can be found in <http://www.pmd.gov.pk/Observatories/index.html>.

We use time series data of monthly total precipitation and monthly mean temperature to compute SPTI at varying time scales. The dataset is collected from Pakistan meteorological department through Karachi data processing centre.

3.2. Calculation of SPTI data

In this study, monthly time series secondary data total precipitation and average temperature from 1955 to 2016 is used to compute SPTI at various time scales by standardizing DAI_i series. We use Kolmogorov-Smirnov, and Chi-Square, Anderson-Darling tests to find the candidate distribution for each time scale with level of significance $\alpha = .05$ by using Easyfit software (Schittkowski, 2002). Distributions with minimum value of AIC are then selected for standardization. Further, as different stations have different climatic features, and in each station, the definition of drought with respect to time scale is varied. For example, in moving average process, the definition of meteorological drought at 1- and 3-month time scales,

may not lead to monitoring agricultural drought at 6- and 12-month time scale in tropical regions (Dracup et al., 1980).

Table 2 shows the AIC values of candidate distributions that are fitted on DAI time series at 1, 3, 6, 9, 12, 24 and 48 time scales of Chirat station (Latitude: 33°49.4052' N, Longitude: 71°53.5752'). At 1-month time scale, 3P Weibull with minimum AIC value -508.52 among all the list of probability functions, have strong candidacy of its standardization. However, generalized extreme values distribution is the best fitted for 3-month time scale, with minimum value of AIC -542.36. Figure 2a and b shows the histograms of those distributions that having minimum values of AIC at all time scale. It is observed that, as the time scale increase, behaviour of data tends to inflate in both sides of tails.

Figure 2b clearly indicates that the Laplace distribution at 48-month time scale doesn't capture the uncertainty on the significant parts of the data. Moreover, substantial deviation can be seen in extreme right part of the plot. Similarly, Genralized Extreme Values (GEV) at 3-month time scale, Gumbel at 9-month time scale, Johnson at 24-month time scale and Generalized Trapezoidal distribution at both 6- and 12-month time scale are not suitable for drawing accurate inferences. Consequently, the process of selecting the more relevant and appropriate distribution for each time scale may create barrier in fluent drought monitoring system.

Hence, the selection of appropriate time scale for the interpolation of the whole region is the most important task for water management and resources. Moreover, the selection of appropriate drought scale for monitoring drought as compared to the selection of drought indices is more important (McKee et al., 1993; Vicente-Serrano et al., 2010). In the next subsection, temporal behaviour of each time scale is assessed using temporal plot and correlation analysis.

3.3. Assesment of temporal behaviour

Figure 3 shows the temporal behaviour of each time scale for Astor, Chirat, Gilgit and Gupis stations. These stations are randomly selected for pictorization of case study computations and analysis. However, similar temporal behaviour of each SPTI time scale was observed for all other stations. On the same line, Fig. 4 presents the correlation and scatters behaviour among each time scale. By careful inspection, it is inferred that as time scale increases, the correlation between indices increases, however, the correlation behaviour decreases from higher to lower time scale. For example, in Chirat station, the correlation between SPTI-48 and SPTI-24 is 0.73, while

SPTI-48 and SPTI-1 has very small value 0.15 of the correlation.

3.4. Selection of time scales

Application of the boruta algorithm with two random forest adapters are applied to confirm and rank the significance of important time scale. The most commonly used time scales SPTI-1, SPTI-3, SPTI-6, SPTI-9, SPTI-12, SPTI-24, and SPTI-48 and spatially reference meteorological stations are used as predictor variables and response class respectively. Results associated with this research show that all time scales are important. However, the SPTI-1 time scale is observed to be the most important by normal permutation and raw permutation adaptor. Figures 5 and 6 show the significant importance of SPTI-1 as compared to other time scales for classification of drought in the selected stations. Hence, SPTI-1 has their own identity and can be used for drought monitoring for spatially referenced geographical studies. Moreover, in both figures the order of relevance, importance of time scales is the same. However, SPTI-9 ranked second by the Gini Index (see Fig. 7). Although the Gini index is a biased method, nevertheless, this method also indicates that all time scales are important.

Therefore, after SPTI-1, it is observed that SPTI-9 in both random forest adapters have a tentative candidate for drought monitoring for this region (see Fig. 7). Higher values of attribute statistics of SPTI-1 confirm its significant importance as compared to other time scales (see Tables 3 and 4). In Table 3, the mean importance of SPTI-1 using random forest raw importance is 0.1558 while median and mode importance is 0.1342 and 0.1398. However, the difference between the minimum and maximum importance is almost negligible.

In Table 4, details about attribute statistics of all scales are described where mean importance of SPTI-1 is significantly higher as compared to other time scales of SPTI. However, Gini index rank 4th to SPTI-1 index (see Fig. 7, Table 5). But, all the time scales are important, and can be used to monitor drought for specific drought type. That is, no single time scale was found unimportant.

4. Discussion

This paper presents the solution of issues regarding interpretation and uses of various time scales of multi-scalar drought indices. To overcome multi-scaling issues raised by Bazrafshan et al. (2014), this research provides an application of Boruta algorithm for searching and ranking the relevant time scale of the SPTI drought index.

In case study, SPTI at 1, 3, 6, 9, 12, 24 and 48 time scales of 52 maeterological station of Pakistan were

computed by standardizing appropriate probability distribution function. To rank and find important time scale of SPTI, Boruta algorithm of machine learning were used to select important time scale. Results of this feature selection approach confirms the significance of all drought time scale. Moreover, experimental results show that SPTI-1, being high ranked and having maximum value of mean importance, determine the exact and joint effect of drought by extracting features from other scales. Two famous and mostly used random forest adapter, suggest the importance of SPTI-1 for spatiotemporal classification of droughts. As different stations have different climatic features, and in each station definition of drought with respect to drought scale are varried. For example, in moving average process, the definition of meteorological drought at 1- and 3-month time scale, may not lead to monitoring agricultural drought at 6- and 12-month time scale in tropical regions (Dracup et al., 1980). So, the selection of unbiased and stable scale for the interpolation of the whole region is the most important task for water management and resources.

However, for single station monitoring, one can use various time scales depending on the nature of the study. Therefore, instead of the calculation of all time scales for all stations, SPTI-1 is suggested as it is the most important time scale for spatiotemporal analysis. There is no need for the calculation of time series data on all the time scales to infer the regionalized behaviour of drought severity. So, unlike MSPI, one doesn't need to calculate all the other time scales of specified regions. For operational point of view, as each drought time scale has its own importance, so instead of using only SPTI-1, one can use other time scales for a specific purpose. However, for advanced spatial and temporal methodologies, unbiased and stable time scales may contribute in reliable and accurate spatiotemporal interpolations. Hence, for regionalized study and spatiotemporal interpolation of drought, machine learning approach may lead to rank and select the most important time scales of multi-scalar drought indices.

5. Conclusions

Current issues in multi-scalar drought indices create several ambiguous problems regarding their computation and interpretation. The purpose of this study is to search more appropriate and unbiased time scale which can be used as a geographical representative, and hence to resolve the issues for accurate and efficient drought monitoring mitigation policies for a particular geography. We present the research framework of how the feature selection wrapper algorithm of Boruta can incorporate for optimal choice of drought time scale in a spatial context.

Preliminary, the study used spatio-temporal secondary data on monthly cumulative precipitation and mean temperature. Calculation procedure of multi-scalar drought index-SPTI at 1, 3, 6, 9, 12, 24 and 48 time scale are according the guidelines provided by Stagge et al. (2015). Exploratory analysis indicates that as the time scale increase, the correlation between successors and predecessors increase. Therefore, to avoid computational labour and interpretation difficulties of using multiple time scale, appropriate technique of feature selection may be incorporated. In this research, spatially referenced meteorological stations are used as the response variable, whilst all the time scales are playing their role as predictors in Boruta algorithm of supervised machine learning environment. Experimental results show that all the time scales are important in terms of classification. Here, SPTI-1 is the topped ranked and have the ability to capture the spatially referenced variation from other time scales. Similarly, the research may expand for other geographical regions and drought indices. Hence, both the time reduction and multi-scaling interpretation problem can be resolved by using selected drought time scale for inferencing spatially referenced meteorological stations.

Ethical statement, financial disclosure, and conflict of interest

The manuscript is prepared in accordance with the ethical standards of the responsible committee on human experimentation and with the latest (2008) version of Helsinki Declaration of 1975. The manuscript is prepared by using secondary data and authors have not received any financial support. The authors of manuscript certify that they have no affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript. The authors declare that they have no competing interests.

Acknowledgement

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group no RG-1439-015.

References

- Abramowitz, M. and Stegun, I. A. 1965. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, Vol. 55. New York, NY: Courier Corporation.
- Adnan, S., Ullah, K., Shuanglin, L., Gao, S., Khan, A. H. and co-authors. 2018. Comparison of various drought indices to

- monitor drought status in Pakistan. *Clim. Dyn.* **51**(5–6), 1885–1899.
- Ali, Z., Hussain, I., Faisal, M., Nazir, H. M., Abd-el Moemen, M. and co-authors. 2017. A novel multi-scalar drought index for monitoring drought: the standardized precipitation temperature index. *Water Resour. Manage.*, **31**(15), 4957–4969.
- Archer, K. J. and Kimes, R. V. 2008. Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.* **52**, 2249–2260. doi:10.1016/j.csda.2007.08.015
- Aswathanarayana, U. 2001. *Water Resources Management and the Environment*. Rotterdam, Netherlands: CRC Press.
- Bazrafshan, J., Hejabi, S. and Rahimi, J. 2014. Drought monitoring using the multivariate standardized precipitation index (MSPI). *Water Resour. Manage.* **28**, 1045–1060. doi:10.1007/s11269-014-0533-2
- Blain, G. C. and Meschiatti, M. C. 2015. Inadequacy of the gamma distribution to calculate the Standardized Precipitation Index. *Rev. Bras. Eng. Agríc. Ambient.* **19**, 1129–1135. doi:10.1590/1807-1929/agriambi.v19n12p1129-1135
- Dawadi, S. and Ahmad, S. 2013. Evaluating the impact of demand-side management on water resources under changing climatic conditions and increasing population. *J. Environ. Manage.* **114**, 261–275. doi:10.1016/j.jenvman.2012.10.015
- Dogan, S., Berktaş, A. and Singh, V. P. 2012. Comparison of multi-monthly rainfall-based drought severity indices, with application to semi-arid konya closed basin, Turkey. *J. Hydrol.* **470**, 255–268.
- Dracup, J. A., Lee, K. S. and Paulson, E. G. 1980. On the definition of droughts. *Water Resour. Res.* **16**, 297–302. doi:10.1029/WR016i002p00297
- Cunnane, C. 1978. Unbiased plotting positions—a review. *J. Hydrol.* **37**, 205–222. doi:10.1016/0022-1694(78)90017-3
- Farahmand, A. and AghaKouchak, A. 2015. A generalized framework for deriving nonparametric standardized drought indicators. *Adv. Water Resour.* **76**, 140–145. doi:10.1016/j.advwatres.2014.11.012
- Ghamghami, M., Hejabi, S., Rahimi, J., Bazrafshan, J. and Olya, H. 2017. Modeling a drought index using a nonparametric approach. *Global Nest J.* **19**, 58–68.
- Gringorten, I. I. 1963. A plotting rule for extreme probability paper. *J. Geophys. Res.* **68**, 813–814. doi:10.1029/JZ068i003p00813
- Guttman, N. B. 1998. Comparing the palmer drought index and the standardized precipitation index. *J. Am. Water Resources Assoc.* **34**, 113–121. doi:10.1111/j.1752-1688.1998.tb05964.x
- Fiorillo, F. and Guadagno, F. M. 2010. Karst spring discharges analysis in relation to drought periods, using the SPI. *Water Resour. Manage.* **24**, 1867–1884. doi:10.1007/s11269-009-9528-9
- Hao, Z. and AghaKouchak, A. 2013. Multivariate standardized drought index: a parametric multi-index model. *Adv. Water Resour.* **57**, 12–18. doi:10.1016/j.advwatres.2013.03.009
- Hao, Z. and Singh, V. P. 2015. Drought characterization from a multivariate perspective: a review. *J. Hydrol.* **527**, 668–678. doi:10.1016/j.jhydrol.2015.05.031
- Hao, Z., Yuan, X., Xia, Y., Hao, F. and Singh, V. P. 2017. An overview of drought monitoring and prediction systems at regional and global scales. *Bull. Amer. Meteor. Soc.* **98**, 1879–1896. doi:10.1175/BAMS-D-15-00149.1
- Hayes, M. J., Wilhelmi, O. V. and Knutson, C. L. 2004. Reducing drought risk: bridging theory and practice. *Nat. Hazards Rev.* **5**, 106–113. doi:10.1061/(ASCE)1527-6988(2004)5:2(106)
- Keyantash, J. and Dracup, J. A. 2002. The quantification of drought: An evaluation of drought indices. *Bull. Am. Meteorol. Soc.* **83**, 1167–1180.
- Kogan, F. N. 2000. Contribution of remote sensing to drought early warning. In: *Early Warning Systems for Drought Preparedness and Drought Management*. World Meteorological Organization, Geneva, 75–87.
- Kursa, M. B. and Rudnicki, W. R. 2010. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13.
- Liaw, A., and Wiener, M. 2002. Classification and regression by random forest. *R. News* **2**, 18–22.
- McKee, T. B., Doesken, N. J., Kleist, J. 1993. The relationship of drought frequency and duration to time scales. In: *Proceedings of the 8th Conference on Applied Climatology*. American Meteorological Society, Boston, MA, USA, vol. 17, pp. 179–183.
- Mishra, A. K. and Desai, V. R. 2005. Drought forecasting using stochastic models. *Stoch. Environ. Res. Ris. Assess.* **19**, 326–339. doi:10.1007/s00477-005-0238-4
- Musuuza, J. L., Van Loon, A. F. and Teuling, A. J. 2016. Multiscale evaluation of the Standardized Precipitation Index as a groundwater drought indicator. *Hydrol. Earth Syst. Sci.* **20**, 1117.
- Ntale, H. K. and Gan, T. Y. 2003. Drought indices and their application to East Africa. *Int. J. Climatol.* **23**, 1335–1357. doi:10.1002/joc.931
- Patel, N. R., Chopra, P. and Dadhwal, V. K. 2007. Analyzing spatial patterns of meteorological drought using standardized precipitation index. *Met. Apps.* **14**, 329–336. doi:10.1002/met.33
- Parker, W. 2014. Values and uncertainties in climate prediction, revisited. *Stud. Hist. Philos. Sci.* **46**, 24–30. doi:10.1016/j.shpsa.2013.11.003
- Rehman, A., Jingdong, L., Shahzad, B., Chandio, A. A., Hussain, I. and co-authors. 2015. Economic perspectives of major field crops of Pakistan: an empirical study. *Pacif. Sci. Rev. B Hum. Soc. Sci.* **1**, 145.
- Schittkowski, K. 2002. Easy-fit: a software system for data fitting in dynamical systems. *Struct. Multidisc. Optim.* **23**, 153–169. doi:10.1007/s00158-002-0174-6
- Schneider, S. H., Root, T. L. and Mastrandrea, M. D. 2011. *Encyclopedia of Climate and Weather*, Vol. 1. Oxford University Press, New York.
- Spate, O. H. K. and Learmonth, A. T. A. 2017. *India and Pakistan: A General and Regional Geography*, Vol. 12. Routledge, UK.
- Sheffield, J., Wood, E. F. and Roderick, M. L. 2012. Little change in global drought over the past 60 years. *Nature* **491**, 435–438. doi:10.1038/nature11575
- Spieß, A. N. 2014. Propagate: Propagation of Uncertainty. R package version 1.0-4. Online at: <https://CRAN.R-project.org/package=propagate>.

- Stagge, J. H., Tallaksen, L. M., Gudmundsson, L., Van Loon, A. F., and Stahl, K. 2015. Candidate distributions for climatological drought indices (SPI and SPEI). *Int. J. Climatol.* **35**, 4027–4040. doi:[10.1002/joc.4267](https://doi.org/10.1002/joc.4267)
- Svoboda, M. and Fuchs, B. and of Nebraska-Lincoln. National Drought Mitigation Center, U. 2016. *Handbook of Drought Indicators and Indices*. Integrated Drought Management Tools and Guidelines Series. World Meteorological Organization, Geneva, Switzerland.
- Tigkas, D., Vangelis, H. and Tsakiris, G. 2016. Introducing a modified reconnaissance drought index (RDIE) incorporating effective precipitation. *Proc. Eng.* **162**, 332–339. doi:[10.1016/j.proeng.2016.11.072](https://doi.org/10.1016/j.proeng.2016.11.072)
- Tsakiris, G., Nalbantis, I., Vangelis, H., Verbeiren, B., Huysmans, M. and co-authors. 2013. A system-based paradigm of drought analysis for operational management. *Water Resour. Manage.* **27**, 5281–5297. doi:[10.1007/s11269-013-0471-4](https://doi.org/10.1007/s11269-013-0471-4)
- van Dorp, J. R. and Kotz, S. 2003. Generalized trapezoidal distributions. *Metrika* **58**, 85–97. doi:[10.1007/s001840200230](https://doi.org/10.1007/s001840200230)
- Vicente-Serrano, S. M. 2006. Differences in spatial patterns of drought on different time scales: an analysis of the Iberian Peninsula. *Water Resour. Manage.* **20**, 37–60. doi:[10.1007/s11269-006-2974-8](https://doi.org/10.1007/s11269-006-2974-8)
- Vicente-Serrano, S. M., Beguería, S. and López-Moreno, J. I. 2010. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *J. Climate* **23**, 1696–1718. doi:[10.1175/2009JCLI2909.1](https://doi.org/10.1175/2009JCLI2909.1)
- Vicente-Serrano, S. M., García-Herrera, R., Barriopedro, D., Azorin-Molina, C., López-Moreno, J. I., Martín-Hernández, N., ... & Nieto, R. 2016. The Westerly Index as complementary indicator of the North Atlantic oscillation in explaining drought variability across Europe. *Climate dynamics* **47**, 845–863. doi:[10.1007/s00382-015-2875-8](https://doi.org/10.1007/s00382-015-2875-8)
- Wang, H., Pan, Y. and Chen, Y. 2017. Comparison of three drought indices and their evolutionary characteristics in the arid region of northwestern China. *Atmos. Sci. Lett.* **18**, 132–139. doi:[10.1002/asl.735](https://doi.org/10.1002/asl.735)
- Wilhite, D. A. 1994. *Preparing for Drought: A Guidebook for Developing Countries*. Diane Publishing, Darby, PA.
- Wilhite, D. A. 2000. Drought as a natural hazard: concepts and definitions. *Drought, a Global Assessment* **1**, 3–18.
- Wright, M. N. and Ziegler, A. 2015. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.