Hindawi
Advances in Meteorology
Volume 2020, Article ID 5014280, 13 pages
https://doi.org/10.1155/2020/5014280

Research Article

A Novel Framework for Selecting Informative Meteorological Stations Using Monte Carlo Feature Selection (MCFS) Algorithm

Rizwan Niaz,[1] Ibrahim M. Almanjahie,[2,3] Zulfiqar Ali,[1] Muhammad Faisal,[4,5] and Ijaz Hussain[1]

[1]*Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan*
[2]*Statistical Research and Studies Support Unit, King Khalid University, Abha, Saudi Arabia*
[3]*Mathematics, College of Science, King Khalid University, Abha, Saudi Arabia*
[4]*Faculty of Health Studies, University of Bradford, Bradford, UK*
[5]*Bradford Institute for Health Research, Bradford, UK*

Correspondence should be addressed to Zulfiqar Ali; zulfiqarali@stat.qau.edu.pk and Ijaz Hussain; ijaz@qau.edu.pk

Received 27 September 2019; Accepted 12 December 2019; Published 17 February 2020

Academic Editor: Giacomo Gerosa

Spatial distribution of meteorological stations has a significant role in hydrological research. The meteorological data play a significant role in drought monitoring; in this regard, accurate and suitable provision of meteorological stations is becoming crucial to improve and strengthen the skill of drought prediction. In this perspective, the choice of meteorological stations in a specific region has substantial importance for accurate estimation and continuous monitoring of drought hazards at the regional level. However, installation and data mining on a large number of meteorological stations require high cost and resources. Therefore, it is necessary to rank and find dependencies among existing meteorological stations in a particular region for further climatological analysis and reanalysis of databases. In this paper, the Monte Carlo feature selection and interdependency discovery (MCFS-ID) algorithm-based framework is proposed to identify the important meteorological station in a particular region. We applied the proposed framework on 12 meteorological stations situated in varying climatological regions of Punjab (Pakistan). We employed the drought index SPTI on 1-, 3-, 6-, 9-, 12-, 24-, and 48-month time-scale data to find the interdependencies among meteorological stations at various locations. We found that Sialkot has significance regional importance for studying SPTI-3, SPTI-6, and SPTI-48 indices. This regional importance is based on scores of relative importance (RI); for example, the RI values for SPTI-3, SPTI-6, and SPTI-48 indices are 0.1570, 0.1080, and 0.0270, respectively. Furthermore, the Jhelum station has more relative importance (RI = 0.1410 and 0.1030) for SPTI-1 and SPTI-9 indices, while varying concentration behaviour is observed in the remaining time scales.

1. Introduction

Drought is a creeping phenomenon and recurrently occurring natural disaster in many regions of the world [1–3]. It is an insidious, slow-moving natural hazard that has an adverse social and environmental impact and can influence the economy of any country [4]. These effects can be well experienced outside the affected area, even at the global level. The complexity of the effects is largely due to the dependence of so many areas on water to provide goods and services such as availability and quality of water which have serious implications for water resource management [5, 6]. Since rising pressure on water and other natural resources leads to drought [7], it is clear that producing more comprehensive assessments over time is challenging [8].

Due to the complex features of drought, it has severe and prolonged adverse effects [9]. Attempts have been made to identify the complexity of these effects at the local, regional, or national level [10]. It is almost impossible to track the databases and trends of the region because of the insidious behaviour of drought in a region. The researchers and policymakers have given different strategies for their

countries to improve the level of preparation for their drought by building better early warning systems and adopting drought policies and response and mitigation plans [11, 12]. It is therefore imperative for scientists and policymakers to develop efficient early warning drought-monitoring tools to avoid the severe effect of drought [13–15]. However, long-term records of drought indices of the regionally representative meteorological station are required for accurate estimation of regional drought forecasting and for developing an early warning tool. In this regard, accurate estimation and continuous monitoring of future drought at the regional level require a dense meteorological network. However, the implications of each meteorological network require high cost and resources. Particularly in developing countries, the high cost of installation and complex sampling design may force to adopt compromise allocation and installation of meteorological stations [16–18].

In the last decades, several algorithms and methods for the optimal selection of metrological stations have been developed [19, 20]. These algorithms and methods reduce the size of the network that provides more accurate and regionally representative estimates of metrological variables. In this perspective, resultant spatial distributions of optimal meteorological stations play a key role, specifically in hydrological research [21, 22]. Therefore, the essential feature of hydrologic research is to achieve optimal meteorological stations based on some meteorological characteristics [20, 23]. To achieve the optimal network, geographical facts such as deserts, hills, and forests of a region are the key factors that significantly contribute to the distribution of meteorological stations. Consequently, optimal selection of meteorological networks requires a standardized climatic indicator.

There are several tools to monitor drought and its characterization [24]. However, the estimation of standardized drought indices (SDIs) and other novel approaches of drought indices on all meteorological stations make a chaotic situation for regional forecasting and early warning management policies. The use of the SDI has been found in many applications [25–27]. SDIs are useful for drought characterization and comparing meteorological stations having different climatological characteristics [28]. Nevertheless, long-term SDI time-series data are crucial for drought characterization at individual meteorological stations [29]. It is thus necessary to rank and find dependencies among existing meteorological stations for climatological and reanalysis databases. In this regard, advanced statistical procedures are helpful to find important and regionally dependent stations under complex meteorological network settings.

In this study, we propose a framework to identify important meteorological stations and to discover dependencies among stations for up-to-date real-time drought-monitoring systems. The core configuration of the framework is based on MCFS-ID [30, 31]. We applied this proposed framework on 12 meteorological stations situated in varying climatological regions of Punjab (Pakistan). The analysis of the study is performed with the drought index SPTI on 1-, 3-, 6-, 9-, 12-, 24-, and 48-month time-scale data to find the interdependencies among meteorological stations at varied stations.

2. Methods

2.1. Standardization Precipitation Temperature Index (SPTI). There are numerous procedures that use the multiscalar drought index to describe the severity of the drought. A drought index is called the Standardized Precipitation Index (SPI), which is based on the overlong time period precipitation records to compute the precipitation scarcity [32]. The SPI can be monitored on different time-scale drought. The SPI is standardized by the suitable probability distributions for the observed monthly cumulative precipitation time series to estimate the quantitative values. Thus, positive and negative SPI values are used to identify greater than and less than median precipitation, respectively. Another water balance model, Standardized Precipitation Evapotranspiration Index (SPEI), where the same mathematical procedure of the SPI is used, is grounded on the basis of a difference between precipitation and potential evapotranspiration (PET) [33]. One significant advantage of the SPEI over the SPI is that it comprises the influence of the evaporation to designate the area being studied. In line with the same methodology of the SPI and SPEI, more recently, a multiscalar drought index SPTI is developed for the characterization of drought in both cold (minimum temperature −5.50) and hot (maximum temperature 45.2) climate regions [34]. In this study, we used the SPTI due to the following three reasons: (1) the regions being selected have both cold and hot climatic weather, (2) the SPTI provides true values for regions observed with low temperature, and (3) there is no mathematical contention in the SPTI mechanism. The procedure for SPTI estimation is as follows: In step one, for each selected station, a De Martonne Aridity Index (DAI) is evaluated by utilizing monthly total precipitation and average temperature as follows:

$$\mathrm{DAI}_i = \frac{P_i}{10 + T_i}, \tag{1}$$

where P_i denotes the monthly total precipitation and T_i stands for the monthly mean temperature. In the second step, the candidacy of appropiriate probability distribution will be considered for DAI_i series of each station. In this work, more specifically, 32 most frequently used probability distributions were applied to perceive the most suitable probability distribution using the *propagate* R package [35]. In step three, distributions are selected for each station's time-series data of DAI_i on the basis of minimum values of the Akaike information criterion (AIC) and Bayesian information criterion (BIC). For each station, we standardize the cumulative distribution function (CDF) of the fitted distribution as follows:

$$H(x) = q + (1 - q)F(x). \tag{2}$$

To adjust the effect of undefined values in the DAI, a little amendment in the CDF is constructed in equation (2). For example, in case of gamma distribution, q is the probability with a value zero for each station in the DAI time-series data. If m represents the ciphers (zero) present in DAI_i time-series

data, then q is likely to be estimated by m/n, where n shows all observations in the DAI_i time series. More specifically, here, other familiar methods of probability plotting position such as that in [36] can be implemented for regulating the probability of undefined values in the CDF instead of the above specified method.

$$\mathrm{SPTI} = -\left(P + \frac{C_0 + C_1 P + C_2 P^2}{1 + d_1 P + d_2 P^2 + d_3 P^3}\right), \quad (3)$$

where

$$P = \sqrt{\ln\left[\frac{1}{\{H(x)\}^2}\right]}, \quad (4)$$

in which

$$0 \le H(x) \le 0.5.$$

$$\mathrm{SPTI} = \left(P - \frac{C_0 + C_1 P + C_2 P^2}{1 + d_1 P + d_2 P^2 + d_3 P^3}\right), \quad (5)$$

where

$$P = \sqrt{\ln\left[\frac{1}{\{1 - H(x)\}^2}\right]}, \quad (6)$$

in which

$$0.5 \le H(x) \le 1. \quad (7)$$

Here, $C_0 = 2.515517$, $C_1 = 0.802853$, $C_2 = 0.010328$, $d_1 = 1.432788$, $d_2 = 0.189269$, and $d_3 = 0.001308$.

2.2. Monte Carlo Feature Selection and Interdependency Discovery (MCFS-ID). In the past two decades, substantial progress has been attained in the zone of feature ranking and selection for high-dimensional classification. Draminski et al. [30] gave efficacious and well-presented features' ranking methods with respect to their classification importance. . Recently, a Bayesian technique of determining automatic relevance is developed with nonfilter approaches (see [37]). Apart from this, the importance of the so-called variable (i.e., feature) can be inferred using random forests [38]. Determination of the importance of variables is not compulsory for the construction of random forests, but it is a subroutine to be made corresponding to the construction of the forest [39, 40]. Features' ranking by variable significance can thus be considered a by-product of the classifier [41]. In this approach, we rely heavily on using the classifier, and we shall not use it for the classification work per se. In fact, we only use classes: (i) according to the characteristics of rank, we differentiate between their classical powers according to their importance and (ii) we find interdependencies between features.

In this algorithm MCFS-ID, we can estimate the important features; specifically, m features are randomly selected out of all the d features, considering fixed subsets of m and $m \ll d$, and for every feature's subset, trees t are built and their enactment is judged. Every tree from t in the inner loop is trained and estimated on unlike, randomly selected training and test sets, which are produced by dividing the complete training data into two subsets. Further detailed steps of the algorithm are given in Figure 1.

The relative importance of features gk, $\mathrm{RI}_{\mathrm{GK}}$, is defined as

$$\mathrm{RI}_{\mathrm{GK}} = \sum_{\tau}^{\mathrm{s.t}} \mathrm{wAcc}_{\tau}^{u} \sum_{n_{\mathrm{gk}}(\tau)} \left(n_{\mathrm{gk}}(\tau)\right) \mathrm{GR}\left(\frac{\mathrm{no.in}\, n_{\mathrm{gk}}(\tau)}{\mathrm{no.in}\, \tau}\right)^{v}, \quad (8)$$

where s.t overall trees are denoted by summation; the tree on which the split is constructed on the feature gk has all nodes $n_{\mathrm{gk}}(\tau)$; the τth tree has weighted accuracy denoted by wAcc_{τ}; the gain ratio for the node $n_{\mathrm{gk}}(\tau)$ is denoted by $\mathrm{GR}(n_{\mathrm{gk}}(\tau))$; $\mathrm{no.in}\, n_{\mathrm{gk}}(\tau)$ stands for the number of samples in the node $n_{\mathrm{gk}}(\tau)$; $\mathrm{no.in}\, \tau$ symbolizes the number of samples in the root of the τth tree; and fixed positive real values of u and v are now set to 1 by default [30]. For computational causes, the normalizing factor $(\mathrm{no.in}\, \tau)$ is present mainly, which has the same value for all τ. Furthermore, m, s, and t are three parameters to be set by an experimenter. The expected constraint is that s is not too large for the choice of subset size m of features selected for each series of t experiments.

Once the ranking of the feature is completed by the MCFS-ID algorithm, a natural issue can be raised about potential interdependencies between informational features. Interdependence between features is frequently modelled using interactions, such as those in experimental design and analysis of variance [42]. Possibly, the most extensively used approach to identifying interdependencies is to find correlations between features or to find groups of characteristics that perform in the same sense [43]. In this approach, concentration is on ascertaining the characteristics that "support" in defining whether a sample is of a particular class. The interdependency (ID) graph is based on collecting information given by all the s.t trees (see Figure 1). To see how to create an ID graph, assume that, in each crowd of classification trees, each node represents an attribute, on which a partition is created. Now, for each node in each classification tree, all its integrated nodes can be kept in mind, on which the node is concerned, and a node is equipped with the attribute that is displayed in this manner, and any directed strand found in this way is actually an edge which combines two distinct characteristics in a directed way. The edges are found in all the path s.t MCFS-ID trees obviously, and the same edge may occur more than once in a single tree. The strong point of interdependence between the two nodes, essentially two features, is connected to a directed edge; the ID weight for a given edge, or the weight of the ID in short, is equal to the gain ratio (GR) in the multiplication node by a fraction. Thus, for node $n_k(\tau)$ in the τ th tree, $\tau = 1, \ldots,$ s.t, and its antecedent node $n_i(\tau)$, the ID weight of the directed edge from $n_i(\tau)$ to $n_k(\tau)$ is denoted $[n_i(\tau) \longrightarrow n_k(\tau)]$, which equals

$$\omega\left[n_i(\tau) \longrightarrow n_k(\tau)\right] = \mathrm{GR}\left(n_{\mathrm{gk}}(\tau)\right)\left(\frac{\mathrm{no.in}\, n_k(\tau)}{\mathrm{no.in}\, n_i(\tau)}\right), \quad (9)$$

where $\mathrm{GR}(n_{\mathrm{gk}}(\tau))$ denotes the gain ratio for the node $n_k(\tau)$, $\mathrm{no.in}\, n_{\mathrm{k}}(\tau)$ stands for the number of samples in the node

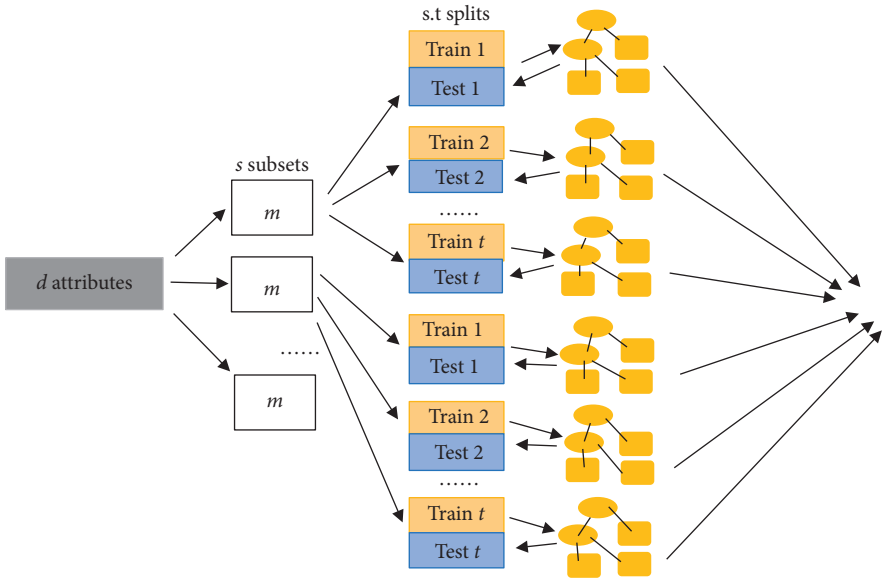


Figure 1: Block diagram of the main step of the MCFS procedure.

$n_k(\tau)$, and no.in $n_i(\tau)$ stands for the number of samples in the node $n_i(\tau)$.

3. The Proposed Framework: MCFS-ID Algorithm-Based Selection Framework

Specifically, a key objective of this study is to develop a new framework for the selection of meteorological stations by incorporating the MCFS-ID algorithm and SDI time-series data. To achieve this work, this section comprises the MCFS-ID-based computational propagation for the choice of meteorological stations. The proposed framework has two steps which are given in the flow chart (see Figure 2). They are detailed as follows:

(1) *Defining Region*. This progression adjudicates the selection of region for drought monitoring. In this step, a specific region is assimilated for regional drought monitoring. In such a manner, a suitable selection of region will strengthen accurate and efficient drought mitigation policies at the province or country level.

(2) *Defining Meteorological Station*. After the selection of a significant region, suitable selection of meteorological stations/monitoring stations is suggested. We know that long climatic information plays a significant role in the model structure and measurable statistical inferences. Along these lines, the meteorological stations, which have a rich drought-monitoring observation history, are suggested. After defining and characterizing the above two points, the stepwise execution of the proposed structure comprises 3 phases. The detailed clarifications and explanation are given in the following sections.

3.1. Phase 1: The Choice of Drought Indices and Their Estimation. This phase comprises the selection of a drought indicator from the list of all available drought indicators of the SDI procedure and the estimation procedures. Numerous studies have given various drought indicators for the standardized procedure of the drought index [24]. In Section 2, we have illuminated a brief summary of various SDI indicators and their applications in various regions. Similar to SDI procedures, recent developments also focus on the parametric- and nonparametric-based estimation [44]. Therefore, this phase is important for accurate regional drought monitoring and its analysis.

The foremost important and major concern of this phase is to select the climatic parameters and the time scale for the estimation of the multiscalar drought index. Conditional in nature, it depends on climate, soil type, and tropical status, and several drought indices required various climatic parameters such as temperature, precipitation, solar radiation, and humidity. Therefore, optimized selection of drought indices and their estimation procedure can meaningfully contribute to accurate and reliable drought monitoring. In particular, this step involves a deep knowledge of the following issues:

(i) The identification of the nature of the gauging station and the accessibility of the time-series data on the climatic parameters.

(ii) The proper selection of the multiscalar drought indicator (i.e., SPI, SPEI, and SPTI) that can be accomplished with the available data.

(iii) The type of drought with its corresponding time scale. In this step, the time scale of multiscalar drought indices is designated. For example, short time scales are suggested for meteorological data [45], whereas a longer time scale is recommended for the monitoring of agricultural and hydrological drought [46].

3.2. Phase 2: Configuring MCFS-ID Algorithm. This phase defines and constitutes the MCFS-ID algorithm on the

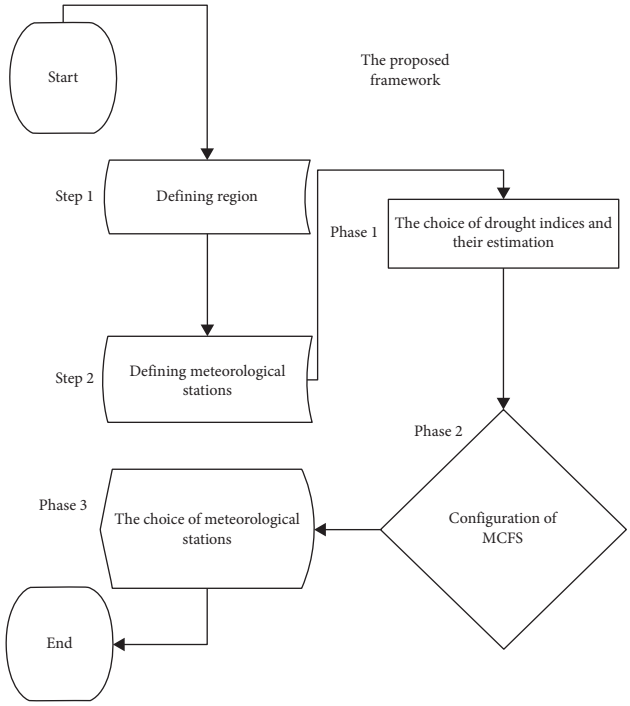


Figure 2: Flow chart of the proposed framework.

time-series data of the SDI of various meteorological stations. By incorporating the MCFS algorithm, we were able to decide which stations are more important for the reanalysis purpose. Selection of important stations is based on relative importance (RI) values. The station which has higher values of RI is considered important accordingly. Furthermore, with the help of fitting graphics, the facts about stations are obtained for selecting the most important station among other stations. The colour concentration of a node is proportional to the corresponding feature's RI. The node size is proportional to the number of edges associated with that node. The level of darkness and width of an edge are proportional to the ID weight of that edge.

3.3. Phase 3: The Choice of Meteorological Stations. In this phase, a meteorological station is identified from the ID graph on the basis of RI values for the station. The station which has a higher RI value is considered the most important than other stations which are being compared in this particular study. By careful implementation of MCFS-ID, this study suggests identifying some important meteorological stations according to their importance. The first step is to find the RI values corresponding to their ID weights by configuring the MCFS-ID algorithm on the time-series data of the SDI.

4. Application

In this section, the application of the proposed framework is discussed. The preliminary application of the proposed framework is presented in Punjab Province of Pakistan (see Figure 3). Long-term time-series data of precipitation and temperature are required for index calculation. Therefore, secondary data of these variables were collected for 46 years, from January 1971 to December 2017. This data set satisfies the requirements of the World Meteorological Organization (WMO) and is used for the analysis in this study.

4.1. Data and Study Area. The data were collected from Punjab Province of Pakistan, and the study area was 12 meteorological stations named Bahawalpur, Bahawalnagar, Faisalabad, Jhelum, Khanpur, Lahore, Mianwali, Multan, Murree, Rawalpindi, Sargodha, and Sialkot (see Figure 3). Agriculture sectors are significantly affected by these stations, and most of these stations have significant importance for crop and farming. The Punjab regions have rich agricultural attributes among other provinces. Therefore, the agriculture sector of Punjab Province continues to play a central role in Pakistan's economy in terms of gross domestic products (GDPs). However, several parts of the country are shockingly suffering from the severe drought condition due to the growing consequences of climate change and the effect of global warming. In 2018, moderate-to-several drought appeared in the arid land of Punjab regions. Although its intensity prevailed in other parts of the country including northern areas, moreover, the direct role of drought has been observed for rice crops.

4.2. Results. This section presents the results for the proposed framework. The framework is proposed for drought monitoring and categorization of stations, and the detailed description is given in Figure 2. The monthly data set was used to calculate the drought index for varying time scales. The 12 stations were taken into account, and the index for seven periods (one, three, six, nine, twelve, 24, and 48 months) was calculated.

4.2.1. Estimation of Drought Indices. There are varying probability distribution concepts that are used to estimate the drought indices at varying time scales. An estimation procedure, the fitting suitable probability distribution of the DAI series, is evaluated using the *propagate* R package. The CDF of those distributions, which has the smallest value of the BIC, is further standardized according to the approximation (as described in Section 2.1). This is repeated for all DAI time scales. Table 1 shows the BIC values for all the time scales of the Sialkot station. We perceive that three-parameter (3P) Weibull distribution has a minimum value of the BIC (−692.1) at the one-month time scale; however, Weibull distribution has several applications in the field of hydrology and related disciplines [47], and it has better candidacy for standardization. The generalized extreme value has the lowest BIC value (−535.0) for the three-month time scale of the DAI. The second and third choices for DAI-3 are generalized normal and Johnson SU distribution. In a similar way, DAI-6, DAI-9, DAI-12, DAI-24, and DAI-48 have gamma distribution with BIC = −543.1,

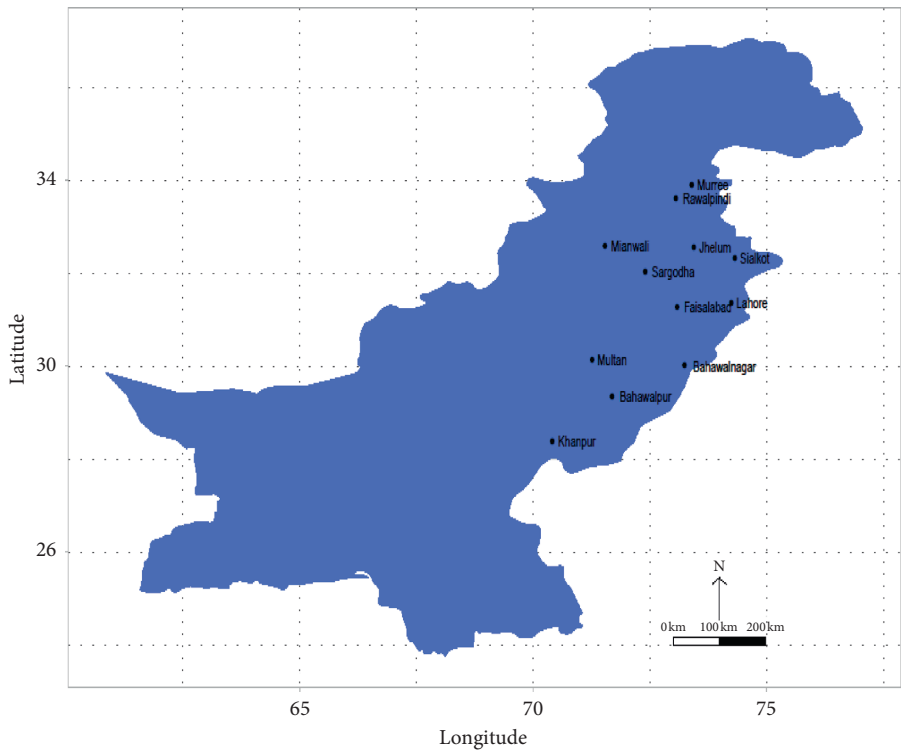


Figure 3: Locations of the study area.

Gumbel distribution with BIC = −595.8, inverse gamma distribution with BIC = −532.2, Rayleigh distribution with BIC = −651.8, and Rayleigh distribution with BIC = −510.2, respectively.

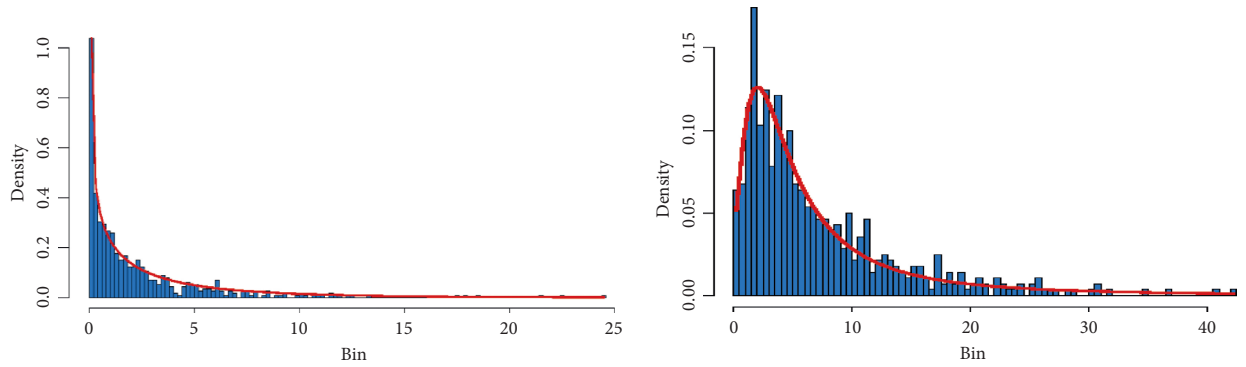
4.2.2. The Choice of the Meteorological Station under MCFS-ID. The MCFS-ID algorithm-based selection framework is designed to identify the important meteorological station in a specific region. Figure 4 shows the theoretical vs. empirical histograms of the selected distribution of all the time scale of the DAI series. It can be observed that DAI-1 has more accuracy between theoretical and empirical histograms, whereas a significant discrepancy still occurs in other time scales. This discrepancy is natural and cannot be controlled due to the behaviour of data. Recently, a probabilistic drought indicator is developed to address this discrepancy issue [48]. However, the analogy and application of this paper are beyond the description of these discrepancies. In this paper, varying distribution concept is used for the estimation of drought indices [49]. . Figure 5 shows the temporal behaviour of the various time scales of the SPTI on the same rationale, and the procedure of the drought index SPTI is estimated for all other stations.

The suitable information for stations can be obtained from the graphics. The colour intensity of a node is proportional to the corresponding features of RI. The size of a node is proportional to the number of edges related to that node. The width and level of darkness of an edge are proportional to the ID weight of that edge. The graphical representations are given for 12 stations on varying time-scale indices such as SPTI-1, SPTI-3, SPTI-6, SPTI-9, SPTI-12, SPTI-24, and SPTI-48, respectively. Since we would like to examine and analyze only the strongest ID weights (see Figure 6(a), ID graph of SPTI-1), we can see the node's size in the Lahore station for SPTI-1 is specifically large because of related edges and the intensity, and the intensity in colour for Jhelum shows the higher value of RI, and the darkness of an edge is proportional to the corresponding weight for that edge. Furthermore, for SPTI-48 (see Figure 6(g), ID graph of SPTI-48), the size of the node for Sargodha is large due to related edges, while the intensity in colour for Sialkot is specifically high because of a higher RI value of this station and also the darkness of the edges is related to the corresponding weights.

In Table 1, some statistics for precipitation and maximum and minimum temperatures are available for all selected stations, while Table 2 shows the BIC values for all stations at varying time scales. We are actually interested in finding out which stations are highly important. We can see from Table 3 that the stations which are more important have higher RI values than other stations. For SPTI-1, it can be anatomized that which six stations are more important among 12 selected stations. The results from Table 3 show that Jhelum, Sialkot, Lahore, Rawalpindi, Murree, and Mianwali have RI values of 0.140582, 0.136804, 0.12947, 0.126862, 0.112541, and

Table 1: Summary statistics of precipitation and temperature (1971–2017).

Variable	Station	Minimum	1st quartile	Median	3rd quartile	Maximum	St. dev.
Precipitation	Bahawalpur	0.00	0.00	3.00	15.55	562.60	36.90
	Bahawalnagar	0.00	0.00	4.15	21.43	263.10	37.35
	Faisalabad	0.00	1.48	12.30	40.63	438.90	52.22
	Jhelum	0.00	12.20	36.20	94.83	648.60	94.87
	Khanpur	0.00	0.00	0.60	7.00	308.00	30.35
	Lahore	0.00	5.23	21.55	69.50	640.00	86.84
	Mianwali	0.00	4.18	23.40	68.78	530.00	63.31
	Multan	0.00	0.01	4.30	20.93	231.20	32.09
	Murree	0.00	47.20	112.15	216.08	704.30	127.19
	Rawalpindi	0.00	19.70	57.30	131.35	743.30	127.24
	Sargodha	0.00	4.00	20.25	49.23	351.20	55.61
	Sialkot	0.00	9.08	37.30	94.88	917.60	128.98
Maximum temperature	Bahawalpur	18.23	26.63	35.10	38.90	44.50	7.13
	Bahawalnagar	16.10	26.35	34.95	38.53	44.60	7.42
	Faisalabad	15.80	24.40	33.50	36.80	42.80	7.25
	Jhelum	16.20	24.38	32.80	35.60	43.20	6.88
	Khanpur	19.30	27.28	35.40	39.10	44.90	7.10
	Lahore	15.20	24.75	33.05	35.60	43.10	6.85
	Mianwali	14.50	24.48	33.75	37.90	43.80	7.65
	Multan	17.96	26.33	34.80	38.50	44.20	7.30
	Murree	3.80	12.68	19.00	22.40	31.83	6.00
	Rawalpindi	14.80	22.18	30.95	34.23	41.10	6.93
	Sargodha	15.40	24.88	33.65	37.13	44.50	7.33
	Sialkot	13.50	23.50	31.80	34.23	42.90	6.99
Minimum temperature	Bahawalpur	3.70	10.68	19.15	26.33	30.30	8.30
	Bahawalnagar	3.10	10.93	19.90	26.70	30.90	8.13
	Faisalabad	1.80	8.90	17.85	25.50	29.30	8.41
	Jhelum	2.80	9.40	17.60	24.60	28.00	7.71
	Khanpur	0.90	9.38	18.10	25.80	29.60	8.44
	Lahore	4.20	11.30	19.95	26.00	29.60	7.55
	Mianwali	0.70	8.55	17.50	25.70	30.20	8.68
	Multan	2.90	10.00	19.40	27.13	30.80	8.70
	Murree	-5.50	3.00	9.95	14.60	18.40	6.32
	Rawalpindi	0.60	7.00	15.15	22.63	26.70	7.86
	Sargodha	1.00	9.39	18.55	25.60	29.20	8.47
	Sialkot	2.50	9.08	17.45	24.12	27.30	7.67



(a) (b)

Figure 4: Continued.

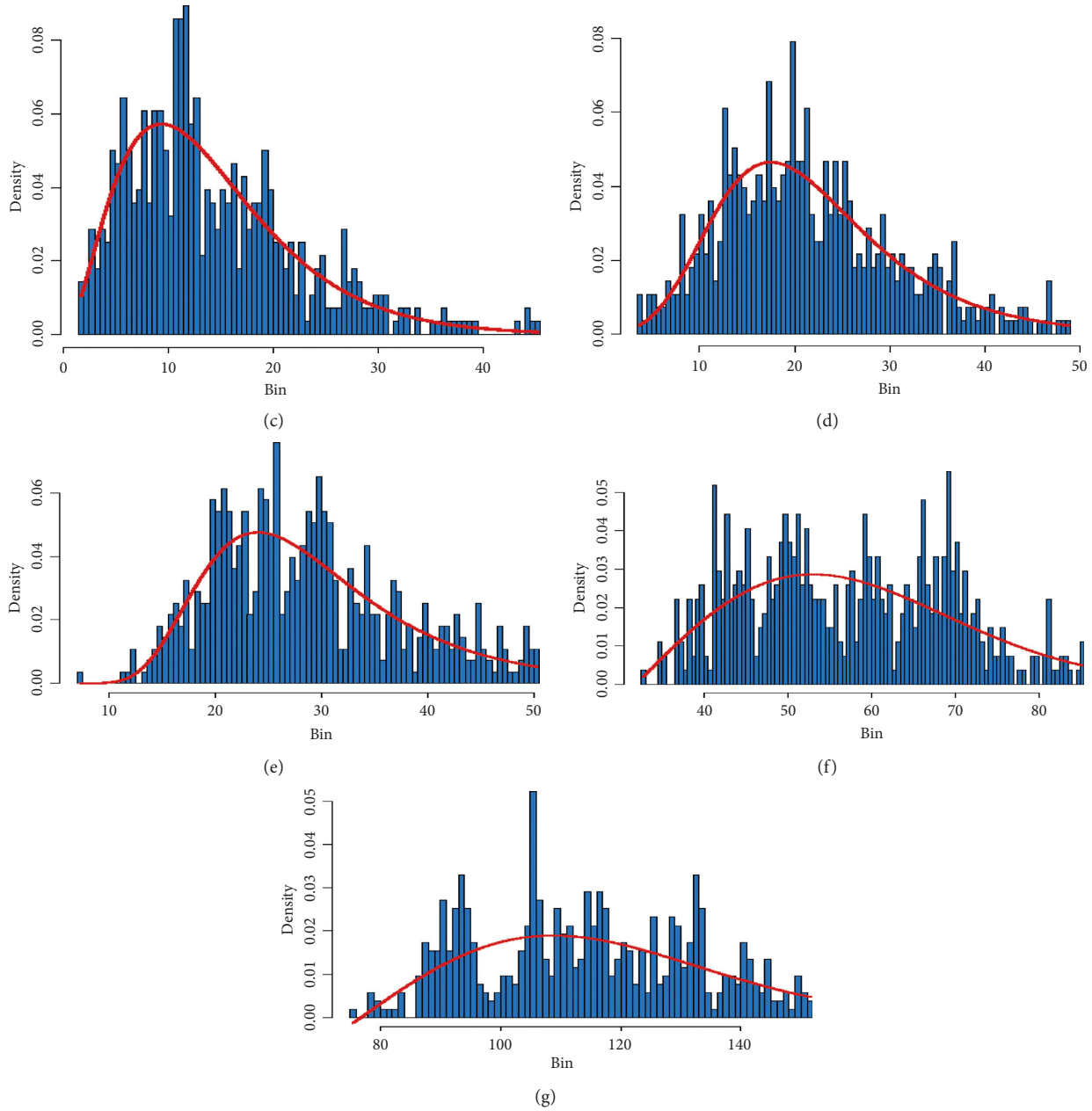


Figure 4: Theoretical vs. empirical histograms of selected distribution of various time scales for the SPTI at the Sialkot station.

0.103813, respectively, considered to be more important among 12 stations. For SPTI-48, we can see the first six stations are more important corresponding to RI values from the stations. The results from Table 2 show that Sialkot, Jhelum, Murree, Mianwali, Faisalabad, and Multan have RI values of 0.0270, 0.0260, 0.0250, 0.0240, 0.0220, and 0.0210, respectively, contemplated as more important among 12 stations.

4.3. Discussion. In this study, we proposed the MCFS-ID algorithm for identifying the important meteorological stations. The findings from this study can be helpful for metrological networks in particular regions from the data mining point of view and for analysis and reanalysis purposes. Moreover, it is better to obtain information for dependencies among existing metrological stations in a region. In the estimation procedure, we standardized the CDF of some suitable distributions on the basis of the smallest BIC. The results which are calculated for the Sialkot station show that 3P Weibull distribution is suitable for the one-month time scale, the generalized extreme value is an appropriate choice for the three-month time scale of the DAI, and so on. Furthermore, the information is obtained by setting an exhaustive framework which works by using the MCFS-ID algorithm. In addition, discovering dependencies among stations, a particular station will be considered the strongest candidate which has a higher RI value.

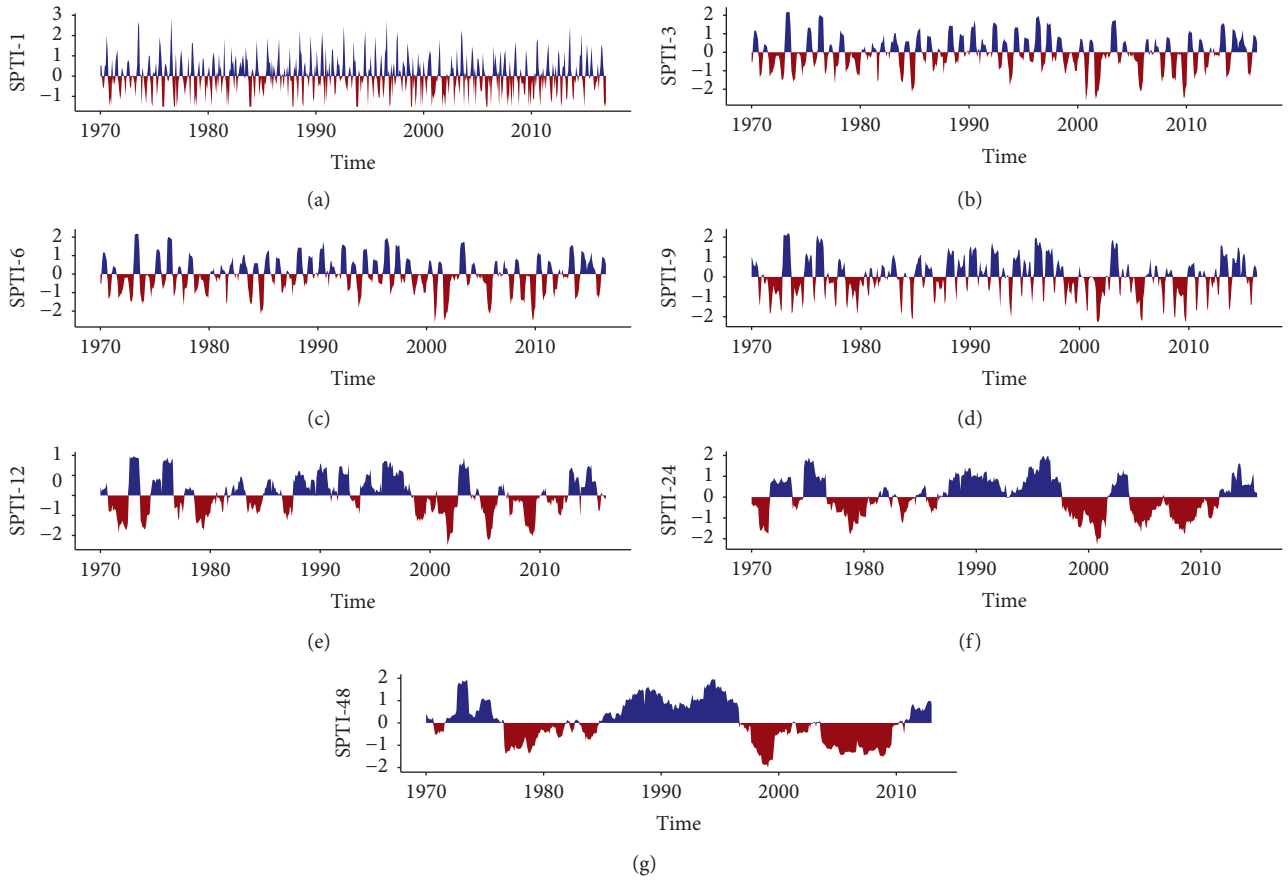


Figure 5: Temporal plots of the SPTI in various time scales for the Sialkot station. (a) DAI-1. (b) DAI-3. (c) DAI-6. (d) DAI-9. (e) DAI-12. (f) DAI-24. (g) DAI-48.

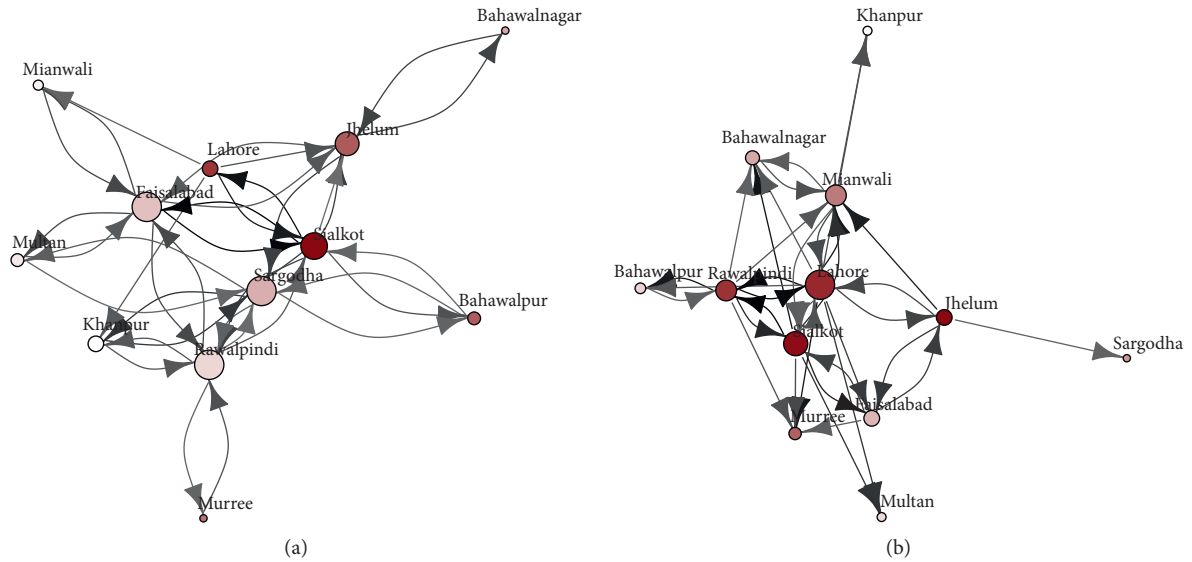


Figure 6: Continued.

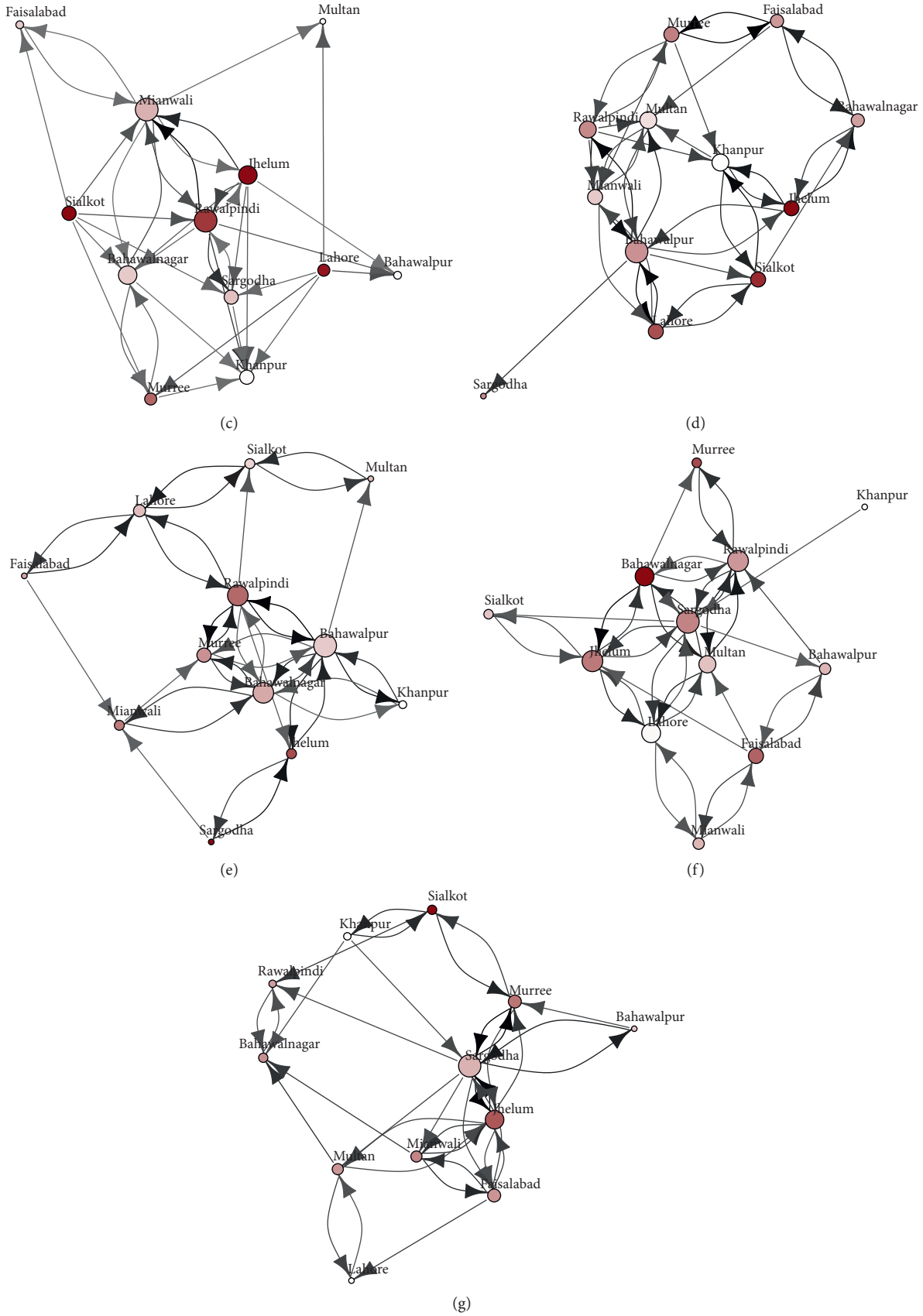


Figure 6: ID graph of (a) SPTI-1, (b) SPTI-3, (c) SPTI-6, (d) SPTI-9, (e) SPTI-12, (f) SPTI-24, and (g) SPTI-48.

Table 2: BIC of various probability distributions for DAI data of varying time scales.

Distribution	DAI-1	DAI-3	DAI-6	DAI-9	DAI-12	DAI-24	DAI-48
2P beta	−381.6	−288.4	−344.4	−379.5	−352.1	−484.3	−403.2
3P Weibull	**−692.1**	−509.2	−536.4	−586.1	−524.6	−647.9	−506.0
4P beta	−689.3	−510.4	−534.1	−585.8	−522.5	−648.4	−507.0
Arcsine	−342.3	−368.5	−400.1	−439.4	−403.4	−541.8	−455.4
Burr	−311.8	−309.4	−350.4	−384.4	−356.6	−489.0	−407.6
Cauchy	−401.7	−477.7	−517.7	−568.2	−502.0	−613.6	−492.8
Chi	−404.2	−293.2	−349.2	−384.1	−356.6	−489.0	−407.6
Chi-square	−526.1	−436.2	−473.4	−384.8	−359.9	−623.7	−408.4
Cosine	−193.5	−365.3	−505.8	−579.2	−510.7	−649.8	−509.2
Curvilinear trapezoidal	−348.7	−464.7	−500.5	−515.8	−471.9	−632.8	−463.5
Exponential	−388.7	−472.4	−459.9	−469.6	−409.3	−540.3	−432.9
F	−577.9	−283.4	−373.8	−402.7	−365.7	−494.3	−407.5
Gamma	−630.8	−518.3	**−543.1**	−594.6	−527.9	−647.7	−508.4
Generalized extreme value	−535.0	**−535.0**	−539.0	−591.6	−527.7	−643.1	−504.3
Generalized normal	−594.4	−534.4	−539.0	−591.3	−527.7	−643.1	−504.1
Gumbel	−373.6	−501.5	−542.9	**−595.8**	−531.6	−646.8	−507.5
Inverse chi-square	−275.9	−363.5	−382.9	−410.2	−372.1	−500.9	−412.7
Inverse gamma	−507.4	−501.5	−522.2	−582.6	**−532.2**	−647.2	−508.3
Inverse Gaussian	−418.2	−511.7	−532.2	−588.8	−531.5	−648.0	−508.5
Johnson SB	−595.1	−530.0	−534.6	−586.8	−523.3	−531.9	−423.9
Johnson SU	−589.6	−534.4	−534.6	−587.9	−523.3	−638.5	−499.8
Laplace	−408.6	−486.4	−538.9	−584.6	−508.3	−623.5	−498.9
Logistic	−353.2	−480.6	−532.0	−589.7	−517.5	−640.6	−505.6
Log-normal	−579.8	−527.0	−537.3	−590.9	−531.3	−647.7	−508.5
Normal	−339.1	−472.5	−529.1	−585.8	−515.6	−645.8	−507.6
Rayleigh	−352.6	−485.9	−537.8	−589.1	−517.7	**−651.8**	**−510.2**
Scaled/shifted t	−409.2	−482.4	−527.5	−585.4	−513.2	−641.1	−503.2
Skewed-normal	−359.5	−497.7	−537.8	−591.2	−527.3	−646.2	−506.4
Trapezoidal	−331.9	−486.9	−529.0	−579.5	−523.9	−651.4	−480.4
Triangular	−334.7	−490.5	−530.9	−583.8	−524.5	−647.0	−506.5
Uniform	−250.4	−288.6	−353.2	−482.1	−361.6	−547.8	−408.8
von Mises	−343.5	−332.8	−344.6	−379.7	−352.1	−484.3	−403.2

Table 3: Scores of relative importance (RI) for each station against each time scale.

Stations	SPTI-1	SPTI-3	SPTI-6	SPTI-9	SPTI-12	SPTI-24	SPTI-48
Bahawalpur	0.0770	0.0770	**0.0900**	0.0720	0.0290	0.0230	0.0190
Bahawalnagar	0.0900	0.0910	**0.0730**	0.0690	0.0310	**0.0310**	0.0200
Faisalabad	0.0860	0.0910	0.0680	0.0700	**0.0310**	**0.0260**	**0.0220**
Jhelum	**0.1410**	**0.1540**	**0.0900**	**0.1030**	**0.0360**	**0.0250**	**0.0260**
Khanpur	0.0670	0.0780	0.0570	0.0510	0.0260	0.0210	0.0180
Lahore	**0.1290**	**0.1500**	**0.0990**	**0.0870**	0.0290	0.0220	0.0170
Mianwali	**0.1040**	0.0990	0.0590	0.0600	**0.0330**	0.0230	**0.0240**
Multan	0.0760	0.0780	0.0610	0.0560	0.0290	0.0230	**0.0210**
Murree	**0.1130**	**0.1250**	**0.0850**	**0.0750**	**0.0320**	**0.0300**	**0.0250**
Rawalpindi	**0.1270**	**0.1400**	0.0640	**0.0740**	**0.0350**	**0.0260**	0.0200
Sargodha	0.0940	**0.0940**	0.0710	**0.0730**	**0.0410**	**0.0250**	0.0200
Sialkot	**0.1370**	**0.1570**	**0.1080**	**0.0950**	0.0280	0.0230	**0.0270**

5. Conclusion

The meteorological data play an important role in drought monitoring. Therefore, the choice of meteorological stations in a specific region has substantial importance. In this study, we found that the Jhelum station has more relative importance for SPTI-1 and SPTI-9 indices, while Sialkot has regional importance for studying SPTI-3, SPTI-6, and SPTI-48 indices based on our proposed method MCFS-ID for identifying the important meteorological stations. In summary, our framework can discover dependencies among stations for up-to-date real-time drought-monitoring systems. It can be useful for making informed mitigation policies and for developing an early warning system for drought monitoring.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Ethical Approval

This study was conducted in accordance with the ethical standards of the responsible committee on human experimentation and with the latest (2008) version of the 1975 Helsinki Declaration.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors are very grateful to the deanship of scientific research at King Khalid University, Abha, Saudi Arabia, for the financial support through the General Research Program under project number GRP-73-41.

References

[1] I. R. Tannehill, "Drought, its causes and effects," *Soil Science*, vol. 64, no. 1, p. 83, 1947.

[2] C. S. Szinell, A. Bussay, and T. Szentimrey, "Drought tendencies in Hungary," *International Journal of Climatology*, vol. 18, no. 13, pp. 1479–1491, 1998.

[3] B. He, A. Lü, J. Wu, L. Zhao, and M. Liu, "Drought hazard assessment and spatial characteristics analysis in China," *Journal of Geographical Sciences*, vol. 21, no. 2, pp. 235–249, 2011.

[4] D. A. Wilhite, *Drought as a Natural Hazard: Concepts and Definitions*, Drought Mitigation Center Faculty Publications, Washington, DC, USA, 2000.

[5] P. G. Whitehead, R. L. Wilby, R. W. Battarbee, M. Kernan, and A. J. Wade, "A review of the potential impacts of climate change on surface water quality," *Hydrological Sciences Journal*, vol. 54, no. 1, pp. 101–123, 2009.

[6] S. Parry, C. Prudhomme, R. L. Wilby, and P. J. Wood, "Drought termination," *Progress in Physical Geography: Earth and Environment*, vol. 40, no. 6, pp. 743–767, 2016.

[7] P. H. Gleick, "Water, drought, climate change, and conflict in Syria," *Weather, Climate, and Society*, vol. 6, no. 3, pp. 331–340, 2014.

[8] L. A. Patterson, B. D. Lutz, and M. W. Doyle, "Characterization of drought in the south Atlantic, United States," *JAWRA Journal of the American Water Resources Association*, vol. 49, no. 6, pp. 1385–1397, 2013.

[9] T. A. Alpino, A. R. M. D. Sena, and C. M. D. Freitas, "Desastres relacionados à seca e saúde coletiva—uma revisão da literatura científica," *Ciência & Saúde Coletiva*, vol. 21, no. 3, pp. 809–820, 2016.

[10] S. Bokal, A. Grobicki, J. Kindler, and D. Thalmeinerova, "From national to regional plans—the integrated drought management programme of the global water partnership for central and eastern Europe," *Weather and Climate Extremes*, vol. 3, pp. 37–46, 2014.

[11] A. Iglesias, L. Garrote, A. Cancelliere, F. Cubillo, and D. A. Wilhite, *Coping With Drought Risk in Agriculture and Water Supply Systems: Drought Management and Policy Development in the Mediterranean*, vol. 26, Springer Science & Business Media, Berlin, Germany, 2009.

[12] N. Gerber and A. Mirzabaev, "Benefits of action and costs of inaction: drought mitigation and preparedness–a literature review," *Drought and Water Crises*, World Meteorological Organization, Geneva, Switzerland, 2017.

[13] V. Acácio, J. Andreu, D. Assimacopoulos, C. Bifulco, A. di Carli, S. Dias et al., "Review of current drought monitoring systems and identification of (further) monitoring requirements," DROUGHT-R&SPI Technical Report, (6), Wageningen Environmental Research, Wageningen, Netherlands, 2013.

[14] D. A. Wilhite, M. V. K. Sivakumar, and R. Pulwarty, "Managing drought risk in a changing climate: the role of national drought policy," *Weather and Climate Extremes*, vol. 3, pp. 4–13, 2014.

[15] J. Andreu, A. Solera, J. Paredes-Arquiola, D. Haro-Monteagudo, and H. van Lanen, "Enhancing drought Monitoring and Early Warning by linking indicators to impacts," in *Drought: Research and Science-Policy Interfacing*, pp. 303–308, CRC Press, Boca Raton, FL, USA, 2015.

[16] J. Estévez, P. Gavilán, and J. V. Giráldez, "Guidelines on validation procedures for meteorological data from automatic weather stations," *Journal of Hydrology*, vol. 402, no. 1-2, pp. 144–154, 2011.

[17] P. Koeniger and A. Margane, "Stable isotope investigations in the jeita spring catchment," *BGR, Special Report on Protection of Jeita Spring*, vol. 12, p. 48, 2014.

[18] J. M. Colston, T. Ahmed, C. Mahopo et al., "Evaluating meteorological data from weather stations, and from satellites and global models for a multi-site epidemiological study," *Environmental Research*, vol. 165, pp. 91–109, 2018.

[19] N. T. Turab, P. T. T. Truc, N. D. Liem, and N. K. Loi, "Optimal selection of number and location of meteo-hydrological monitoring networks on vu gia–thu bon river basin using GIS," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, no. 3, pp. 324–328, 2016.

[20] R. Arsenault and F. Brissette, "Determining the optimal spatial distribution of weather station networks for lumped and distributed hydrological modelling purposes using RCM datasets," *Journal of Hydrometeorology*, vol. 15, no. 1, pp. 517–526, 2014.

[21] M. Borga, E. N. Anagnostou, G. Blöschl, and J.-D. Creutin, "Flash flood forecasting, warning and risk management: the HYDRATE project," *Environmental Science & Policy*, vol. 14, no. 7, pp. 834–844, 2011.

[22] A. K. Mishra and P. Coulibaly, "Developments in hydrometric network design: a review," *Reviews of Geophysics*, vol. 47, no. 2, 2009.

[23] A. Berne, G. Delrieu, J. D. Creutin, and C. Obled, "Temporal and spatial resolution of rainfall measurements required for urban hydrology," *Journal of Hydrology*, vol. 299, no. 3-4, pp. 166–179, 2004.

[24] M. D. Svoboda, B. A. Fuchs, C. C. Poulsen, and J. R. Nothwehr, "The drought risk atlas: enhancing decision support for drought risk management in the United States," *Journal of Hydrology*, vol. 526, pp. 274–286, 2015.

[25] H. Akbari, G. Rakhshandehroo, A. H. Sharifloo, and E. Ostadzadeh, "Drought analysis based on standardized precipitation index (SPI) and streamflow drought index (SDI) in Chenar Rahdar river basin, Southern Iran," in *Proceedins of the Conference Watershed Management Symposium 2015*, pp. 11–22, Reston, VA, USA, August 2015.

[26] Y. Bayissa, S. Maskey, T. Tadesse et al., "Comparison of the performance of six drought indices in characterizing historical drought for the upper blue nile basin, Ethiopia," *Geosciences*, vol. 8, no. 3, p. 81, 2018.

[27] J. Bezdan, A. Bezdan, B. Blagojević et al., "SPEI-based approach to agricultural drought monitoring in Vojvodina region," *Water*, vol. 11, no. 7, p. 1481, 2019.

[28] R. L. Nikolić-Đorić, C. W. Dawson, and E. M. Barrow, "SDSM—a decision support tool for the assessment of regional climate change impacts," *Environmental Modelling & Software*, vol. 17, no. 2, pp. 145–157, 2002.

[29] H. Hisdal and L. M. Tallaksen, "Estimation of regional meteorological and hydrological drought characteristics: a case study for Denmark," *Journal of Hydrology*, vol. 281, no. 3, pp. 230–247, 2003.

[30] M. Dramiński, M. Kierczak, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection and interdependency discovery in supervised classification," in *Advances in Machine Learning II*, pp. 371–385, Springer, Berlin, Germany, 2010.

[31] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[32] T. B. McKee, N. J. Doesken, and J. Kleist, "The relationship of drought frequency and duration to time scales," in *Proceedingsof the 8th Conference on Applied Climatology*, American Meteorological Society¸ vol. 17, no. 22, pp. 179–183, Boston, MA, USA, January 1993.

[33] S. M. Vicente-Serrano, S. Beguería, and J. I. López-Moreno, "A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index," *Journal of Climate*, vol. 23, no. 7, pp. 1696–1718, 2010.

[34] Z. Ali, I. Hussain, M. Faisal et al., "A novel multi-scalar drought index for monitoring drought: the standardized precipitation temperature index," *Water Resources Management*, vol. 31, no. 15, pp. 4957–4969, 2017.

[35] A. N. Spiess, *Propagate: Propagation of Uncertainty. R Package Version 1.0-4*, 2014, https://CRAN.R-project.org/package=propagate. R package version, 1-0.

[36] C. Cunnane, "Unbiased plotting positions—a review," *Journal of Hydrology*, vol. 37, no. 3-4, pp. 205–222, 1978.

[37] Y. Li, C. Campbell, and M. Tipping, "Bayesian automatic relevance determination algorithms for classifying gene expression data," *Bioinformatics*, vol. 18, no. 10, pp. 1332–1339, 2002.

[38] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[39] L. Breiman and A. Cutler, *Random Forests-Classification/Clustering Manual*, 2008.

[40] C. Strobl and A. Zeileis, *Danger: High Power!–Exploring the Statistical Properties of a Test for Random Forest Variable Importance*, Ludwig Maximilian University of Munich, Munich, Germany, 2008.

[41] S. Bornelöv and J. Komorowski, "Selection of significant features using Monte Carlo feature selection," in *Challenges in Computational Statistics and Data Mining*, pp. 25–38, Springer, Cham, Switzerland, 2016.

[42] M. Dramiński, M. J. Dabrowski, K. Diamanti, J. Koronacki, and J. Komorowski, "Discovering networks of interdependent features in high-dimensional problems," in *Big Data Analysis: New Algorithms for a New Society*, pp. 285–304, Springer, Cham, Switzerland, 2016.

[43] M. Dramiński and J. Koronacki, "Rmcfs: an R package for Monte Carlo feature selection and interdependency discovery," *Journal of Statistical Software*, vol. 85, no. 1, pp. 1–28, 2018.

[44] T. Soľáková, C. De Michele, and R. Vezzoli, "Comparison between parametric and nonparametric approaches for the calculation of two drought indices: SPI and SSI," *Journal of Hydrologic Engineering*, vol. 19, no. 9, Article ID 04014010, 2013.

[45] N. B. Guttman, "Comparing the palmer drought index and the standardized precipitation index," *Journal of the American Water Resources Association*, vol. 34, no. 1, pp. 113–121, 1998.

[46] E. Gidey, O. Dikinya, R. Sebego, E. Segosebe, and A. Zenebe, "Analysis of the long-term agricultural drought onset, cessation, duration, frequency, severity and spatial extent using Vegetation Health Index (VHI) in Raya and its environs, Northern Ethiopia," *Environmental Systems Research*, vol. 7, no. 1, p. 13, 2018.

[47] D. R. Nielsen, M. Kutílek, and M. B. Parlange, "Surface soil water content regimes: opportunities in soil science," *Journal of Hydrology*, vol. 184, no. 1-2, pp. 35–55, 1996.

[48] Z. Ali, I. Hussain, M. Faisal et al., "Annual characterization of regional hydrological drought using auxiliary information under global warming scenario," *Natural Hazards and Earth System Sciences Discussions*, pp. 1–20, 2019.

[49] J. H. Stagge, L. M. Tallaksen, L. Gudmundsson, A. F. Van Loon, and K. Stahl, "Candidate distributions for climatological drought indices (SPI and SPEI)," *International Journal of Climatology*, vol. 35, no. 13, pp. 4027–4040, 2015.