Evaluation of spatial–temporal characteristics of precipitation using discrete maximal overlap wavelet transform and spatial clustering tools
Kiyoomars Roushangar, Mohsen Moghaddas, Roghayeh Ghasempour and Farhad Alizadeh

ABSTRACT
In the present study, classical and proposed methods were used to investigate the monthly precipitation characteristics of 30 stations in the southeastern United States during 1968–2018. Maximal overlap discrete wavelet transform (MODWT) as preprocessing method and K-means clustering method were used. First, the monthly precipitation time series of stations were decomposed into several subseries using MODWT and considering db as the mother wavelet. Then, the energy values of these subseries were calculated and used as inputs in K-means and radial basis functions (RBF) methods. The optimum number of clusters obtained for the considered stations in both classical and proposed methods was five clusters. In order to use the data as the input of the RBF method, the data correlation was evaluated by variogram. Based on the results of clustering and in accordance with the latitude and longitude variations of the stations, it was found that with increasing the energy of the clusters, the amount of precipitation in the stations decreased and vice versa. The silhouette coefficient of clustering for the classical method obtained was 0.3 and for the proposed method it was 0.8, which indicates better clustering of the selected area using the proposed method.

Key words: clustering, MODWT, precipitation, RBF, spatial variation

HIGHLIGHTS
- Discussing important climatic elements which affect the hydrological cycle.
- Discussing temporal-spatial variations of precipitation.
- Two classical and proposed methods were used to investigate the monthly precipitation characteristics.
- Maximal Overlap Discrete Wavelet Transform (MODWT) as pre-processing method and K-means clustering method were used.
- The data correlation was evaluated by variogram and covariance graphs.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (http://creativecommons.org/licenses/by/4.0/).

INTRODUCTION

Precipitation variation assessment over a large area can provide valuable information for water resources management and engineering issues, particularly in a changing climate (Mishra et al. 2009; Wei et al. 2017). The variability of precipitation through its participation in the global hydrologic and energy cycles is important to understand the behavior and the Earth’s climatic system changes (Lettenmaier et al. 1994; Karl & Knight 1998; Lins & Slack 1999; Douglas et al. 2000; Mauget 2005; Zhang & Mann 2005; Small et al. 2006). Also, understanding spatial and temporal variation of precipitation will provide knowledge into the management and arranging of precipitation subordinate exercises (Kolivras & Comrie 2007; Nischitha et al. 2014). Many researchers have studied global and regional precipitation variations and tried to find the impacts of climate changes on the hydrological cycle and human life. Buytaert et al. (2006) investigated rainfall data from 14 rain gauges in the western mountain range of the Ecuadorian Andes. They studied spatial and temporal rainfall patterns and showed that spatial variability in average rainfall was very high. Also, significant correlations were found between average daily rainfall and geographical location. Cheng et al. (2008) evaluated the rain-gauge network using geo-statistical methods to calculate the mean precipitation in areas without stations. They showed that annual rainfall exhibits a significant orographic effect and less spatial variability, whereas hourly rainfall exhibits higher variability in space and the spatial variation structures vary among different storm types. Chu et al. (2010) investigated the temporal–spatial distribution of the precipitation of several watersheds in China. The produced various distribution maps showed that for the entire watersheds the precipitation of 1958–2007 decreased except for the spring season. The decline trend was significant in summer. The annual and seasonal precipitation amounts and changing trends were found to be different in different regions and seasons. Unal et al. (2012) analyzed annual, wet and dry seasons’ precipitation records for the period of 1961 to 2008 from 271 stations in Turkey via the rotated empirical orthogonal function (REOF), the Mann–Kendall trend test, and the continuous wavelet transform (WT) method. The obtained results showed that the decreasing wet/dry season precipitation that was observed throughout the country, except the northeast coasts and eastern parts of Turkey, had a strong impact on the economic livelihood of the region, especially on agricultural production, drinking water supply, and hydroelectricity production. Zhang et al. (2017) investigated the spatial and temporal patterns of rainy season precipitation from 1960 to 2014 using principal component analysis. The results showed that Niño-3.4 SST in winter positively impacts the subsequent year’s rainy season precipitation. Mengmeng & Baoxiang (2019) analyzed the trend of precipitation change in Shandong Province through the climate trend rate. Through multiple methods, they identified the interannual, seasonal abrupt points. The results showed that in terms of space, the annual precipitation in each region of Shandong Province was basically reduced to a different extent, and the southeast coastal region had a larger reduction than the northwest inland region. Aydin & Raja (2020) evaluated the changing spatial and temporal characteristics of precipitation on August 24, 2015 and November 11–12, 2015 that triggered flash floods and landslides in Artvin in Turkey. They used spatiotemporal (ST) kriging as a tool to investigate the ST patterns. They showed that the analysis of spatiotemporal characteristics of heavy rainfall events represents an indispensable basis to ensure that necessary precautions regarding flash flood events in the city are taken.

Since most hydrological time series are non-stationary, trend dependent, or with seasonal fluctuations, therefore, the use of other methods for identifying the dominant periodicities of time series is necessary. Wavelet analysis (WA) is one of the most common approaches used by researchers for this aim (Kumar & Foufoula 1993; Adamowski et al. 2009; Chou 2013; Roushangar et al. 2018). Hybrid wavelet analysis has been used to improve the ability of models to capture the multiscale features of hydrological time series (e.g., Agarwal et al. 2017; Farajzadeh & Alizadeh 2017). By applying the WA to a signal, it is transformed into time–frequency space. Such a transformation can be useful to identify the dominant periodicities of variability and
alteration of them over time (Adamowski et al. 2013; Joshi et al. 2016; Farajzadeh & Alizadeh 2017; Roushangar et al. 2019). Therefore, WA is a beneficial method to analyze the localized changes of power of a specific signal. WA has been used as a useful method for breaking down and excavating complex, periodic, and irregular hydrological time series (Roushangar & Alizadeh 2018). According to Adamowski et al. (2015), the wavelet transform (WT) is able to provide information on the non-stationary time series’ localized characteristics both in temporal and frequency domains. Different methods have been developed for WT application in modeling and analysis which sometimes have resulted in the wrong application and incorrect outcome (Du et al. 2017). Based on Quilty & Adamowski (2018), time series decomposition into subseries generated by WT is the primary mistake made by some researchers in using WT, where isolation and extraction of relevant features from a given time series using WT occurs. This leads to adding error to the wavelet and scaling coefficients due to the boundary extensions’ wrong selection (Maheswaran & Khosa 2012; Roushangar & Alizadeh 2019).

On the other hand, for identifying structure in an unlabeled precipitation data set clustering techniques can be applied. Clustering techniques objectively organize data into homogeneous groups in which the within-group-object similarity is maximized and the between-group-object similarity is minimized. Most classical clustering methods use static data where features do not change with time. The static data clustering methods are partitioning, hierarchical, density-based, grid-based, and model-based methods (Jain et al. 1999). Clustering analysis is the same as the homogeneity test (Viglione et al. 2007; Mousavi et al. 2015). Recently, time series clustering methods have been developed, including raw data-based, feature-based, and model-based methods (Liao 2003; Zhou & Chan 2014; Barton et al. 2016). Since time series clustering considers features in both temporal and spatial domains, the analysis results show the transient changes and local characteristics that usually occur in climate data. The two most common cluster analysis methods are the K-means clustering method and Ward’s method. In both cases, the number of clusters needs to be determined in advance.

In general, input data for clustering have a significant effect on output results. However, all properties of input variables are not required in area clustering. Therefore, appropriate methods are needed to extract the desired characteristics from the input data. The original precipitation data contain a large amount of information, so, a multiscale approach is needed to select some of the desired data and eliminate undesirable or additional information. Due to the dynamic characteristics and non-uniform distribution of precipitation data and also the need for identifying the homogeneous precipitation areas in water resources management, a temporal–spatial model is proposed to investigate the precipitation characteristics. In this regard, monthly precipitation data from 30 precipitation stations in the United States during the period 1968–2018 were used and the temporal–spatial characteristics of precipitation were investigated using two methods, classical and proposed. In the proposed method, data were first decomposed into several subseries using maximal overlap discrete wavelet transform (MODWT) method and the energy values of each subseries were calculated. In fact, for correcting the WT misapplication, precipitation datasets were analyzed using MODWT and suitable boundary extensions were selected. These data were then used as inputs in the K-means clustering method. The main objects of this study are as shown below:

- Investigating the benefits of the MODWT in evaluating the spatial–temporal characteristics of precipitation.
- Clustering the study area based on the most dominant subseries and identifying the homogeneous precipitation areas based on the subseries energy values.
- Finding the relationship of precipitation with subseries energy values and latitude and longitude of the selected stations.

**MATERIALS AND METHODS**

**Study area**

In this study, precipitation data from 30 stations in the southeast United States and around Atlanta City in the state of Georgia were used to investigate the temporal–spatial variability of precipitation. Most of the southeast part of the United States is dominated by humid subtropical
climate and receives uniform precipitation throughout the year. Atlanta has hot and humid summers and mild winters. The mean annual precipitation in Atlanta is 50.2 inches (1,280 mm). Atlanta’s area is 347.1 square kilometers, of which, 344.9 square kilometers is land and 2.2 square kilometers water. Atlanta is located among the foothills of the Appalachian Mountains. Among major cities in the east part of the Mississippi River, Atlanta has the highest elevation.

Figure 1 and Table 1 show the location of the selected stations.

Maximal overlap discrete wavelet transform (MODWT)

MODWT was considered as the discrete wavelet transform (DWT) modified version. Both DWT and MODWT allow performance of a multi-resolution analysis which is a scale-based additive decomposition. MODWT has several merits in comparison with DWT. For example, MODWT can be properly defined for arbitrary signal length, while the DWT is limited to a signal length with an integer multiple of a power of two.

According to Percival & Walden (2000), MODWT projects a time series onto a collection of non-orthogonal basis functions (wavelets) for generating a set of wavelet coefficients. MODWT makes it possible to develop a multi-resolution analysis (MRA), which is a scale-based additive decomposition. For a time series \( P \) with \( n \) samples, the MODWT can be expressed as:

\[
P = \sum_{j=1}^{N} W_j + V_N
\]  

where \( W_j \) is wavelet coefficients which capture local fluctuations over the whole period of a time series at that

Figure 1 | Location of the study area and selected stations.
Table 1 | Information of the selected stations in the study

<table>
<thead>
<tr>
<th>Station number</th>
<th>Station name</th>
<th>State</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Station number</th>
<th>Station name</th>
<th>State</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Anderson Faa airport</td>
<td>SC</td>
<td>34.495</td>
<td>−82.71</td>
<td>S16</td>
<td>Cornelia</td>
<td>GA</td>
<td>34.515</td>
<td>−85.527</td>
</tr>
<tr>
<td>S2</td>
<td>Anderson</td>
<td>SC</td>
<td>34.532</td>
<td>−82.328</td>
<td>S17</td>
<td>Coweeta Experiment station</td>
<td>NC</td>
<td>35.073</td>
<td>−83.422</td>
</tr>
<tr>
<td>S3</td>
<td>Appling 2 NW</td>
<td>GA</td>
<td>33.557</td>
<td>−82.328</td>
<td>S18</td>
<td>Eastman 1 W</td>
<td>GA</td>
<td>32.205</td>
<td>−83.197</td>
</tr>
<tr>
<td>S4</td>
<td>Athens Ben Epps Airport</td>
<td>GA</td>
<td>33.956</td>
<td>−83.317</td>
<td>S19</td>
<td>Gainesville</td>
<td>GA</td>
<td>34.309</td>
<td>−85.864</td>
</tr>
<tr>
<td>S5</td>
<td>Athens</td>
<td>TN</td>
<td>35.434</td>
<td>−83.575</td>
<td>S20</td>
<td>Hawkinsville</td>
<td>GA</td>
<td>32.281</td>
<td>−83.452</td>
</tr>
<tr>
<td>S6</td>
<td>Atlanta Hartsfield International Airport</td>
<td>GA</td>
<td>33.645</td>
<td>−84.441</td>
<td>S21</td>
<td>Jasper 1nw</td>
<td>GA</td>
<td>34.47</td>
<td>−84.441</td>
</tr>
<tr>
<td>S7</td>
<td>Calhoun Falls</td>
<td>SC</td>
<td>34.099</td>
<td>−82.583</td>
<td>S22</td>
<td>Jonesboro</td>
<td>GA</td>
<td>33.532</td>
<td>−84.336</td>
</tr>
<tr>
<td>S8</td>
<td>Carnesville</td>
<td>GA</td>
<td>34.365</td>
<td>−83.25</td>
<td>S23</td>
<td>Louis Ville 1e</td>
<td>GA</td>
<td>33.019</td>
<td>−82.388</td>
</tr>
<tr>
<td>S9</td>
<td>Cartersville Number 2</td>
<td>GA</td>
<td>34.173</td>
<td>−85.778</td>
<td>S24</td>
<td>Macon Middle Ga Regional Airport</td>
<td>GA</td>
<td>32.692</td>
<td>−83.647</td>
</tr>
<tr>
<td>S10</td>
<td>Chattanooga Airport</td>
<td>TN</td>
<td>35.42</td>
<td>−85.197</td>
<td>S25</td>
<td>Montezuma 2nw</td>
<td>GA</td>
<td>32.325</td>
<td>−84.059</td>
</tr>
<tr>
<td>S11</td>
<td>Clarks Hill 1 W</td>
<td>SC</td>
<td>33.663</td>
<td>−82.186</td>
<td>S26</td>
<td>Taylorsville</td>
<td>GA</td>
<td>34.074</td>
<td>−84.098</td>
</tr>
<tr>
<td>S12</td>
<td>Clayton 1sw</td>
<td>GA</td>
<td>34.885</td>
<td>−83.385</td>
<td>S27</td>
<td>Toccoa</td>
<td>GA</td>
<td>34.581</td>
<td>−85.34</td>
</tr>
<tr>
<td>S13</td>
<td>Clemson university</td>
<td>SC</td>
<td>34.661</td>
<td>−82.823</td>
<td>S28</td>
<td>Walhalla</td>
<td>SC</td>
<td>34.753</td>
<td>−83.077</td>
</tr>
<tr>
<td>S14</td>
<td>Cleveland Filter Plant</td>
<td>TN</td>
<td>35.214</td>
<td>−84.785</td>
<td>S29</td>
<td>Washington</td>
<td>GA</td>
<td>33.719</td>
<td>−82.71</td>
</tr>
<tr>
<td>S15</td>
<td>Cleveland</td>
<td>GA</td>
<td>34.593</td>
<td>−83.782</td>
<td>S30</td>
<td>Woodbury</td>
<td>GA</td>
<td>32.987</td>
<td>−84.598</td>
</tr>
</tbody>
</table>

As seen in Figure 2, the MODWT-based MRA decomposes an original signal P into scaling (V3) wavelet coefficients (W1, W2, and W3) (Maslova et al. 2016). Based on Cornish et al. (2006), unlike the WDT, the MODWT does not decimate the coefficients and the number of wavelet and scaling coefficients is the same as the number of sample observations at each level of the transform. In fact, the MODWT coefficients consider the result of a simple changing in the pyramid algorithm used in computing DWT coefficients through not down sampling the output at each scale and inserting zeros among coefficients in the scaling and wavelet filters. Details on MODWT can be found in Percival & Walden (2000).

Calculating the energy of decomposed subseries via MODWT

The wavelet and scaling coefficients obtained from the MODWT method were used as inputs in the clustering and geostatistical methods in the proposed model. In the proposed model, in order to reduce the numbers of inputs

[Further content follows...]

particular scale. $V_N$ is scaling coefficients and shows the overall trend of the original signal. N is the decomposition level number. For a discrete signal $P = \{P_t, t = 0, 1, \ldots, N−1\}$, the elements of the jth level MODWT wavelet and scaling coefficients, $W_j$ and $V_j$, can be written as Equation (2):

$$W_{j,t} = \sum_{l=0}^{N-1} h_{j,l}^{\alpha} P_{t-1} \text{mod} N, \quad V_{j,t} = \sum_{l=0}^{N-1} g_{j,l}^{\alpha} P_{t-1} \text{mod} N \quad (2)$$

where $W_{j,t}$ is the tth element of the jth level MODWT wavelet coefficient; $V_{j,t}$ is the tth element of the jth level MODWT scaling coefficient; $\{h_{j,l}^{\alpha}\}$ and $\{g_{j,l}^{\alpha}\}$ are the jth level MODWT high- and low-pass filters (wavelet and scaling filters) yielded by periodizing $\{h_{j,l}\}$ and $\{g_{j,l}\}$ to length n, respectively; $\{h_{j,l}\}$ and $\{g_{j,l}\}$ are the jth level MODWT high-pass filter ($h_{j,l} = h_{j,1}/2^{j/2}$) and low-pass filter ($g_{j,l} = g_{j,1}/2^{j/2}$); $\{h_{j,l}\}$ and $\{g_{j,l}\}$ are the jth level DWT high- and low-pass filters; and L is the highest decomposition level. The filters are determined depending on the mother wavelets, as in DWT. Figure 2 represents a flowchart of a three-level MODWT.
and prevent possible errors in the output of the used models, the energy values of the subseries decomposed by MODWT were calculated for each of the wavelet and scaling coefficients using Equation (3). Then, the obtained energy values were used as a representative of each coefficient in clustering methods.

$$E = \frac{\sum (P_n^2)}{\sum P_n^2} \quad n = 1, 2, \ldots, 536$$

where $P$ is the precipitation amount, $n$ represents the month of the time series, and $E$ denotes the energy value.

**K-means clustering method**

Clustering is the process of dividing or grouping a specific set of patterns into separate groups, in which, similar patterns remain in the same cluster and different patterns locate in the other clusters. The k-means clustering is used for vector quantization and, generally, it is used in data mining cluster analysis. The aim of the k-means clustering is to partition $n$ observations into $k$ clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This leads to a partitioning of the data space into Voronoi cells. According to Pham et al. (2005), k-means minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. Given an initial set of $k$ means, the algorithm proceeds by alternating between two steps:

- Assignment step: assigning each observation to the cluster whose mean has the least squared Euclidean distance.
- Update step: calculating the new means (centroids) of the observations in the new clusters. When the assignments no longer change, the algorithm has converged (see Pham et al. 2005 for more details).

**Radial basis functions (RBF) method**

Geostatistical methods can be used to extend the obtained results to the whole studied area. Radial basis function (RBF) is a geostatistical model and has its origin in methods for providing the exact interpolation of datasets’ points in a multi-dimensional space (Powell 1987). RBF interpolation is an advanced method in approximation theory for constructing high-order accurate interpolants of unstructured data, possibly in high-dimensional spaces. The interpolant takes the form of a weighted sum of RBF. RBF is often spectrally accurate and stable for large numbers of nodes even in high dimensions (Buhmann & Dyn 1993). According to Flyer & Wright (2009), RBF interpolation has been used to approximate differential operators, integral operators, and surface differential operators. The RBF mappings give an interpolating function which passes exactly through every data point. If there is noise present on the data, the interpolating function which averages over the noise gives the best extension.

**Considered methodology**

The main aim of this study was to investigate the capability of new methods in temporal and spatial analysis of
precipitation in an area to reduce computational time, cost, errors, and noise in temporal and spatial investigation of precipitation. The MODWT-K-means framework was used for this aim. MODWT was used to extract dynamic and multiscale features of the non-stationary precipitation time series, and K-means was applied to objectively identify spatially homogeneous clusters on the high-dimensional wavelet-transformed feature space. During the modeling process, an attempt was made to reduce the number of input data in the analyzing process. The following points could be considered to reduce the noise values:

1. The number of inputs in modeling process should be reduced.
2. The probability of selecting inputs should be reduced (i.e., the selection of inputs should not be done by chance).
3. The methods with higher accuracy should be used.

Figure 3 illustrates the various stages of the research to estimate the temporal–spatial characteristics of monthly precipitation using both classical and proposed models. In this study, monthly precipitation datasets from 30 rainfall stations for the period 1968–2018 were used. Each station had 612 data. In the classical method, without any preprocessing in the precipitation data, all data were used as inputs for area clustering. The silhouette coefficient was used to validate the results. In the proposed model for investigating the precipitation variations over the selected time period, the precipitation time series were first decomposed using the MODWT method. The MODWT method was used to reduce the number of inputs. In this regard, between the subseries obtained from MODWT, the best input combination was used for clustering.

RESULTS AND DISCUSSION

Decomposition of precipitation time series and calculating the subseries energy values

To investigate the temporal–spatial variations of precipitation in the study area, time series were first decomposed to several subseries via MODWT. In this study, the db was used as mother wavelet, since this type of mother wavelet has been widely used in hydrological studies. The db wavelets provide complete support for time series, indicating that these wavelets have non-zero basic functions over a given interval. In the proposed model, in order to find the best value of the decomposition level and the best type of db, monthly time series were decomposed in several levels between two and six and numbers two to five were considered for db. Among the 20 different considered states, the best case was selected by calculating the values of root mean square error (RMSE) criteria. Therefore, the decomposition level of four and the db number of three was selected for time series decomposition. In order to use the MODWT method, some data must be removed from the left side of the time series, whereas in the DWT method it is necessary to remove data from both left and right sides. In fact, the MODWT needs an infinite signal \( P_t \), in which \( t = \ldots, -1, 0, 1, \ldots, N - 1, N \), while in reality the data are measured in a finite interval at discrete times. To use this method, the extension of time series is needed for unobserved amounts determining, \( P_0, P_1, \ldots, P_{N+1}, P_{N+2} \), prior to preprocessing. Therefore, the right end of the \( P_s \) series (i.e., \( P_{N+1}, P_{N+2}, \ldots \)) should be extended properly and special attention should be paid to the values affected by the boundary conditions. There are two methods for considering the boundary effect, which include data modification and wavelet modification. In this study, boundary conditions’ handling is performed based on Percival et al. (2011). The extension of data is done using the following equation:

\[
L_j = (2^j - 1) \times (L - 1) + 1
\]  

(4)

where \( j \) is decomposition level, \( L \) is twice the db number, and \( L_j \) is the number of omitted data. In Figure 4, the results of station S2 time series decomposition with and without data removing are shown.

Signal decomposition via MODWT will lead to two wavelet and scaling coefficients. The scaling coefficients (\( V \)) represent the wavelet transform coefficients with large resolution which show smooth trends in the series and wavelet coefficients (\( W_1, W_2, W_3, \ldots \)) provide detailed information of trends in hydrological time series. Each of the \( W \) components provides a specific period of time. For example, in monthly data, \( W_1, W_2, W_3, \) and \( W_4 \) indicate 2, 4, 8, and
16-month periods, respectively. After time series decomposition, in the next step, the energy of the subseries was calculated using Equation (3). Figure 5 shows the energy values of each subseries decomposed by MODWT for all stations. The magnitude of the energy variations in the $W_1$ was $3.53 \times 10^{20}$ to $5.78 \times 10^{20}$, in the $W_2$ was $1.84 \times 10^{20}$ to $3.58 \times 10^{20}$, in the $W_3$ was $9.36 \times 10^{19}$ to $1.73 \times 10^{20}$, and in $W_4$ was $6.74 \times 10^{19}$ to $1.01 \times 10^{20}$. Also, $V_4$ was between $19.04 \times 10^{-4}$ and $19.35 \times 10^{-4}$. As can be seen, $V_4$ and $W_2$ had the least variations and $W_4$ had the most variations among the subseries.

**Clustering the study area**

After calculating the subseries energy values, clustering was performed for both classical and proposed models and the
obtained results were compared. In the clustering process, first, the number of clusters should be determined, therefore, the number of clusters was selected between 2 and 10 and K-means operation was performed. The best number of clusters was selected based on silhouette coefficient (SC) and dispersion of stations. Accordingly, the best cluster number was found to be 5 clusters. In the classical method, all monthly precipitation data were considered as input. In this case, 612 data were selected as input for each station and clustering was performed. The silhouette coefficient for the classical method obtained was 0.31 which indicated relatively poor correlation and clustering. Figure 6(a) shows the silhouette coefficient of each cluster in the classical method. As can be seen, the cluster number 5 is a single-member cluster that showed dissimilarity to other stations. Also, clusters 2 and 3 had negative silhouette coefficients. In the proposed method, for selecting the best input data for clustering with 5 clusters, the energy

Figure 4 | Comparing the precipitation time series decomposition with and without data removing for station S2.

Figure 5 | The variations of energy values of decomposed subseries by MODWT.
values of decomposed subseries and combination of them (15 different states) were considered and the silhouette coefficient and spatial distribution of the stations were assessed. The results are presented in Table 2 and Figure 6. It was observed that when variables \( W_2 \) and \( V_4 \) were used as inputs, the value of silhouette coefficient improves up to 0.8. Another reason for selecting variables \( W_2 \) and \( V_4 \) as the best inputs was their lower data variability in comparison with the other variables. According to the results, the S18 and S24 stations with \( S = 0.98 \) had the highest, and the S30 and S22 stations with \( S = 0.11 \) and \( S = 0.37 \) had the lowest SC, respectively. In Figure 6(b), the clusters were marked on the map with different shapes. From the results, it seems that the stations are clustered appropriately. The clusters were located as: the cluster number 2 (with rhombus shape) in the southern part, the cluster number 3 (with square shape) in the north and northwest parts, the cluster number 5 (with triangle shape) in the east part, and the cluster numbers 1 and 4 (with star and circle shapes, respectively) from the east to the west parts. Also, Figure 7 shows the energy values of the decomposed subseries. According to this figure, it could be deduced...
that the subseries energy values had wide variations; however, in each cluster the subseries energy variations were almost similar to each other. For all stations, the W1 subseries had the highest energy amount and the W4 and V4 subseries had the lowest energy amount. Also, the most similarity was obtained for cluster 1 and the least similarity was obtained for clusters 3 and 4.

Hsu & Li (2010) used the wavelet transform self-organizing map (WTSOM) framework to cluster and explore spatial–temporal characteristics of the 22 years of precipitation data for Taiwan. They showed that the hybrid methods performed successfully in recognizing homogeneous hydrologic regions. Comparing the results of the classical and proposed models applied in this study showed high efficiency of the considered methodology in the selected area clustering. From the MODWT, the best input combination (i.e., $W_2 + V_4$) was used for clustering. This reduced the number of input data to one-fifth. The amount of the silhouette coefficient for this case was almost three times more than the classical method, which shows the efficiency of the selected method in zonning the area. Reducing the number of inputs, not using the usual classical methods to select the mother wavelet type and the level of decomposition in the MODWT method (which creates additional subseries), and testing a different number of clusters and selecting the most appropriate $k$ based on trial-and-error process increased the accuracy of the proposed method.

Figure 8 shows the relationship of mean precipitation with energy values for each cluster and central stations of clusters. The obtained results indicated that as the mean precipitation of the clusters increased, the energy values decreased. The highest precipitation was measured at cluster 3 (square shape) stations, where northern stations of the

<p>| Table 2 | Cluster information |</p>
<table>
<thead>
<tr>
<th>Station number</th>
<th>Cluster number</th>
<th>Monthly mean precipitation variation</th>
<th>Energy variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>1,051.4–951.4</td>
<td>$2.41 \times 10^{20}$–$2.28 \times 10^{20}$</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>978.2–959.4</td>
<td>$3.58 \times 10^{20}$–$3.3 \times 10^{20}$</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>1,525.8–1,081.8</td>
<td>$1.98 \times 10^{20}$–$1.84 \times 10^{20}$</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>1,266.7–988.1</td>
<td>$2.25 \times 10^{20}$–$2.07 \times 10^{20}$</td>
</tr>
</tbody>
</table>

Figure 7 | The energy values of decomposed subseries in each cluster.
study area and north of Atlanta are located. Also, the lowest precipitation was measured at cluster 2 (rhombus shape) stations, where the southern stations and south of Atlanta are located. According to the results, for stations and clusters located in the northern part of the selected area, the amount of precipitation increased and the amount of energy decreased. Also, for the stations and clusters located in the southern part, the amount of precipitation decreased and the amount of energy increased. As can be seen, the highest precipitation after cluster 3 was obtained for cluster 4 (circle shape). Each cluster had several stations as the members of that cluster; among them, one station was selected as the cluster central station, which had the lowest distance from the center of the cluster and the highest silhouette coefficient among the other members. Details of the central stations for each cluster are given in Table 3. According to Figure 8(b), it can be seen that with increasing energy values the precipitation values decreased and vice versa.

### Radial basis functions (RBF) method results

In this study, two important parameters of precipitation and energy were used as inputs for the RBF method which is an interpolation method. Before interpolating the precipitation and energy values of stations without data, first, data distribution should be investigated. One of the useful methods for determining the data distribution is the QQ plot. In this study, the QQ plot was used and the results showed that the data used had normal distribution. After preliminary analysis, the semi-variance graphs were plotted. Semi-variance analysis reveals the data correlations, trends, and similarity covariance between points. Figure 9(a) shows the semi-variance and covariance graphs for monthly precipitation and energy. Also, in Table 4, the characteristics of precipitation and energy semi-variance models are listed. To express the robustness of the spatial structure of a variable, the ratio $C_0/(C + C_0)$ ($C_0$ is nugget effect and $C$ is partial sill) can be used and investigated to see how much of the total variability justifies the nugget effect. Since the obtained values for both monthly precipitation and energy parameters were less than 1/2, it could be deduced that the role of the unstructured component is less than the structured component. Therefore, the investigated parameters had strong spatial structure. According to Figure 9(b), and based on the covariance between mean monthly precipitation and station distance, it could be indicated that there was a significant positive correlation between these two variables. This correlation was higher at the initial distances and as the distance increased, the

---

**Table 3 | Information of the central stations of each cluster**

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>Station number</th>
<th>Silhouette coefficient</th>
<th>Central station distance from the cluster center</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S2</td>
<td>0.94</td>
<td>$3.95 \times 10^{35}$</td>
</tr>
<tr>
<td>2</td>
<td>S18</td>
<td>0.98</td>
<td>$3.58 \times 10^{35}$</td>
</tr>
<tr>
<td>3</td>
<td>S12</td>
<td>0.96</td>
<td>$2.29 \times 10^{36}$</td>
</tr>
<tr>
<td>4</td>
<td>S13</td>
<td>0.88</td>
<td>$1.91 \times 10^{36}$</td>
</tr>
<tr>
<td>5</td>
<td>S3</td>
<td>0.85</td>
<td>$1.85 \times 10^{36}$</td>
</tr>
</tbody>
</table>
effect of station distance on the mean monthly precipitation decreased. At a distance of about 142 km, the similarity of these two parameters was zero and, after this point, variation on monthly precipitation was approximately ineffective. The covariance between energy and station distance also showed that there was a significant positive correlation between these two variables. The correlation was higher at the initial distances and, as the distance increased, the effect of the station distance on the amount of energy decreased.
The RBF method has five models. These models can be evaluated using different performance criteria to select the most appropriate model. In this study, different models with different parameters were tested and after interpolation via the RBF method the model with the lowest error (RMSE) and the highest correlation coefficient ($R^2$) was considered as the best model. The results are shown in Figure 9(c).

From the results, it was found that the best model for both mean monthly precipitation and energy variables was spline with tension, with $R^2 = 0.93$ and RMSE = 4.15 for precipitation variable and $R^2 = 0.77$ and RMSE = $2.53 \times 10^{-19}$ for energy variable.

The study area zoning was performed for the W2 subseries and mean monthly precipitation values using the spline with tension model as the best model. Due to the lesser variability of the W2 subseries compared to other subseries, this parameter was selected as the best input for the RBF method. According to Figure 10, it was observed that in the south and southeast of the studied area, monthly precipitation of the stations was lower than the other stations, and in the north and northwest parts, the monthly precipitation increased and reached the highest amount of precipitation. In the state of zoning based on energy values, it was found that the south and

---

**Table 4** | Characteristics of precipitation and energy semi-variance models

<table>
<thead>
<tr>
<th>Energy</th>
<th>Precipitation</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>479,313 (m)</td>
<td>192,932 (m)</td>
<td>Effect amplitude (R)</td>
</tr>
<tr>
<td>1.470 (mm)</td>
<td>0.891 (mm)</td>
<td>Partial sill (c)</td>
</tr>
<tr>
<td>0.378 (mm)</td>
<td>0.225 (mm)</td>
<td>Nugget effect ($c_0$)</td>
</tr>
</tbody>
</table>

---

**Figure 10** | Zoning of the studied area based on the precipitation and energy values using RBF method.
southeast of the area had the highest energy values and the northern part had the lowest energy values. The results showed an inverse relationship between the energy and monthly precipitation values.

For investigating the relationship between precipitation and latitude and longitude of the selected stations, Figure 11 was drawn. This figure shows that with increasing the stations’ longitude, the amount of precipitation increased and the amount of energy decreased. Also, with decreasing the stations’ latitude, the amounts of precipitation increased and the energy values decreased. This issue verified the obtained results of the proposed model.

**CONCLUSION**

In this study, two proposed and classical methods were used to investigate the monthly precipitation characteristics of the selected area in the United States. In the proposed method, the time series were decomposed via MODWT, different combinations of the wavelet (W) and scaling (V) coefficients were used to determine the input dataset as a basis of spatial clustering. These combinations were determined in way to cover all possible scales captured from MODWT. The proposed model’s efficiency in spatial clustering stage was verified using silhouette coefficient index. Results demonstrated superior performance of MODWT-K-means in comparison to historical-based K-means approach. It was observed that the clusters captured by MODWT-K-means approach determined homogenous precipitation areas very well (based on physical analysis). In the classical method, monthly precipitation data were used as input for clustering the study area. The results showed that in the proposed method, clustering based on the combination of W2 and V4 subseries led to better performance.
results. The best number of clusters obtained was 5. The silhouette coefficient in the classical method obtained was 0.3 and in the proposed method 0.8, which indicated appropriate clustering of the selected area using the proposed method. In the RBF modeling, data distribution was first evaluated and the correlation of the data was verified. Then, the five most commonly used RBF methods were modeled and the best model was selected based on RMSE and $R^2$. The results showed that the best model for both mean monthly precipitation and energy variables was spline with tension model. According to the results, an inverse relationship between monthly precipitation and subseries energy was obtained. It was found that the southeast of the selected area had the highest energy and the lowest precipitation values, and the northern parts had the highest precipitation and the lowest energy values. Also, variations of the precipitation and energy parameters were investigated in terms of stations’ latitude and longitude. It was observed that with increasing the stations’ longitude and decreasing their latitude the amount of precipitation increased and the energy values decreased. In general, the proposed model yielded better results than the classical model due to its higher silhouette coefficient and station similarities. Also, the proposed model performed better than the classical method due to less input data and computational time.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories (https://gis.ncdc.noaa.gov/maps/ncei/summaries/monthly).

REFERENCES


Du, K., Zhao, Y. & Lei, J. 2017 The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. Hydrology 552, 44–51.


Downloaded from http://waponline.com/hr/article-pdf/52/2/414/872794/nn9520414.pdf by guest


