Hydrology Research

British Hydrological Society

IWA PUBLISHING

Hydrology Research Vol 55 No 3, 280 doi: 10.2166/nh.2024.098

Identification of hydrologically homogenous watersheds and climate-vegetation dynamics in the Blue Nile Basin of Ethiopia

Temesgen Tsehayeneh Mihret [a,b], Fasikaw A. Zemale [a,*], Abeyou W. Worqlul [c], Ayenew D. Ayalew [d] and Nicola Fohrer [d]

[a] Faculty of Civil and Water Resources Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia
[b] Department of Water Resources and Irrigation Engineering, Assosa University, Assosa, Ethiopia
[c] International Center for Agricultural Research in the Dry Areas (ICARDA), Tunis, Tunisia
[d] Department of Hydrology and Water Resources Management, Institute for Natural Resource Conservation, Kiel University, Kiel, Germany
*Corresponding author. E-mail: Fasikaw.Atanaw@bdu.edu.et

TTM, 0000-0001-9793-9070; FAZ, 0000-0001-9778-2712; AWW, 0000-0002-7990-8446; ADA, 0000-0002-7331-8170; NF, 0000-0002-7456-6301

ABSTRACT

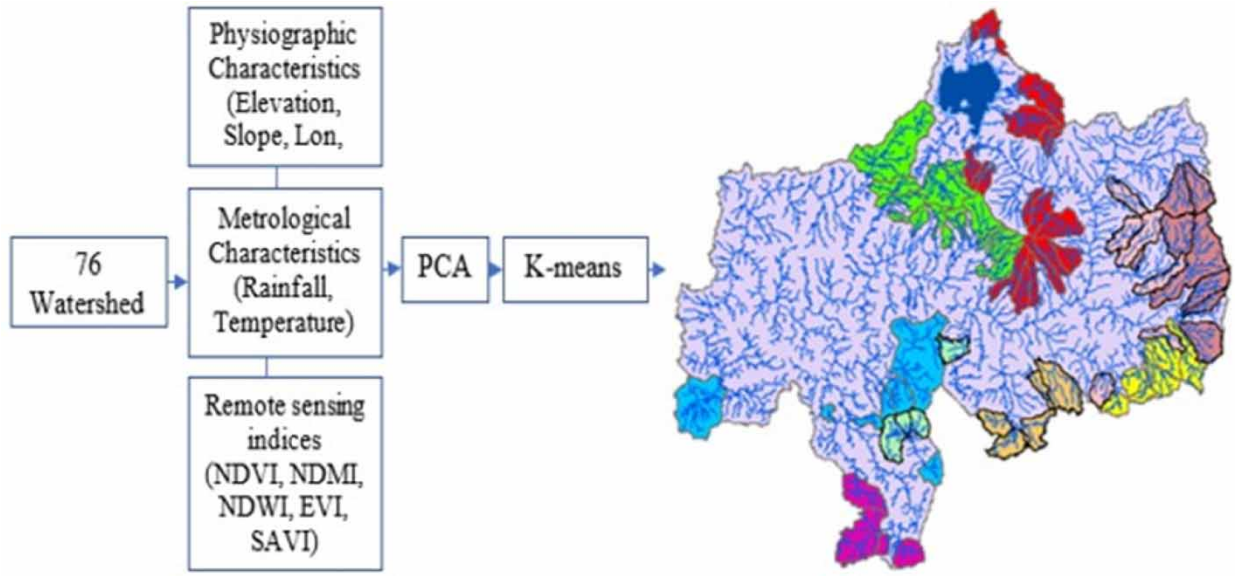
Identification of hydrologically homogenous watersheds in the Upper Blue Nile Basin of Ethiopia is challenging due to the large number of watersheds and the lack of consistent and reliable data. Traditional methods, such as expert-based classification, are time-consuming, subjective, and often not reproducible. Therefore, this study aims to identify homogenous gauged watersheds using hydrometeorological and remote sensing data. In this study 76 watersheds were delineated from a 30-m digital elevation model (SRTM-DEM). Twelve watershed characteristics were selected to aid the classification process. Three homogenous climate regions were identified using rainfall data from 42 stations, and for each homogeneous climate region, gauged watersheds were identified. Principal component analysis (PCA) and *K*-means clustering were used for classification. The PCA reduced 12 watershed characteristics into three principal components using a threshold of 80% accounted variance and eigenvalues greater than one. *K*-means clustering classified the 76 watersheds into nine homogenous clusters. In the classified regions, vegetation dynamics within three decades have also been analyzed. This helped identify trends in vegetation cover and its spatial and temporal dynamics. The results of the investigation will potentially be used for runoff prediction of ungauged watersheds and for water resource management models in the future.

Key words: Blue Nile Basin, homogeneous watersheds, *K*-means clustering, principal component analysis, remote sensing indices, vegetation dynamics

HIGHLIGHTS

- Principal component analysis and *K*-means cluster analysis were used for homogenous watershed classification.
- Use of hydrometeorological data and remote sensing indices were used for homogenous watershed classification.
- Seventy-six watersheds in the upper Blue Nile were classified to three climatological regions and nine homogenous watersheds.

GRAPHICAL ABSTRACT



INTRODUCTION

Watershed classification is a method used to group watersheds based on similar attributes such as land use, soil type, climate, topography, geology, and hydrology (Wolfe *et al.* 2019). This process is useful for predicting streamflow in ungauged basins (Choubin *et al.* 2019), sustainable environmental planning (Pascucci *et al.* 2018), and flood frequency analysis (Pallard *et al.* 2008; Farsadnia *et al.* 2014). Predicting runoff in ungauged basins also requires the classification of watersheds into hydrologically comparable groups before regionalization (Kanishka & Eldho 2017). This is especially true when there are limited financial resources and a lack of technical knowledge to undertake on ground work (Choubin *et al.* 2017). The precision of regionalization heavily relies on the precise categorization of comparable watersheds (Razavi & Coulibaly 2013; Kanishka & Eldho 2017; Ayalew *et al.* 2022). Choubin *et al.* (2017), Mosavi *et al.* (2021), and Sardooi *et al.* (2019) have demonstrated watershed classification based on average hydrological indices, physiographic features, and meteorological characteristics specific to each watershed within the basin. With the help of different machine learning clustering algorithms, such as *K*-means clustering, agglomerative hierarchical clustering, and hybrid clustering, watersheds are generally grouped into a similar region (Farsadnia *et al.* 2014; Kanishka & Eldho 2017; Sardooi *et al.* 2019) using the hydrological characteristics of each watershed. So, it is important to choose these watershed characteristics carefully. However, in the Blue Nile Basin, it is generally difficult to determine and understand the hydrological indices of the ungauged watersheds due to the lack of sufficient data. This issue may be resolved using remote sensing datasets that are widely used for runoff prediction (Choubin *et al.* 2017). Kanishka & Eldho (2020), Palcon (2021), and Chaudhary & Pandey (2022) have explored the potential of clustering watersheds into similar regions, aiming to enhance predictions for ungauged watersheds. These studies aimed to identify homogenous groups of watersheds using various dimensionality reduction techniques on watershed characteristics. Dimensionality reduction methods include both linear and nonlinear techniques. Principal component analysis (PCA) is a linear dimensionality reduction technique that reduces high-dimensional data into smaller dimensions. Several hydrology researchers have demonstrated the effectiveness of utilizing PCA prior to watershed classification (Farhan *et al.* 2017; Kanishka & Eldho 2017, 2020; Kunnath-Poovakka & Eldho 2018; Palcon 2021). In each of the studies mentioned, PCA reduced the dimensionality of the data so that the majority of the variations within the data were reduced in lower dimensions. One example of nonlinear dimensionality reduction techniques is self-organizing maps (Swain *et al.* 2016). Watershed classification in the Blue Nile River Basin, Ethiopia, is challenging due to the large number of watersheds and the lack of consistent and reliable data. Traditional methods, such as expert-based classification, are time-consuming, subjective, and often not reproducible. The Upper Blue Nile River Basin faces recurrent drought and famine due to inadequate infrastructure, such as a lack of water impounding systems and reliable irrigation schemes to address extended periods of low precipitation

(Kim & Kaluarachchi 2008; Gebregiorgis *et al.* 2013). The region's dependence on rain-fed agriculture and small-scale irrigation makes runoff estimation crucial for small watersheds. In addition, estimating runoff for ungauged watershed is essential for planning long-term strategies like hydropower generation, large-scale irrigation, and ecological protection. Understanding the temporal and spatial variability of water yield in the study area is vital for local economies and downstream countries. Despite the significance, previous studies have primarily concentrated on estimating runoff only at the outlet of the gauged watersheds (Tigabu *et al.* 2015, 2020; Ayele *et al.* 2016). In addition Kim & Kaluarachchi (2008) have attempted to develop regionalization models without identifying hydrologically homogenous watersheds. Therefore, before developing a hydrological model, there is a need for a more objective and efficient method for watershed classification in the Blue Nile River Basin of Ethiopia. Recognizing these challenges in this study, a more objective and efficient linear classification technique, PCA with *K*-means clustering, is used for classifying 76 watersheds in Blue Nile River Basin, Ethiopia, using the existing physiographic and meteorological characteristics as well as remote sensing-based watershed characteristics. One of the most important objectives and interests of the International Association of Hydrological Sciences (IAHS) is the use of remote sensing datasets to improve runoff prediction in ungauged watersheds (Sivapalan *et al.* 2003; Choubin *et al.* 2017). The use of remote sensing datasets is particularly important in regions where hydrological data may be limited, as they can provide additional information about land use, vegetation, and soil properties that can be used to improve the accuracy of runoff prediction models.

MATERIAL AND METHODS

Study area

The Blue Nile River is the most important tributary of the Nile River, providing over 60–70% of the Nile's flow at Aswan Dam (Nawaz *et al.* 2010). Both Egypt and, to a lesser extent, Sudan are almost entirely dependent on water from the Nile. This dependency creates the challenges of water resources management in these regions and is currently a subject of the international law of transboundary rivers (Waterbury 2008). The Upper Blue Nile Basin refers to the uppermost part of the

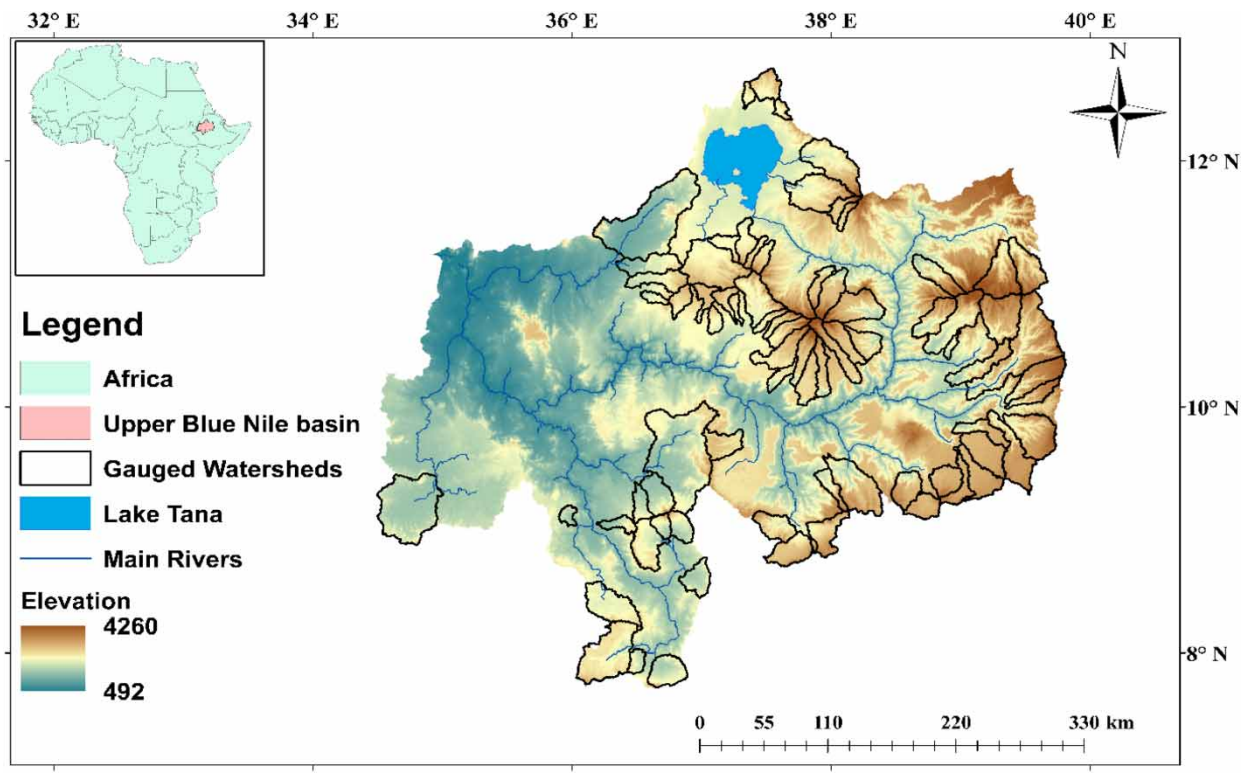


Figure 1 | Study area: Upper Blue Nile Basin and 76 selected gauged watersheds for hydrological similarity study.

Blue Nile Basin, located in Ethiopia, that originates from Lake Tana, which is located at an elevation of just under 1,800 m (Figure 1). It leaves the southeastern corner of the lake, flowing first southeast, before looping back on itself, flowing west and then turning northwest close to the border with Sudan. Until the main stream reaches the lowlands at the Ethiopian–Sudanese border at El-Diem, numerous tributaries join the main stream in the central and southern highlands of Ethiopia. By gaining a better understanding of the Upper Blue Nile Basin's homogeneous watersheds, managers can more effectively plan the utilization of water resources and mitigate natural disasters such as erosion, drought, and others, which may be influenced by the geography and climate of the watershed.

Methodology

The methodology of the study included (i) deriving the required watershed characteristics from hydrometeorological and different remote sensing datasets, (ii) normalization of watershed characteristics, (iii) multicollinearity assessment, (iv) *K*-means clustering, and (v) finding the optimum number of classes according to the classification validation criteria. Various research studies have utilized different methods to characterize watersheds, indicating the need to determine which watershed characteristics significantly affect runoff responses. Expert judgment is required to identify such characteristics. According to previous studies (Choubin *et al.* 2017; Sardooi *et al.* 2019; Wolfe *et al.* 2019), a total of 12 potentially useful watershed characteristics were selected to identify homogeneous watersheds in the Upper Blue Nile Basin. The selected watershed characteristics for the classification of watersheds into homogenous groups are presented in Table 1.

Physiographic and metrological characteristics

The Upper Blue Nile Basin boundary was obtained from the Hydro SHEDS dataset (Lehner *et al.* 2008). The physiographic characteristics, elevation and slope, were extracted from SRTM-DEM, and the spatial–temporal variations of long-term mean annual rainfall data in the Blue Nile Basin was determined by applying the inverse distance weight method on GIS environment (Figure 2).

Table 1 | Selected watershed characteristics to identify homogenous watersheds in the study area

	Description	Units	Data source
Watershed attribute			
Area	The size of each watershed	km^2	30-m SRTM-DEM (https://gdex.cr.usgs.gov/gdex)
Longitude	Longitudinal centroid value for each watershed	Degrees	
Latitude	Latitudinal centroid value for each watershed	Degrees	
Physiographic characteristics			
Elevation	Average elevation for each watershed	M	
Slope	Average slope for each watershed	%	
Meteorological characteristics			
Precipitation	Mean areal precipitation for each watershed	Mm	Ethiopian National Metrological Agency (ENMA)
Temperature	Mean temperature	°C	
Remote sensing indices			
NDVI	Mean area normalized difference vegetation index	–	12-year mean annual Landsat 8 https://earthexplorer.usgs.gov/
EVI	Mean areal enhanced vegetation index	–	
SAVI	Mean areal soil-adjusted vegetation index	–	
NDMI	Mean areal normalized difference moisture index	–	
NDWI	Mean areal normalized difference water index	–	

Note: All data were obtained on October 2, 2023.

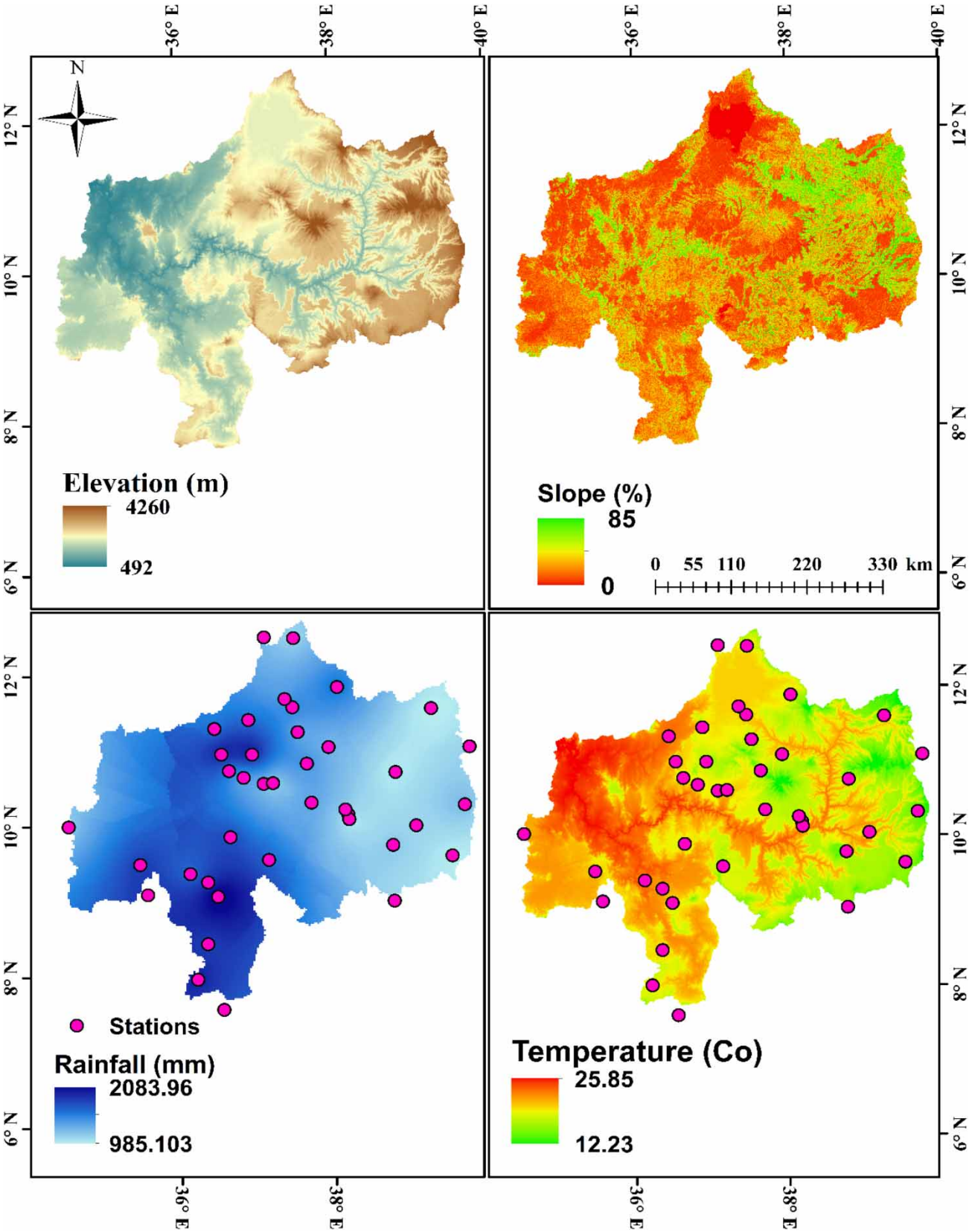


Figure 2 | The physiographic and metrological input data of 12-year mean annual spatial variation (over 2010–2021) in the Upper Blue Nile Basin.

Remote sensing indices

Normalized difference vegetation index

Normalized difference vegetation index (NDVI) is used to quantify vegetation greenness. It is used to understand vegetation cover and assess changes in plant health. NDVI value ranges from −1 to +1, where a high NDVI value indicates the presence of healthy vegetation, while negative values indicate poor and sparse vegetation. The NDVI, which runs from −1 to +1, is also crucial for identifying various land cover types in the study area. A pixel-based NDVI value was computed using Equation (1) as the normalized difference between the Red (Band 4:0.64–0.67 μm) and near infrared (NIR) (Band 5: 0.85–0.88 μm) from the Landsat 8 satellite data (see Figure 3 for a visual representation):

$$\mathrm{NDVI} = \frac{(\mathrm{NIR} - \mathrm{Red})}{(\mathrm{NIR} + \mathrm{Red})}. \tag{1}$$

Soil-adjusted vegetation index

Soil-adjusted vegetation index (SAVI) is a vegetation index that includes a soil brightness correction factor to reduce the effects of soil brightness (Huete 1988). This is often used in arid regions where vegetative cover is low. It is calculated using two bands of Red (Band 4:0.64–0.67 μm) and NIR (Band 5: 0.85–0.88 μm) from the Landsat data (see Figure 3 for a visual representation):

$$\mathrm{SAVI} = \frac{(\mathrm{NIR} - \mathrm{Red})}{(\mathrm{NIR} + \mathrm{Red} + L) \times (1 + L)} \tag{2}$$

where NIR are the pixel values from the NIR, Red are the pixel values from the near red band, and L is the amount of green vegetation cover. The L value is determined by how much green vegetation is present. L is typically equal to 1 in places with no green vegetation, 0.5 in areas with moderate green vegetation, and 0 in areas with very high vegetation coverage. To account for the majority of land cover types, a soil brightness adjustment factor (L) of 0.5 was adopted for this research.

Normalized difference moisture index

Changes in moisture content are analyzed at the landscape element level, especially for soil and vegetation using a vegetation distinguishing index called the normalized difference moisture index (NDMI). NDMI is sensitive to the moisture levels in vegetation. It is used in a variety of applications, including drought monitoring, crop yield estimation, and vegetation health assessment (Wilson & Sader 2002). It is calculated using two bands of NIR (Band 5: 0.85–0.88 μm) and mid-infrared (MIR) (Band 6: 1.57–1.65 μm) from the Landsat 8 satellite data. This index ranges from −1 to +1, where positive values indicate a high moisture level and negative values indicate a low moisture level:

$$\mathrm{NDMI} = \frac{(\mathrm{NIR} - \mathrm{MIR})}{(\mathrm{NIR} + \mathrm{MIR})} \tag{3}$$

The NIR band is sensitive to vegetation reflectance, while the MIR band is sensitive to water and moisture content in vegetation and soil. By subtracting the MIR band from the NIR band and normalizing the result, the NDMI index is able to highlight areas of high moisture content and discriminate them from areas of low moisture content.

Enhanced vegetation index

Enhanced vegetation index (EVI) is a spectral index that measures vegetation density and health using data from satellite imagery. It was developed to improve upon the NDVI, which has limitations in areas with dense vegetation or high levels of atmospheric interference (Huete *et al.* 2002). EVI is calculated from the reflectance values of three spectral bands, typically the Blue (Band 4:0.64–0.67 μm), Red (Band 4:0.64–0.67 μm), and NIR (Band 5: 0.85–0.88 μm) from the Landsat 8 satellite data as in Equation (4):

$$\mathrm{EVI} = 2.5 \times \frac{(\mathrm{NIR} - \mathrm{Red})}{(\mathrm{NIR} + 6 \times \mathrm{Red} - 7.5 \times \mathrm{Blue} + 1)}. \tag{4}$$

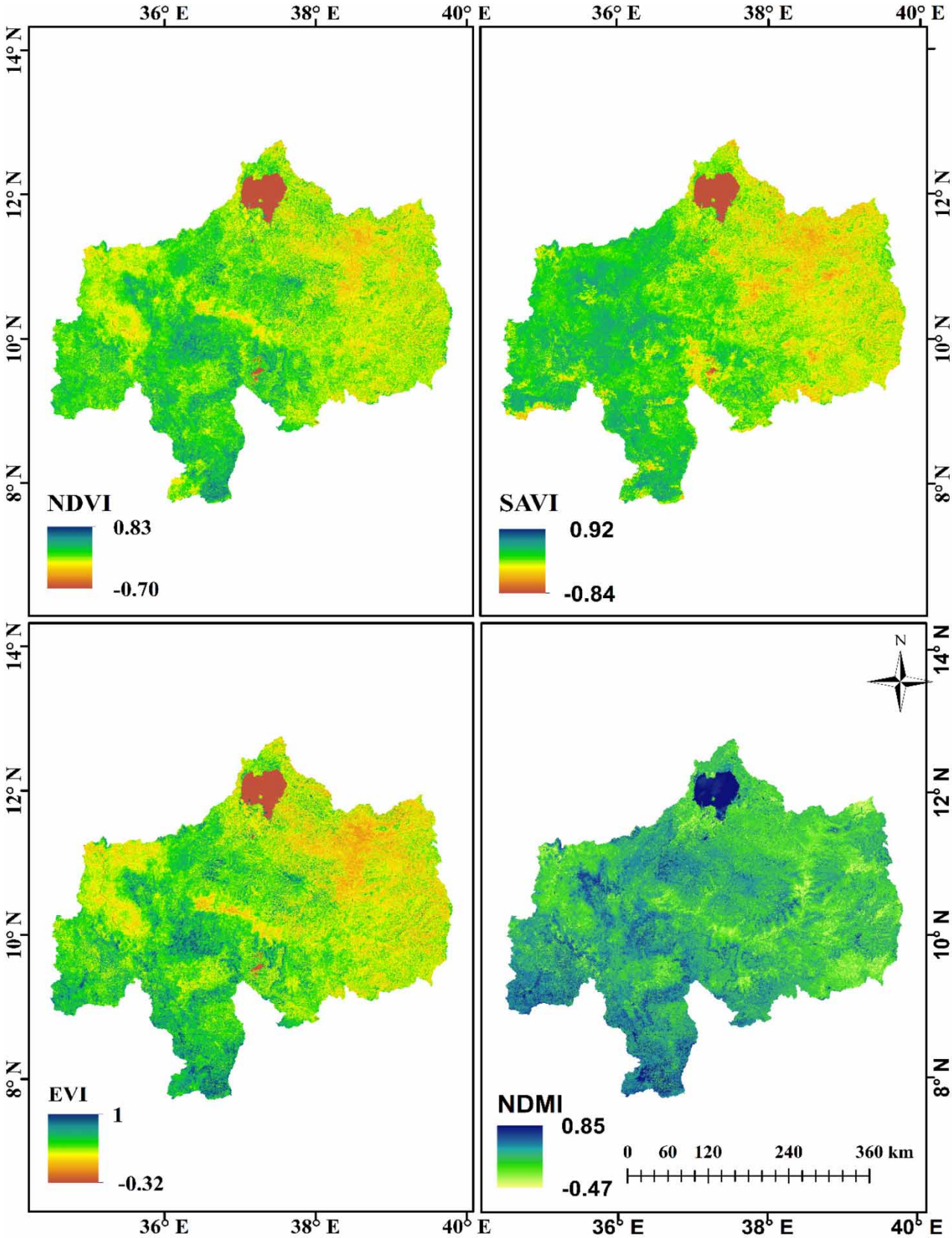


Figure 3 | The remote sensing indices input data long term annual mean spatial variation (over 2010–2021) in the Upper Blue Nile Basin.

The Red band is sensitive to chlorophyll absorption, while the NIR band is sensitive to vegetation structure and biomass. The Blue band helps correct for atmospheric interference and soil background effects. EVI values range from −1 to 1, with higher values indicating greater vegetation density and health (Huete *et al.* 2002). EVI is widely used in a variety of applications, including monitoring crop yields, tracking deforestation, and assessing the impacts of climate change on vegetation (Fensholt & Proud 2012).

Identifying homogenous climate and watersheds

Homogenous climate zones and watersheds were identified using the PCA. PCA converts possibly correlated multiple variables into linearly uncorrelated variables and considerably reconstructs the variability in the original dataset with numerous variables using fewer new variables (Jackson 2005). In this study, the watershed attributes (Table 1) were dimensionally reduced using PCA. However, the concepts and the algorithms used to execute a cluster analysis with PCA are inherently different. Due to our data matrix being huge, performing eigen decomposition to calculate the eigenvalues of the covariance matrix proved challenging and prone to round-off errors. As an alternative, singular value decomposition (SVD) is a reliable computational technique frequently used to compute PCAs of a dataset (Ayalew *et al.* 2022). This involves reducing the less significant basis vectors in the initial SVD matrix. Therefore, the analysis was performed using SVD in R environment. To determine the number of primary components, in this study, a scree plot of the elbow rule was used (Peres-Neto *et al.* 2005). This approach involves locating the 'elbow' shape on the curve and keeping all components until the curve flattens out (Holland 2008; Zambelli 2016). In the process of identifying similar watersheds, it is crucial to take into account different factors that may impact the precision of outcomes, including variation in climate zones. To reduce this influence, initially, the homogeneous climate zones were identified using data from 42 rainfall stations, and subsequently, for each homogeneous climate zone, the homogeneous watersheds were identified.

Once PCA was employed to reduce the dimensionality of the dataset by identifying the most important variables that contribute to the variation in the data, *K*-means clustering was used to group the watersheds based on similarities in these variables. *K*-means clustering, introduced by MacQueen (1967), is a frequently used unsupervised machine learning algorithm to partition a given dataset into a set of *K* clusters, or groups (Kassambara 2017). It classifies watersheds into multiple groups (clusters), such that watersheds within the same cluster are as similar as possible (i.e., high intraclass similarity), whereas watersheds from different clusters are as dissimilar as possible (i.e., low interclass similarity) (Kassambara 2017). In *K*-means clustering, each cluster is represented by its center (i.e., centroid), which corresponds to the mean of points assigned to the cluster. *K*-means clustering was applied to the derived dimensions obtained using the dimensionality reduction techniques to obtain clusters. For determining the optimal number of clusters in a dataset, numerous indices have been published in the literature (Zhou *et al.* 2016; Shahapure & Nicholas 2020; Kassambara 2021). For this study, the most used Elbow and Average Silhouette methods were applied. The Elbow method is based on determining the within-cluster-sum of squared errors (WSS) for various numbers of clusters (K) and choosing the K for which change in WSS first begins to decrease, that is the plot of the WSS with the K value resembles an elbow. The graph then begins to travel nearly parallel to the x-axis from this point forward. The best K value, or the most clusters, is the one that corresponds to this location (Zambelli 2016). In the Average Silhouette method, optimum value of K refers to the point that maximizes the Average Silhouette over a range of possible values for K. Generally, K is the minimum number of clusters that were formed from the 25 watersheds from the homogenous climate zone 1, 15 watersheds from the homogenous climate zone 2, and 36 watersheds from the homogenous climate zone 3. This is iteratively until the distance of the watersheds to each cluster centroids is minimal and the distance between the cluster centroids is maximal. The distance of the line segment between two points was computed using Equation (5) and accounts for correlation within the data (Danielsson 1980; Swain *et al.* 2016; Kanishka & Eldho 2017):

$$D = \sqrt{(x2 - x1)2 + (y2 - y1)^2} \tag{5}$$

where ($x1$, $y1$) are the coordinates of one point, ($x2$, $y2$) are the coordinates of the other point, and D is the distance between two points.

Data normalization

To avoid the impact of measurement units, before performing classification, the raw data should be normalized. This step makes the variables comparable and eliminates any heterogeneous impact on distance measurement. Here, the variables were normalized to the range of [0, 1] using a commonly used standard normalization technique, *Z*-score (Al Shalabi *et al.* 2006) as below:

$$Z = \frac{X - X_{\text{mean}}}{\delta} \tag{6}$$

where Z is the standard normal inputs of watershed attributes, X is the continuous random inputs of watershed attributes, δ is the standard deviation of inputs of watershed attributes, and X_{mean} is the mean of inputs of watershed attributes. Z standardization makes the mean to be 0 and the standard deviation to be 1. A flowchart represented in Figure 4 shows the whole procedure and methodology.

RESULTS AND DISCUSSION

The results of PCA analysis depicted that the 42 stations within the Blue Nile Basin were categorized into three distinct climate regions (Figure 5). Through identifying homogenous climate regions, we were able to reduce the uncertainty of homogenous watershed identification. By doing so, we were able to obtain watershed classification results that were more precise and dependable. This allowed us to achieve more accurate and reliable results in our watershed classification analysis. The identification of homogenous climate regions provides valuable information about the spatiotemporal distribution of rainfall within the basin, which can serve as a useful reference for future investigations into the hydrological mechanisms at work in the region.

According to the PCA presented in Figure 5, three homogenous rainfall regions/clusters can be identified. The vector lines in the figure represent highly correlated weather stations and are considered part of a single, homogeneous climate zone. The points indicate temporal variability at a monthly timescale. Based on their relevance degrees, 11 stations in the upper basin were categorized under homogenous climate region I, 11 stations in the lower basin were categorized under homogenous climate region II, and 20 stations in the central basin were categorized under homogenous climate region III. The first two principal components (PCs) explain over 95% of the total variability in the dataset, and their standard deviation is greater than 1, providing valuable insights into rainfall variability.

Variable reduction for watershed classification

After driving all datasets and computing values for each watershed, the variables used in cluster analysis were reduced using PCA and then normalized using *Z*-score, and 76 watersheds were classified using a normalized watershed variable and *K*-means clustering algorithm, which uses the Average Silhouette and Elbow methods (Figure 6) to determine the number of component analysis (dimension).

In Figure 6, the *x*-axis shows the PCs (dimensions), which are eight in this case, and the proportion of the explained variance (green color) and eigenvalue (blue color) for each PC is displayed on the *y*-axis. From this, we observed that the elbow is located at the second PC (eigenvalue greater than 1). This suggests that the analysis should continue with the first two components. Furthermore, in Figure 6 we can observe that PC1 accounts for 58.3% of total variation of the data and PC2 accounts for 11.9% of total variation of the data. Therefore, the first two PCs may be considered the most significant because they account for more than a threshold of 90% variance of the original data. As a result, we only take into account PC1 and PC2 in our analysis and leave out the other PCs from the analysis. This is how dimensionality reduction works. We efficiently reduced our PCs (dimensions) from 12 to 2 PCs. To reduce variables, we considered the mean contribution of variables for each PC and visualized it using the scree plot (Figure 7).

In Figure 7, the horizontal dashed line represents the expected contribution (average contribution of variables) to each PC since all variables might be contributed evenly. Therefore, variables that have a contribution to that PC are located above this line. In the first component, eight watershed characteristics (EVI, Longitude, NDMI, NDVI, Rainfall, and SAVI) have a significant contribution, while Elevation and Temperature have a low contribution (as shown in Figure 7). In contrast, the second PC is determined by only two variables, namely, Area and Slope, as illustrated in Figure 7. After variable reduction, the cluster analysis was performed using *K*-means clustering algorithm based on these 10 most significant variables.

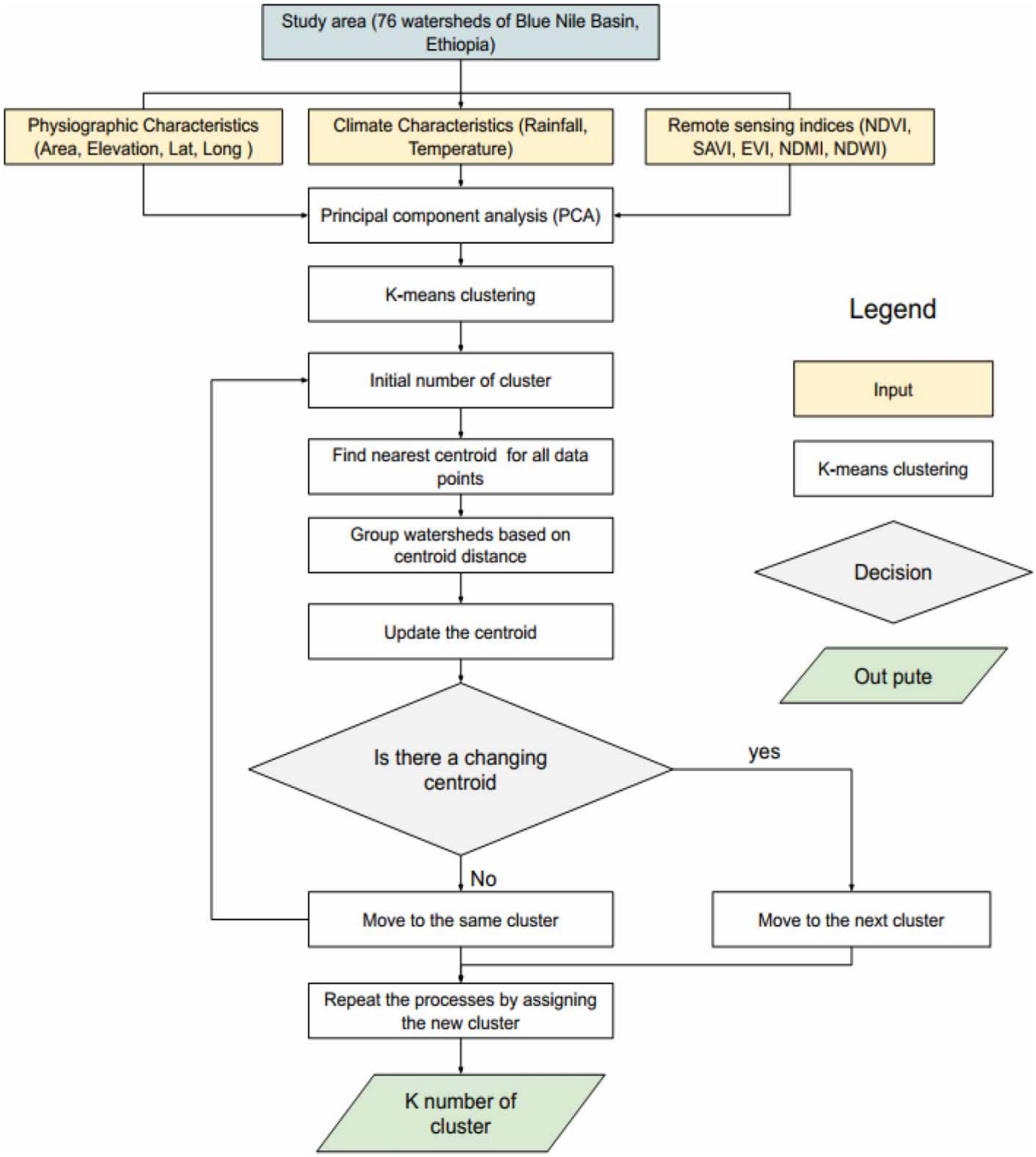


Figure 4 | *K*-means watershed clustering methodology flowchart.

Determine optimal number of clusters

The optimal number of clusters was determined using the Elbow and Average Silhouette methods as shown in Figure 8.

The optimal cluster (K) is determined by selecting the point at which the Average Silhouette is maximized and the total WSS is minimized, across a range of possible values for K. As shown in Figure 8(a), for homogenous climate region I, $K = 4$, the WSS tends to fluctuate more slowly than it does for other Ks. Therefore, $K = 4$ should be a good decision for the number of clusters for region I homogenous climate. For homogenous climate region II, $K = 3$ indicates that the WSS

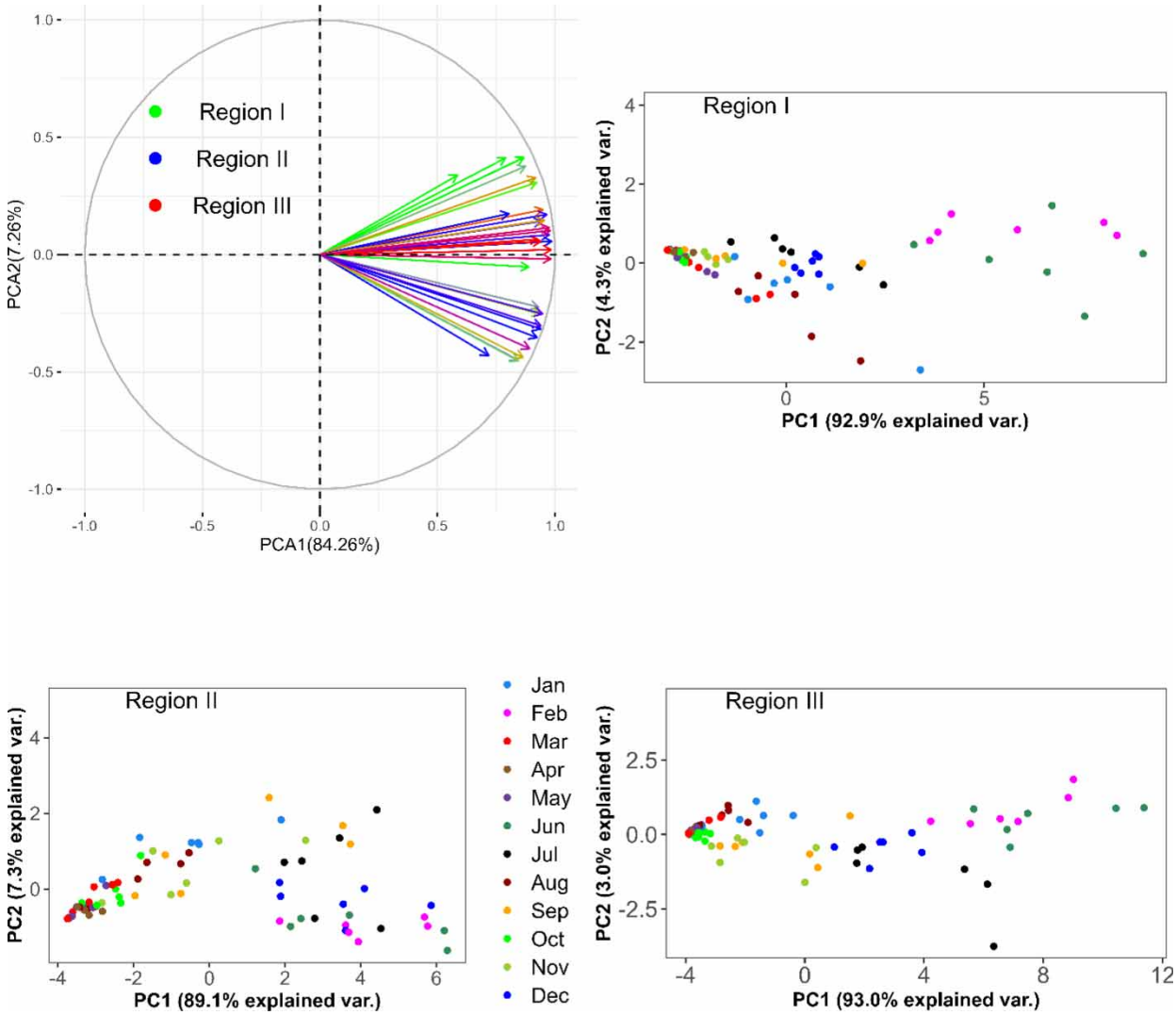


Figure 5 | Homogenous climate region and spatiotemporal variability of rainfall across the basin using a PCA plot of the first two components (PC1 and PC2) accounting for more than 90% of the variance.

tends to fluctuate more slowly than it does for other *K*s. Therefore, $K = 3$ is considered a good decision for the number of clusters for homogenous climate region II. For homogenous climate region III, $K = 2$, the WSS tends to fluctuate more slowly than it does for other *K*s.

K-means clustering result

We determined the optimal number of clusters for classifying watersheds in the Upper Blue Nile Basin using the Elbow and Average Silhouette methods (as shown in Figure 8). Based on our analysis, we classified the watersheds into nine homogeneous groups using the *K*-means classification method. This classification method allowed us to group the watersheds based on their similarities in terms of hydrometeorological and remote sensing data. By doing so, we were able to identify groups of watersheds that have similar hydrological processes and characteristics, which can be valuable for runoff prediction in ungauged watersheds using the regionalization method.

In Table 2, we present the cluster membership of the watersheds in the Blue Nile Basin, which were classified into several clusters based on their similarities in terms of hydrometeorological and remote sensing data. Climate homogeneous region I consists of four homogeneous clusters, namely, cluster 1, cluster 2, cluster 3, and cluster 4, with seven, six, seven, and five

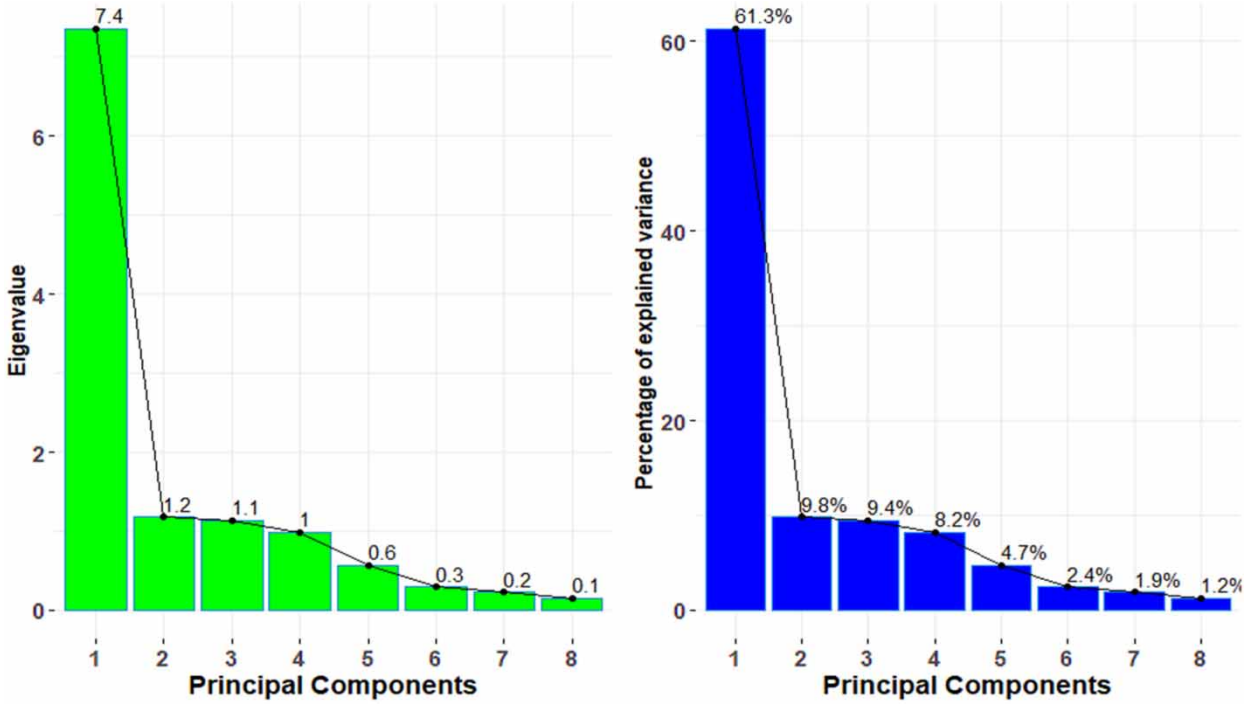


Figure 6 | Optimum number of component analysis using eigenvalue (green) and Elbow method (blue).

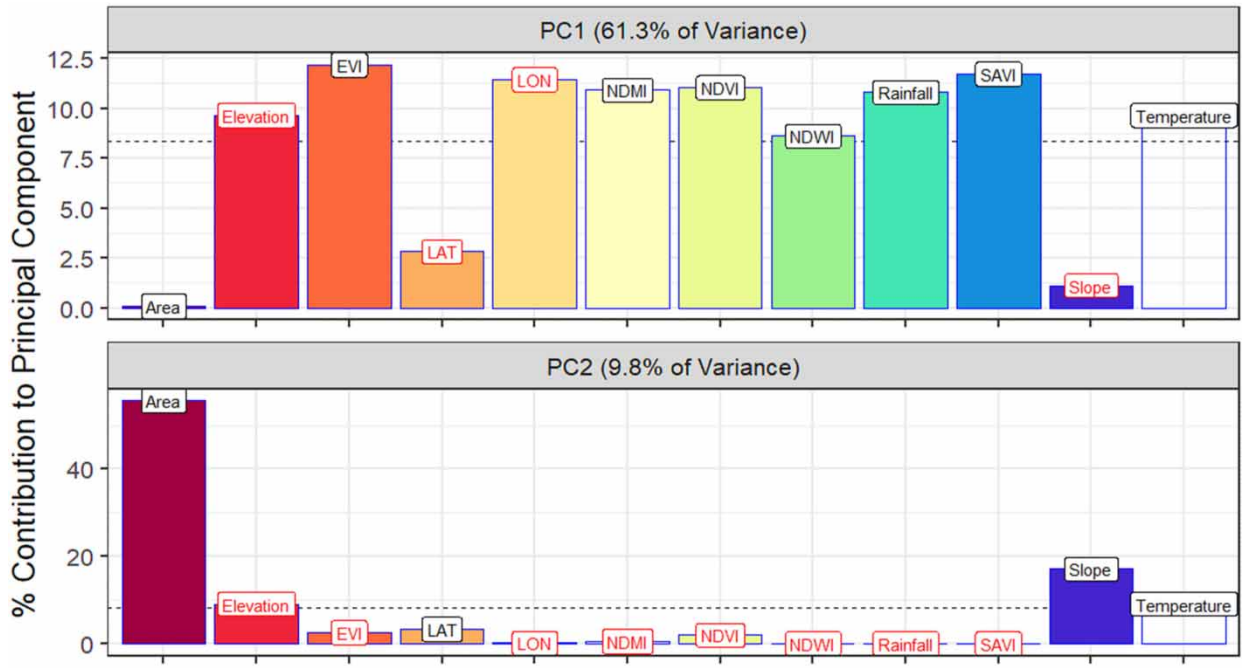


Figure 7 | Plot of variable reduction and grouping of variables into components using the scree plot with the consideration of average contribution of variables.

watersheds, respectively; climate homogeneous region II comprises of three homogeneous clusters, namely, cluster 1, cluster 2, and cluster 3. These clusters consist of seven, four, and four watersheds, respectively; and climate homogeneous region III comprises of two homogeneous clusters, namely, cluster 1 and cluster 2, consisting of 20 and 16 watersheds, respectively. The

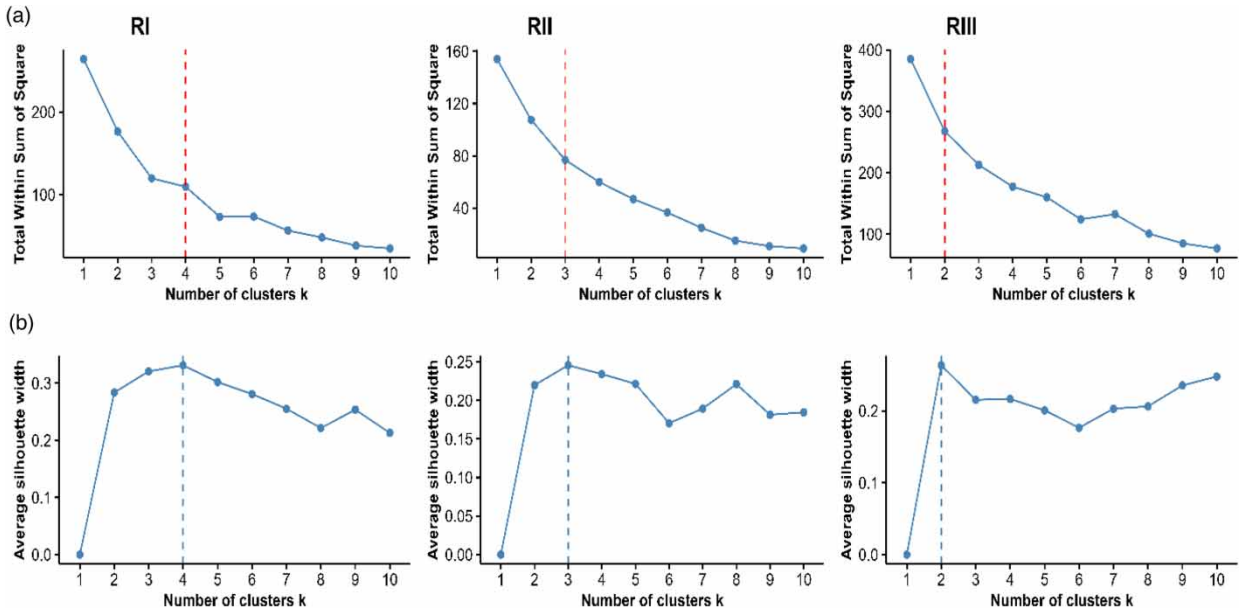


Figure 8 | Optimal cluster for *K*-means clustering algorithm using the Elbow (a) and Average Silhouette (b) methods. RI is climate homogenous region I, RII is climate region II, and RIII is climate region III.

cluster membership information presented in Table 2 provides insight into the grouping of watersheds based on their similarities. This information can be valuable for designing effective water management strategies and decision-making processes for the Upper Blue Nile Basin.

Validation of cluster analysis

The *K*-means clustering method resulted in four clusters in the homogenous climate region I, which were validated against watershed characteristics such as annual rainfall variability, temperature, elevation, slope, NDVI, and NDMI. Cluster 1 watersheds have the lowest elevation (2,020.47–2,512.04 m) and rainfall (1,335.94–1,427.86 mm/year) among all clusters in region I. These watersheds are located downstream in the eastern part of the Blue Nile Basin and have minimal vegetation cover (NDVI range 0.27–0.31), indicating that they are mainly covered by agricultural land. On the other hand, Cluster 2 watersheds are situated in the southeastern portion of the basin and have intermediate elevation (2,666.45–2,769.77 m),

Table 2 | Homogenous watersheds of the Upper Blue Nile Basin

Climate region	Homogenous watersheds			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Region I	Aleltu, Boreda, Desso, Gebreguracha, Jemma, Mechela, and Wenchit	Chacha, Gorfo, Mugher, Roba, Robi-Jida, and Robi-gumero	Beressa, Gerado, Jogola, Kelina, Selgi, Shy, and Wizer	Debis, Guder, Huluka, Tilku Duber, and Tinshu Duber
Region II	Adiya, Angar, Dabana, Dabus, Little Ang, Uke, and Wama	Didessa, Tamsa, Urgessa, and Yebu	Indris, Neshi, Sifa, and Tato	
Region III	Abahim, Andassa, Azuari, Bogena, Chemoga, Chena, Dirma, Gemero, Gumara, Megech, Muga, Ribb, Sedie, Shina, Suha, Teme, Tigdar, Tul, Wenka, and Yeda	Abbay, Birr, Chereka, Dondor, Dura, Fettam, Gelgel Abay, Gilgel Beles, Gudla, Jedeb, Koga, Lah, Leza, Main Beles, Missini, and Temcha		

moderate rainfall (1,456.74–1,476.06 mm/year), and moderate to high vegetation cover (NDVI range 0.3–0.33), indicating that they are primarily covered by forest and vegetation. Cluster 3 watersheds are located in the upland of climate homogenous region I, with high elevation (2,544.39–2,984.89 m), abundant rainfall (1,433.91–1,516.28 mm/year), and excellent vegetation cover (NDVI range 0.28–0.34), suggesting that they are covered by forest and natural vegetation. Cluster 4 watersheds are found in the southern region of the Blue Nile Basin and have intermediate elevation (2,278.68–2,557.78 m), moderate rainfall (1,384.22–1,436.41 mm/year), and moderate to high vegetation cover (NDVI range 0.33–0.4), indicating that they are also mainly covered by forest and vegetation. The spatial distribution of the clusters within region I shows that cluster 1 watersheds are mainly located in the downstream eastern part of the Blue Nile Basin, while cluster 2 watersheds are situated in the southeastern portion of the basin. Cluster 3 watersheds are located in the upland areas of the region, and cluster 4 watersheds are primarily found in the southern region of the basin (Figure 9). The clustering results provide insights into the variability of hydrological conditions across the region, with watersheds in the eastern and southern parts of the basin having relatively lower elevations and moderate rainfall, while those in the upland areas have higher elevations and abundant rainfall (Figure 10). Moreover, the variation in vegetation cover across the clusters highlights the importance of land use and land cover in influencing hydrological processes in the region.

In homogenous climate region II, the *K*-means clustering analysis resulted in three distinct clusters of watersheds. Cluster 1 of this climate zone includes watersheds with elevation ranges from 1,561.55 to 1,784.32 m, the lowest of the three clusters; rainfall ranges from 1,250.12 to 1,291.78 mm/year, and NDVI values range from 0.45 to 0.5. The second cluster includes watersheds with elevation range from 1,823.03 to 1,825.48 meters, rainfall ranges from 1,299.47 to 1,364.40 mm/year, and NDVI values range from 0.38 to 0.56, indicating moderate vegetation cover. Finally, cluster 3 includes watersheds with elevation ranges from 2,172.68 to 2,367.43 m, rainfall ranges from 1,299.02 to 1,400.82 mm/year, and NDVI values range from 0.37 to 0.42, indicating relatively low vegetation cover. These three clusters are spatially distributed in the downstream of the upper Blue Nile Basin. Figure 11 showed that cluster 3 has high rainfall as a result of its watersheds being situated at higher elevations where atmospheric moisture is more likely to condense and precipitate. Nonetheless, the cluster's comparatively low vegetation coverage could imply that the plants are not taking full advantage of the heavy rainfall, leading to erosion and runoff. Multiple factors such as soil type, land use practices, and vegetation type could contribute to this phenomenon. Furthermore, the prevalence of bare land and degraded soil in the area could cause reduced vegetation coverage and soil moisture. To comprehend these trends, more research would be necessary.

In homogeneous climate region III, the watershed classification based on *K*-means clustering method resulted in two clusters. Cluster 1 includes watersheds with elevation ranges from 2,097.66 to 2,727.18 m, rainfall ranges from 1,350.37 to 1,468.09 mm/year, and NDVI ranges from 0.29 to 0.37. These watersheds are mainly located in the eastern part of the Upper Blue Nile Basin. On the other hand, cluster 2 includes watersheds with elevation ranges from 1,471.52 to 2,572.94 m, rainfall ranges from 1,233.28 to 1,439.25 mm/year; and NDVI values range from 0.36 to 0.48 (Figure 12). These watersheds are mainly located in the central part of the Upper Blue Nile Basin (Figure 9). The moderate vegetation cover in cluster 1 may be due to the combined effect of climate and land use, as areas with higher rainfall tend to have more vegetation cover. The watersheds in Cluster 2 have a wider range of elevations, which may result in a more varied hydrological response compared to cluster 1. The higher NDVI values in cluster 2 suggest that these watersheds may have more vegetation cover than cluster 1. The lower NDVI values in cluster 1 are despite higher rainfall and elevation ranges in the presence of bare land and degraded soil. Areas with agricultural land and degraded soil are more susceptible to erosion and runoff, which can lead to loss of topsoil and reduced soil moisture content. These factors can make it more difficult for vegetation to grow and thrive, resulting in lower NDVI values. In contrast, cluster 2 may have a higher NDVI value due to a combination of factors such as more favorable soil conditions, better land use practices, and a wider range of elevations, which can support a greater variety of vegetation types in compared to cluster 1.

The spatial distribution of the clusters within each region provides important insights into the hydrological conditions of the Blue Nile Basin. For instance, the presence of cluster 2 in all three regions indicates that there are areas with moderate elevation, abundant rainfall, and good vegetation cover throughout the basin. This suggests that these areas may play a critical role in the hydrology of the basin, such as in the generation of runoff and the maintenance of water quality. In addition, the spatial distribution of cluster 1 in the eastern part of the basin across all three regions suggests that this area may be more prone to water scarcity and drought conditions, while the presence of cluster 2 in the western part of the basin across all three regions suggests that this area may be more resilient to water scarcity due to the presence of moderate rainfall and vegetation cover. From this, we observed that the *K*-mean clustering algorithm has successfully divided the watershed into nine

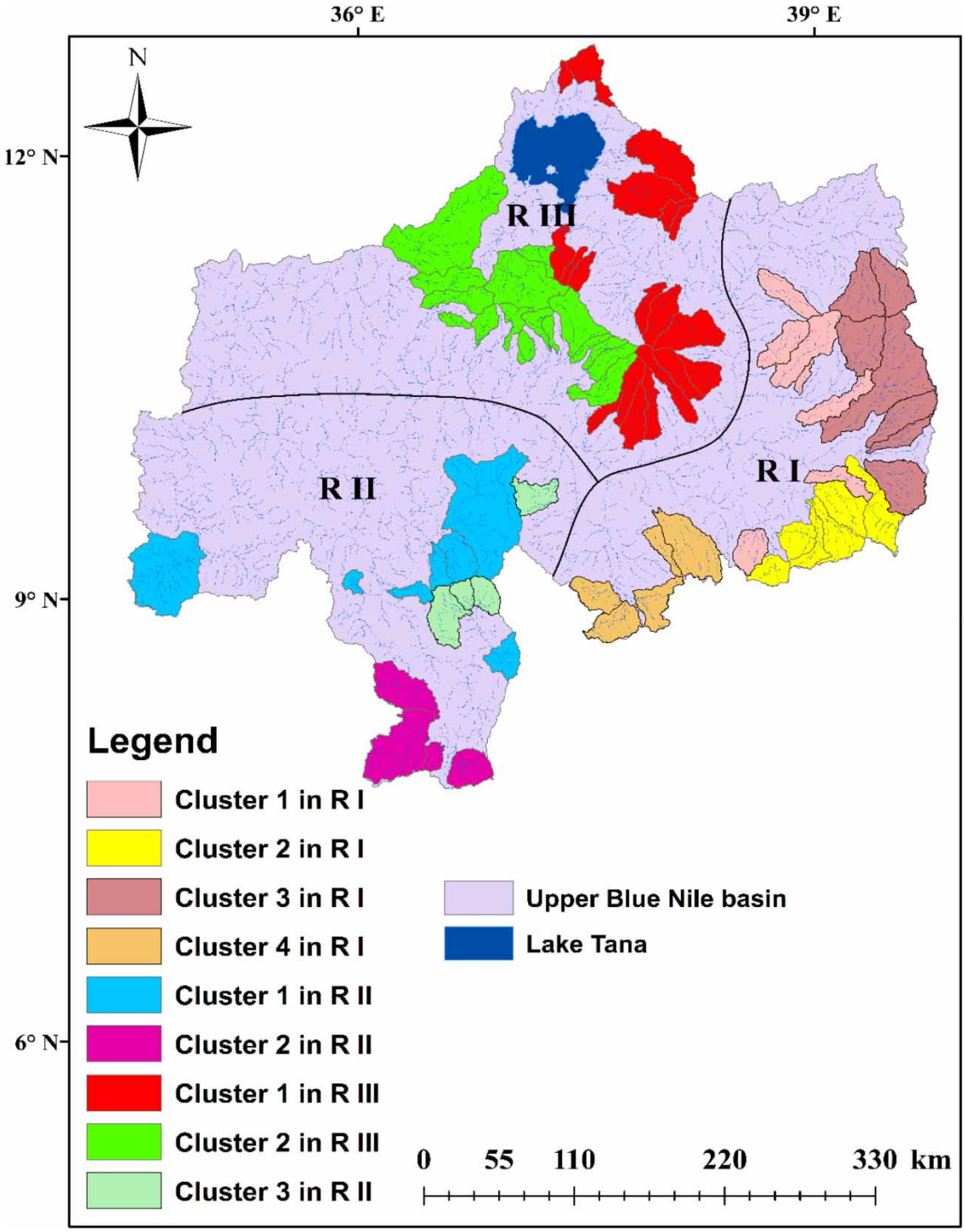


Figure 9 | Map of homogenous watersheds of the Upper Blue Nile Basin within the three homogenous climate regions.

groups that are relatively similar in terms of physiographic and meteorological variability. This is beneficial for better understanding the overall condition of the watersheds, as it allows for more targeted and specific management practices to be implemented in each group based on their particular characteristics. Therefore, watersheds in homogenous climate region I are characterized by steep topography and low vegetation cover; we may want to implement management practices that

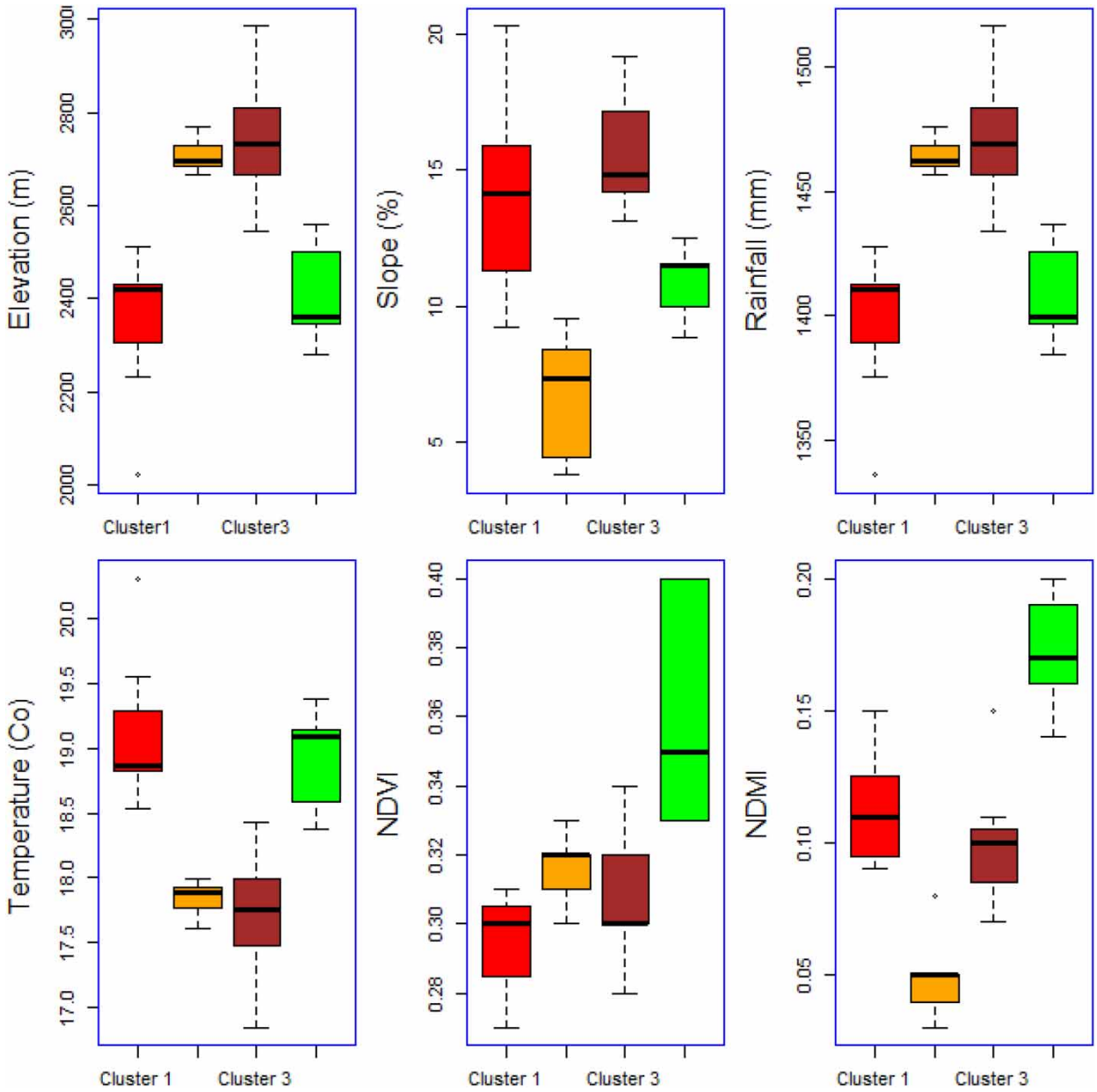


Figure 10 | Spatial distribution of hydrological similarity identifying parameter (mean annual rainfall, temperature, elevation, slope, NDVI, and NDMI) in homogenous climate region I contains four clusters.

focus on controlling erosion and sedimentation; for homogenous climate region II with gentle topography and high vegetation cover, we may want to implement practices that focus on water conservation and increasing soil moisture retention, while for homogenous climate region III with intermediate elevation and medium rainfall, we may want to implement both water conservation practice and practices that focus on controlling erosion and sedimentation. This watershed clustering approach may also have been considered to improve problems of runoff prediction in the ungauged watersheds.

Vegetation dynamics in three homogenous climate regions

The upper Blue Nile Basin exhibits significant spatial variability in NDVI values within the three identified regions (Figure 13). In region II, for example, there are areas of high NDVI values, particularly in the western part of the basin,

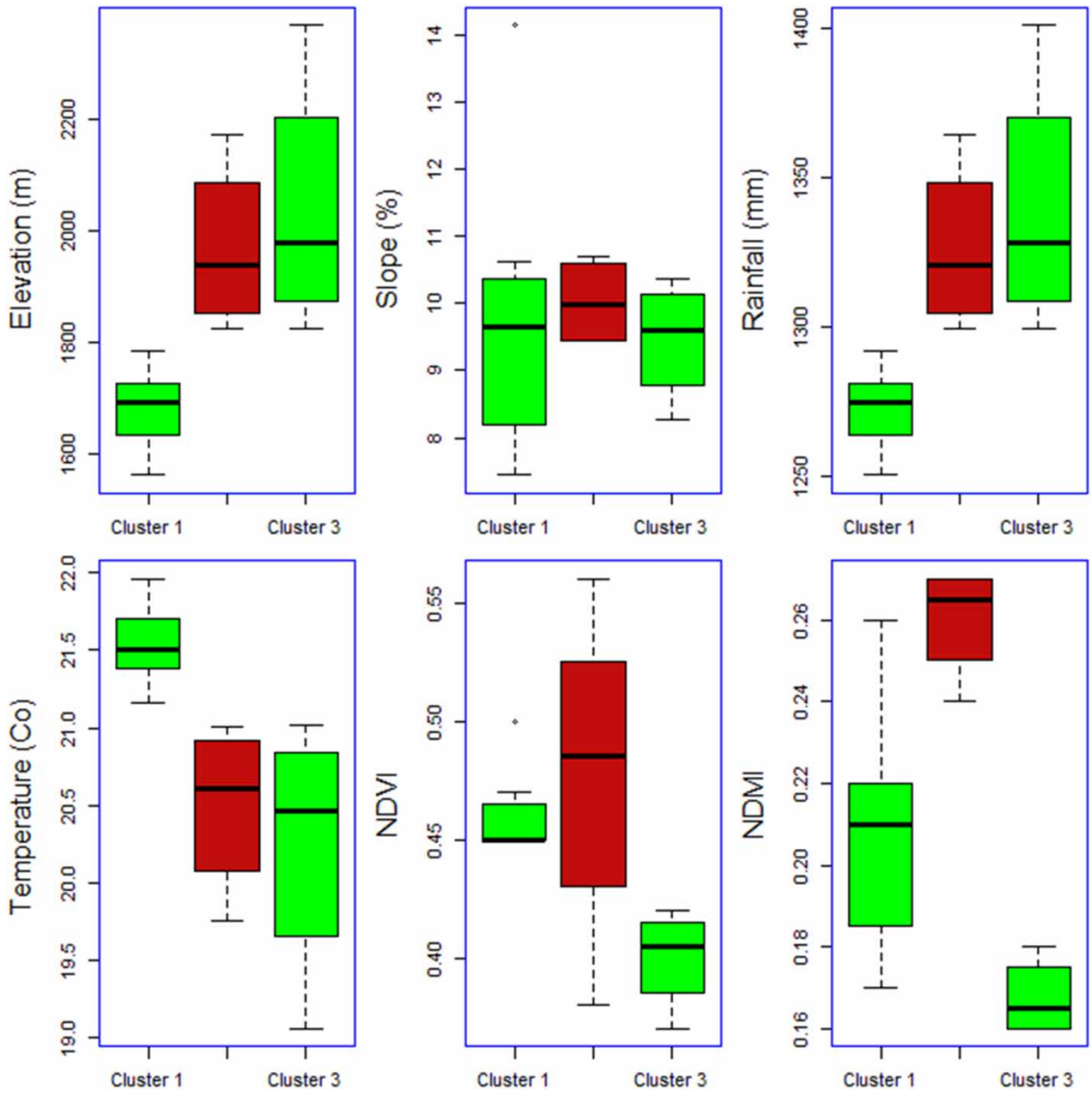


Figure 11 | Spatial distribution of hydrological similarity identifying parameter (mean annual rainfall, temperature, elevation, slope, NDVI, and NDMI) in homogenous climate region II contains three clusters.

where forests and grasslands are dominant. However, there are also areas of low NDVI values, particularly in the eastern and southern parts of the basin, where cropland and grazing land use dominate. In region III, there is also significant spatial variability in NDVI values, with the highest values observed in the irrigated agricultural areas in the central and western parts of the basin. However, there are also areas of lower NDVI values in the eastern and southern parts of the basin, where rain-fed agriculture and grazing land use are dominant. In region I, there is less spatial variability in NDVI values compared to the other two regions, with the lowest values observed in the barren upland areas and the highest values observed in the south-eastern cropland areas. Overall, the spatial variability in NDVI values within the three regions reflects the complex interactions between land use, topography, and environmental factors that influence vegetation growth in the upper Blue Nile Basin. These findings have important implications for sustainable land use management, particularly in the context of climate change and other environmental challenges.

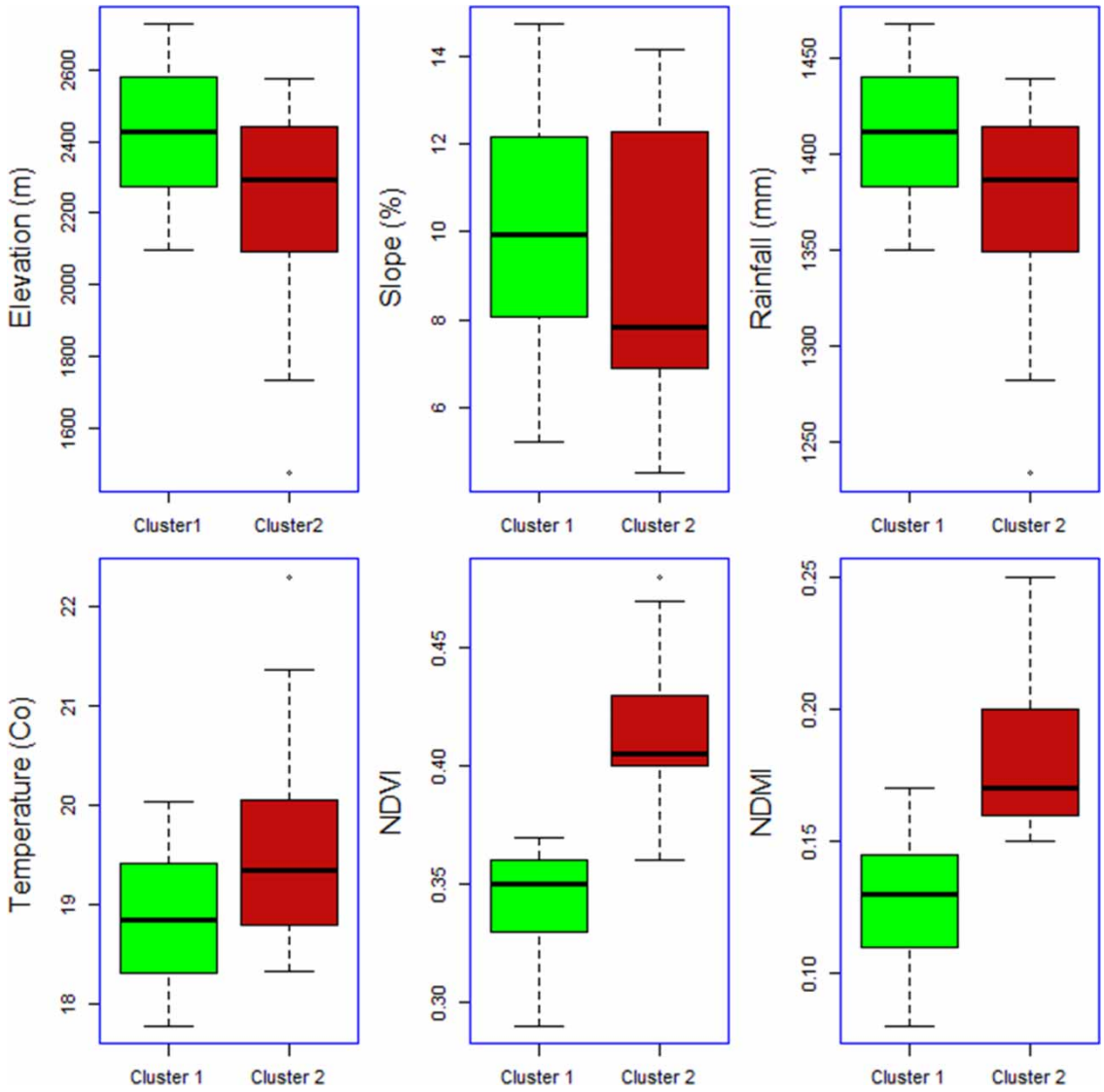


Figure 12 | Spatial distribution of hydrological similarity identifying parameter (mean annual rainfall, temperature, elevation, slope, NDVI, and NDMI) in homogenous climate region III contains two clusters.

NDVI values were also analyzed for the three identified climate regions in the upper Blue Nile Basin for the years 2000, 2010, and 2020 to investigate trends in vegetation cover (Figure 14). In region II, NDVI values showed a relatively stable trend across the study period, with a slight increase from 2000 to 2010 and a slight decrease from 2010 to 2020 (Figure 14). This indicates that vegetation cover in this region has remained relatively stable over the past two decades. In Region III, NDVI values showed a clear increasing trend across the study period (2000–2010) and a more slightly decrease from 2010 to 2020 (Figure 14); this may be due to overgrazing or other environmental stressors. This indicates that agricultural land use in this region has intensified over the past two decades, leading to a decrease in vegetation cover. In region I, NDVI values showed a slight decrease from 2000 to 2010, but a stable trend from 2010 to 2020 (Figure 14). This indicates that vegetation cover in region I showed a stable trend over the past decade than region II and region III. Overall, the trend analysis of NDVI values within the three regions highlights the dynamic nature of land use change in the Blue Nile Basin. While vegetation cover has remained relatively stable in region I, there has been significant intensification of agricultural land use in region III and a decline in vegetation cover in region II. These findings have important implications for sustainable land use management in the basin, particularly in the face of climate change and other environmental challenges.

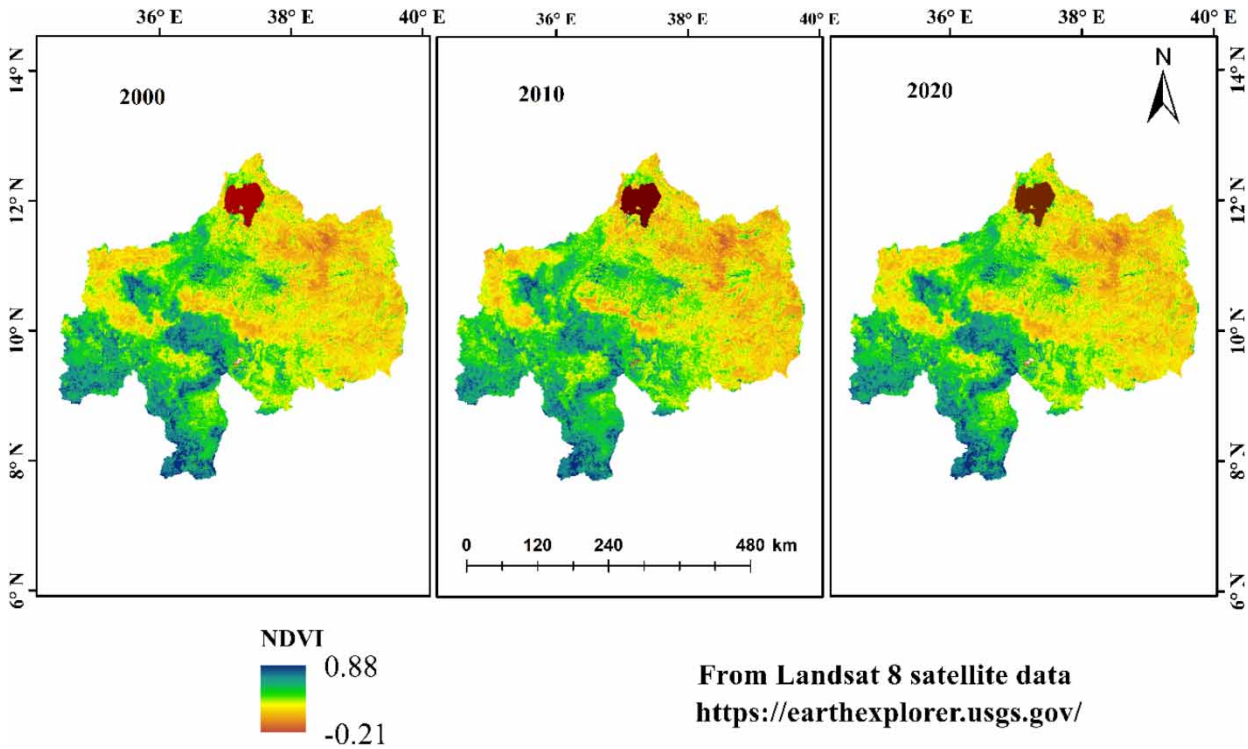


Figure 13 | The Upper Blue Nile Basin spatiotemporal distribution map of NDVI.

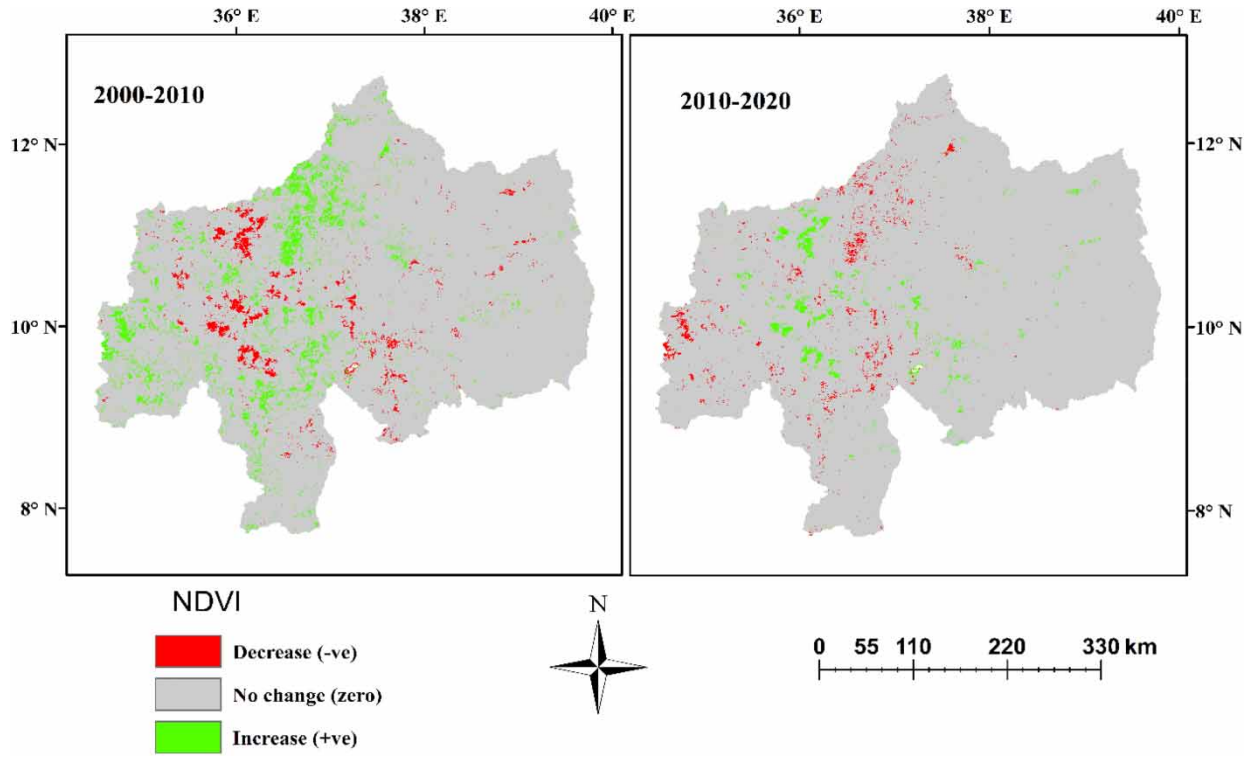


Figure 14 | Vegetation cover dynamics map in the Upper Blue Nile Basin for two time periods from 2000 to 2010 and from 2010 to 2020.

CONCLUSION

In this study, a clustering algorism was used to identify hydrologically homogenous watersheds of the upper Blue Nile Basin, Ethiopia. To achieve the goal, the physiographic parameters (area, longitude, latitude, elevation, and slope), metrological parameters (rainfall and temperature), and remote sensing indices (NDVI, SAVI, EVI, NDMI, and NDWI) were used. For watershed classification, a linear dimensionality reduction technique, PCA followed by *K*-means clustering, was used to classify 76 watersheds of the basin. The number of primary PCs was determined using the plot of the percentage of variance arranged from largest to smallest (scree plot). In the analysis, 10 parameters were taken into account for the first two PCs, which were then used for conducting *K*-means clustering. In addition to the PCA, the Elbow and Average Silhouette methods were employed to determine the optimal value of *K* for *K*-means clustering. After determining the optimal number of regions using the Elbow and Average Silhouette methods, the 76 Upper Blue Nile Basin watersheds of the three homogenous climate regions were also classified into nine watershed clusters using the *K*-means clustering algorithm. The results of the clustering analysis conducted in this study aligned with the existing physiographic and meteorological patterns that are known to exist in the Blue Nile Basin. The upper region of the basin (region I) is distinguished by low vegetation coverage, which can be attributed to the area's high percentage of agricultural land use and dense population. Conversely, the lower region of the basin (region II) is recognized for its relatively flat terrain and greater water resources, which fosters a greater abundance of vegetation and more extensive agricultural practices. The central part of the basin (region III) possesses intermediate features, making it suitable for a mix of land uses. Overall, the findings of this study have important implications for sustainable management of water resources in the Upper Blue Nile Basin and other similar basins. The results highlight the need to consider the spatial variability and heterogeneity of land use and environmental factors in hydrological modeling and decision-making. Future studies could further explore the use of clustering analysis and additional remote sensing indices and hydrological parameters to improve the accuracy and applicability of regionalization models in similar watersheds.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Al Shalabi, L., Shaaban, Z. & Kasasbeh, B. 2006 Data mining: A preprocessing engine. *Journal of Computer Science* **2** (9), 735–739.

Ayalew, A. D., Wagner, P. D., Sahlu, D. & Fohrer, N. 2022 Land use change and climate dynamics in the Rift Valley Lake Basin, Ethiopia. *Environmental Monitoring and Assessment* **194** (11), 791.

Ayele, H. S., Li, M. H., Tung, C. P. & Liu, T. M. 2016 Impact of climate change on runoff in the Gilgel Abbay watershed, the upper Blue Nile Basin, Ethiopia. *Water* **8** (9), 380.

Chaudhary, S. & Pandey, A. C. 2022 PCA driven watershed prioritization based on runoff modeling and drought severity assessment in parts of Koel river basin, Jharkhand (India). *Water Supply* **22** (2), 2034–2054.

Choubin, B., Solaimani, K., Habibnejad Roshan, M. & Malekian, A. 2017 Watershed classification by remote sensing indices: A fuzzy c-means clustering approach. *Journal of Mountain Science* **14** (10), 2053–2063.

Choubin, B., Solaimani, K., Rezanezhad, F., Roshan, M. H., Malekian, A. & Shamshirband, S. 2019 Streamflow regionalization using a similarity approach in ungauged basins: Application of the geo-environmental signatures in the Karkheh River Basin, Iran. *Catena* **182**, 104128.

Danielsson, P. E. 1980 Euclidean distance mapping. *Computer Graphics and Image Processing* **14** (3), 227–248.

Farhan, Y., Al-Shaikh, N. & Ayed, A. 2017 Multivariate statistical analysis of hydro-morphometric parameters of some arid watersheds, Jordan using ASTER DEM and GIS. In: *Applied Morphometry and Watershed Management Using RS, GIS and Multivariate Statistics (Case Studies)*. Scientific Research Publishing, Inc., USA. p. 155.

Farsadnia, F., Kamrood, M. R., Nia, A. M., Modarres, R., Bray, M. T., Han, D. & Sadatinejad, J. 2014 Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. *Journal of Hydrology* **509**, 387–397.

Fensholt, R. & Proud, S. R. 2012 Evaluation of earth observation based global long term vegetation trends – Comparing GIMMS and MODIS global NDVI time series. *Remote Sensing of Environment* **119**, 131–147.

Gebregiorgis, A. S., Moges, S. A. & Awulachew, S. B. 2013 Basin regionalization for the purpose of water resource development in a limited data situation: Case of Blue Nile River Basin, Ethiopia. *Journal of Hydrologic Engineering* **18** (10), 1349–1359.

Holland, S. M. 2008 *Principal Components Analysis (PCA)*. Department of Geology, University of Georgia, Athens, GA. 30602, 2501.

Huete, A. R. 1988 A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment* **25** (3), 295–309.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X. & Ferreira, L. G. 2002 Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment* **83** (1–2), 195–213.

Jackson, J. E. 2005 *A User's Guide to Principal Components*. John Wiley & Sons.

Kanishka, G. & Eldho, T. I. 2017 Watershed classification using isomap technique and hydrometeorological attributes. *Journal of Hydrologic Engineering* **22** (10), 04017040.

Kanishka, G. & Eldho, T. I. 2020 Streamflow estimation in ungauged basins using watershed classification and regionalization techniques. *Journal of Earth System Science* **129**, 1–18.

Kassambara, A. 2017 Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning, Vol. 1. Sthda. Available from: https://xsliulab.github.io/Workshop/2021/week10/r-cluster-book.pdf.

Kassambara, A. 2021 Determining the Optimal Number of Clusters: 3 Must Know Methods. Cluster Validation Essentials. Available from: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/ (accessed: May 2022)..

Kim, U. & Kaluarachchi, J. J. 2008 Application of parameter estimation and regionalization methodologies to ungauged basins of the Upper Blue Nile River Basin, Ethiopia. *Journal of Hydrology* **362** (1–2), 39–56.

Kunnath-Poovakka, A. & Eldho, T. 2018 *Catchment Classification in Data-Scarce Regions Using a Linear Classification Technique*. (unpublished). Available from: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=Hu2QLL8AAAAJ&sortby=pubdate&citation_for_view=Hu2QLL8AAAAJ:YsMSGLbcyi4C.

Lehner, B., Verdin, K. & Jarvis, A. 2008 New global hydrography derived from spaceborne elevation data. *Eos, Transactions, American Geophysical Union* **89** (10), 93–94.

MacQueen, J. 1967 Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1, No. 14, pp. 281–297.

Mosavi, A., Golshan, M., Choubin, B., Ziegler, A. D., Sigaroodi, S. K., Zhang, F. & Dineva, A. A. 2021 Fuzzy clustering and distributed model for streamflow estimation in ungauged watersheds. *Scientific Reports* **11** (1), 8243.

Nawaz, N. R., Bellerby, T., Sayed, M. & Elshamy, M. 2010 Blue Nile runoff sensitivity to climate change. *Open Hydrology* **4**, 137–151.

Palcon, A. X. 2021 Classification of watersheds in Occidental Mindoro and Oriental Mindoro using principal component analysis and k-means clustering. *Ecosystems and Development Journal* **11** (1 and 2), 21–31.

Pallard, B., Castellarin, A. & Montanari, A. 2008 A look at the links between drainage density and flood statistics. *Hydrology & Earth System Sciences Discussions* **5**, 5.

Pascucci, S., Carfora, M. F., Palombo, A., Pignatti, S., Casa, R., Pepe, M. & Castaldi, F. 2018 A comparison between standard and functional clustering methodologies: Application to agricultural fields for yield pattern assessment. *Remote Sensing* **10** (4), 585.

Peres-Neto, P. R., Jackson, D. A. & Somers, K. M. 2005 How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis* **49** (4), 974–997.

Razavi, T. & Coulibaly, P. 2013 Streamflow prediction in ungauged basins: Review of regionalization methods. *Journal of Hydrologic Engineering* **18** (8), 958–975.

Sardooi, E. R., Azareh, A., Choubin, B., Barkhori, S., Singh, V. P. & Shamshirband, S. 2019 Applying the remotely sensed data to identify homogeneous regions of watersheds using a pixel-based classification approach. *Applied Geography* **111**, 102071.

Shahapure, K. R. & Nicholas, C. 2020 Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 747–748.

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., Liang, X., Mcdonnell, J., Mendiondo, E. & O'connell, P. 2003 IAHS decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal* **48**, 857–880.

Swain, J. B., Sahoo, M. M. & Patra, K. C. 2016 Homogeneous region determination using linear and nonlinear techniques. *Physical Geography* **37** (5), 361–384.

Tigabu, T. 2020 *Water Resources in Lake Tana Basin: Analysis of Hydrological Time Series Data and Impact of Climate Change with Emphasis on Groundwater, Upper Blue Nile Basin, Ethiopia*. *Doctoral Dissertation*.

Tigabu, T. B., Hörmann, G. & Fohrer, N. 2015 Water Resources in Lake Tana Basin: Statistical Analysis of Hydrological and Meteorological Time Series. In *AGU Fall Meeting Abstracts*, Vol. 2015, p. GC51E-1130.

Waterbury, J. 2008 *The Nile Basin: National Determinants of Collective Action*. Yale University Press, Michigan, USA.

Wilson, E. H. & Sader, S. A. 2002 Detection of forest harvest type using multiple dates of Landsat TM imagery. *Remote Sensing of Environment* **80** (3), 385–396.

Wolfe, J. D., Shook, K. R., Spence, C. & Whitfield, C. J. 2019 A watershed classification approach that looks beyond hydrology: Application to a semi-arid, agricultural region in Canada. *Hydrology and Earth System Sciences* **23** (9), 3945–3967.

Zambelli, A. E. 2016 A data-driven approach to estimating the number of clusters in hierarchical clustering. [version 1; peer review: 2 approved, 1 approved with reservations]. F1000Research 2016, 5(ISCB Comm J):2809 (https://doi.org/10.12688/f1000research.10103.1).

Zhou, S., Xu, Z. & Liu, F. 2016 Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. *IEEE Transactions on Neural Networks and Learning Systems* **28** (12), 3007–3017.

First received 15 June 2023; accepted in revised form 23 January 2024. Available online 13 February 2024