J. Earth Syst. Sci. (2020) 129:186
https://doi.org/10.1007/s12040-020-01451-8

Streamflow estimation in ungauged basins using watershed classification and regionalization techniques

Ganvir Kanishka and T I Eldho*

Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India.
**Corresponding author. e-mail: eldho@civil.iitb.ac.in*

MS received 18 January 2020; revised 22 April 2020; accepted 2 June 2020; published online 5 September 2020

Classifying watersheds prior to regionalization improves streamflow predictions in ungauged basin. Present study aims to assess the ability of combining watershed classification using dimensionality reduction techniques with regionalization methods for reliable streamflow prediction using soil and water assessment tool (SWAT). Isomap and principal component analysis (PCA) are applied to watershed attributes of 30 watersheds from Godavari river basin in India to classify them. The best classification technique is determined by calculating similarity index (SI). The results showed that Isomap is better at classifying hydrologically similar watersheds than PCA with an average SI value of 0.448. The regionalization methods such as global mean, inverse distance weighted (IDW) and physical similarity were applied to transfer the parameters from watersheds of best watershed classification group to the pseudo-ungauged watersheds, using SWAT model. The present study suggests that classifying watersheds with Isomap and regionalization using physical similarity improves the efficiency of streamflow estimation in ungauged basins.

Keywords. Ungauged basins; regionalization; watershed classification; Isomap; similarity index; SWAT model.

1. Introduction

The assessment of water resources in a watershed is very important for appropriate planning and management, decision making for policy makers, water allocation for agricultural, industrial and domestic sectors, design of bridges, dams, etc., and disaster management. For this, continuous streamflow records for the watersheds are essential. Although hydrologists across the globe have achieved success to a great extent in modelling the rainfall-runoff relationship, estimation of streamflow in ungauged watersheds still remains a crucial problem. The major problem is unavailability of streamflow series data for calibration and validation of the rainfall-runoff models. Estimation of model parameters for an ungauged watershed from a donor gauged watershed is called as regionalization (Blöschl and Sivapalan 1995). Regionalization methods are widely used for the estimation of streamflow in ungauged basins. The past decade saw a paradigm shift in regionalization methods under 'Decade on Predictions in Ungauged Basins (PUB): 2003–2012' established by the International Association of Hydrological Sciences. The advances made in the field of regionalization suggests that performance of regionalization methods is different in different regions and there is no universal regionalization method that is applicable to all the regions (Oudin *et al.* 2008; Samuel *et al.* 2011).

Many of the river basins in the world are either ungauged or poorly gauged (Sivapalan *et al.* 2003;

Young 2006; Blöschl *et al.* 2013). In sub-continental country like India, the situation is even worse where limited streamflow data is available for most of the river basins. In spite of such a diversified hydroclimatic conditions and poorly gauged river basins, the problem of estimating streamflow in ungauged basin is less addressed. Singh *et al.* (2001) developed regional flow duration models for over 1,200 watersheds from 13 states in Himalayan region, India for planning micro-hydropower projects. Using these models, flows of the desired level of dependability can be estimated. Rees *et al.* (2002) developed a regionalized flow estimation method to estimate flow duration curves using catchment characteristics in snow-fed and rain zones of Nepal and Himachal Pradesh (India). Rees *et al.* (2004) developed a hydrological model for estimating dry season flows in ungauged catchments of northern India based on recession curve behaviour using calibration set of 26 catchments. The regional models predicted the volume of water with an average of 8% error during the recession period. Recently, Swain and Patra (2017) estimated streamflow in 32 Indian catchments using regional flow duration curve by applying area – index, inverse distance weighted (IDW), kriging and stepwise regression. Kriging and IDW performed better than the remaining two techniques with average Nash Sutcliffe Efficiency (NSE) of 0.6.

He *et al.* (2011) and Razavi and Coulibaly (2013a) reviewed broad spectrum of regionalization studies where streamflow is estimated by implementing various regionalization methods, rainfall-runoff models and for different study areas in the past decade. The most common regionalization techniques include spatial proximity, kriging, inverse distance weighted (IDW), physical similarity, regression methods, global average and hydrological similarity. Streamflow prediction in either gauged or ungauged catchments are carried out using distributed physically-based models (e.g., BTOPMC [Block wise use of TOPMODEL with Muskingum–Cunge flow routing method], Mike 11 NAM, MIKE-SHE), conceptual and semi-distributed models (e.g., HBV [Hydrologiska Byrans Vattenbalansavdelning model], MAC-HBV [McMaster-Hydrologiska Byrans Vattenbalansavdelning model], SimHyd, IHACRES [identification of unit hydrographs and component flows from rainfall, evapotranspiration and streamflow], GR4J [Génie Rural à 4 paramètres Journalier], AWBM [Australian Water Balance Model], SWAT [Soil and Water Assessment Tool]) and data-driven models (ANNs [Artificial Neural Networks], ARMA [Auto Regressive Moving Average], MLR [Multiple Linear Regression]) (Razavi and Coulibaly 2013a). Due to high level of uncertainty associated with the physically distributed models, conceptual or semi-distributed models are extensively used with combination of regression techniques in most of the studies (Razavi and Coulibaly 2013a). Heuvelmans *et al.* (2004) evaluated the transferability of soil and water assessment tool (SWAT) model parameters on 25 Belgian catchments to predict daily streamflow in ungauged catchments. The results indicated spatial and temporal loss when parameters are transferred in time and space. Gitau and Chaubey (2010) estimated streamflow in seven catchments of Arkansas using SWAT with global average and regression methods. Satisfactory range of model evaluation criteria suggests suitability of SWAT model for streamflow prediction in ungauged catchments. Sellami *et al.* (2014) addressed the issues related to uncertainty of models for streamflow prediction using SWAT combined with spatial proximity. Athira *et al.* (2016) derived regionalized probability distribution parameters for SWAT model. They proposed a methodology for developing ensemble hydrological model simulations in ungauged basin.

Several studies examined the potential of classifying watersheds into homogenous regions for improving the predictions in ungauged basins. These studies focused on identifying homogenous groups of watersheds by applying different linear and non-linear dimensionality reduction techniques on watershed characteristics (e.g., Nathan and McMahon 1990; Chiang *et al.* 2002a, b; Di Prinzio *et al.* 2011; Kileshye *et al.* 2012; Razavi and Coulibaly 2013b). Ganvir and Eldho (2017) investigated the potential of Isomap, a non-linear dimensionality reduction technique, for classifying the watersheds of Mississippi river basin. Isomap performed better than PCA in classifying the watersheds into homogenous groups for regionalization.

In this study, regionalization methods were applied to pre-identified homogenous regions of watersheds to access the potential improvement in estimating streamflow in ungauged basins. Global average, inverse distance weighted (IDW) and physical similarity were used to regionalize SWAT model parameters of some of the watersheds of a major Indian river basin. The main objective is to investigate the potential improvements by applying regionalization methods to groups of systematically classified watersheds obtained using dimensionality reduction techniques with SWAT model and suggest best option.

2. Study area and data

The study area comprises of 30 watersheds of the Godavari river basin in India (figure 1). Godavari river basin is the second largest river basin in India after Ganga river basin with an area of 312,812 km^2. Godavari river flows through the states of Maharashtra, Chhattisgarh, Madhya Pradesh, Odisha, Karnataka and Telangana and finally drains into Bay of Bengal. The 30 watersheds chosen from the basin are from the upper reaches of the basin that do not have inflow from neighbouring watersheds and have minimal flow regulations to avoid the uncertainty caused by the inflow from neighbouring catchments. The study area lies between 73°24′–83°4′ east longitudes and 16°19′–22°34′ north latitudes. India experiences winter season from January to February, pre-monsoon season from March to May, southwest monsoon season from June to September and post-monsoon season from October to December. Rainfall in India primarily occurs during 4 months of monsoon season. Mean annual rainfall in the 30 selected watersheds ranges from 600 to 1700 mm. The temperature is as high as 45°C during pre-monsoon season and falls up to 4°C during winter season. Basic description of the 30 watersheds such as latitude, longitude, gauging station, mean elevation, mean slope and mean annual precipitation are listed in table 1.

The National Aeronautics and Space Administraion (NASA) Shuttle Radar Topographic Mission (SRTM) 90 m Digital Elevation Model's (DEM) with a resolution of 90 m at the equator is used for the delineation of watersheds (Jarvis *et al.* 2008). Area of watersheds and mean slope of watersheds were extracted using ArcMap Geographical Information System (GIS) software from the DEM. Land use data for the present study area is extracted from land use dataset of United States Geological Survey (USGS) (http://www.landcover.usgs.gov/) which is 0.5 km Moderate Resolution Imaging Spectrometer (MODIS)-based Global Land Cover Climatology. The 17 land use

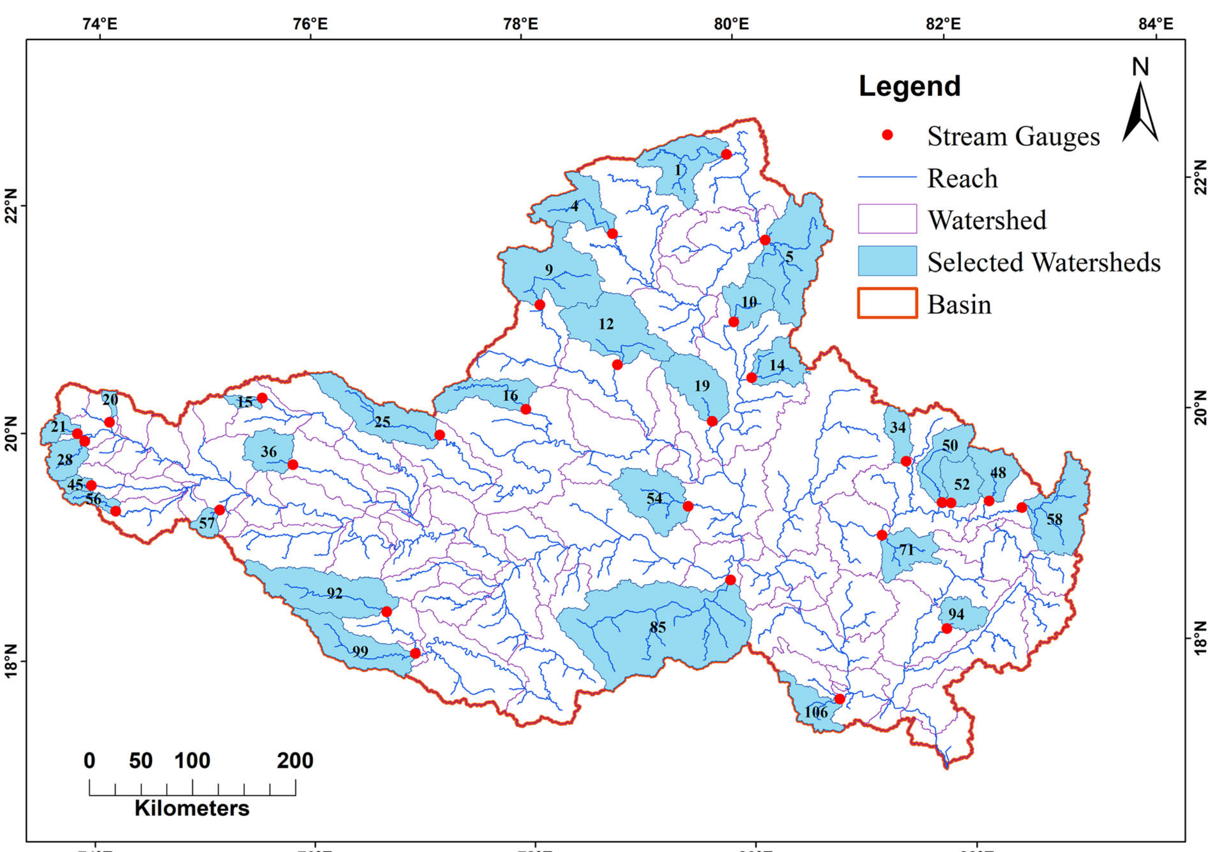

Figure 1. Godavari River Basin with selected 30 watersheds for study.

Table 1. *General details of the 30 selected watersheds and their watershed numbers used in the study.*

Sl. no.	Watershed no.	Centroid Latitude	Centroid Longitude	Name of gauging station	Watershed area (km^2)	Mean elevation (m)	Mean slope (%)	Mean annual precipitation (mm)
1	1	22.39	79.91	Keolari	3061.53	610.19	4.27	1160.86
2	4	21.72	78.82	Ramakoma	2498.97	694.41	6.48	1179.24
3	5	21.62	80.25	Rajegaon	5424.66	426.18	5.79	1449.40
4	9	21.11	78.13	Bishnur	4939.16	469.74	4.45	949.95
5	10	20.92	79.93	Salebardi	1725.83	291.94	3.76	1459.15
6	12	20.57	78.84	Nandgaon	4335.65	314.35	3.20	1089.75
7	14	20.42	80.08	Wairagarh	1757.48	299.99	5.19	1548.13
8	15	20.32	75.53	Chinchkhed	331.89	709.11	6.97	826.76
9	16	20.20	77.98	Mangrul	2164.59	395.44	2.83	938.90
10	19	20.05	79.71	Rajoli	2669.10	239.03	2.28	1394.79
11	20	20.11	74.11	Niphad	225.34	689.46	3.69	632.28
12	21	20.00	73.81	Nasik	631.82	678.65	6.50	1470.03
13	25	19.98	77.18	Karnergaon	3540.81	562.67	2.11	881.08
14	28	19.94	73.88	Samangaon	1210.24	653.61	9.66	1757.15
15	34	19.65	81.48	Cherribeda	1028.40	649.40	4.76	1509.69
16	36	19.73	75.81	Golapangri	1594.27	577.76	4.07	735.00
17	45	19.55	73.95	Mahaldevi	408.77	810.15	17.80	1418.37
18	48	19.27	82.23	Kosarguda	1683.69	616.94	3.37	1410.82
19	50	19.28	81.80	Ambabal	1970.23	621.61	3.05	1425.74
20	52	19.27	81.88	Sonarpal	1488.30	603.98	3.08	1422.33
21	54	19.31	79.46	Bhatpalli	3128.08	382.12	8.22	1115.71
22	56	19.32	74.17	Ghargaon	604.66	817.51	15.95	984.44
23	57	19.34	75.13	Samangaon M	548.42	565.11	4.37	591.15
24	58	19.20	82.53	Nowrangpur	3640.69	800.47	15.01	1469.66
25	71	19.01	81.24	Tunmnar	1860.39	535.79	8.78	1424.76
26	85	18.66	79.83	Somanpally	13141.65	351.70	2.44	933.35
27	92	18.43	76.67	Bhatkheda	4686.32	675.77	1.67	804.66
28	94	18.18	81.80	Potteru	1161.60	243.02	9.35	1553.67
29	99	18.07	76.93	Aurad (Sh)	3138.54	637.23	1.72	887.51
30	106	17.59	80.80	Sangam	1619.46	187.40	6.27	1067.93

classes in the data were re-classified into five land use classes, viz., urban, agriculture, forest, pasture and water. The aridity index (AI) and potential evapotranspiration (PET) data were procured from the Consultative Group on International Agricultural Research–Consortium for Spatial Information (http://www.cgiar-csi.org) (Trabucco and Zomer 2009). The temperature and rainfall data were extracted from the Worldclim-global climate data site (http://www.worldclim.org) (Hijmans *et al.* 2005). The watershed attributes listed in table 2 were derived from the various data sources mentioned above and used for the classification of watersheds.

Table 2. *Selected watershed attributes for classification of watersheds.*

Catchment attribute	Unit
Area	km^2
Slope	%
Area covered with forest	km^2
Area covered with pasture	km^2
Area covered with agriculture	km^2
Urban area	km^2
Mean temperature	Degree celsius
Mean precipitation	mm
Mean potential evapotranspiration	mm/year
Aridity index	–
Drainage density	km/km^2

The India Meteorological Department (IMD) $0.5° \times 0.5°$ resolution gridded data for temperature and $0.25° \times 0.25°$ resolution gridded data for rainfall was collected for the period of 11 years from 1995 to 2005 for SWAT calibration and validation. Climate data such as relative humidity, solar radiation and wind data was obtained from global weather data for SWAT database (https://globalweather.tamu.edu). The daily streamflow was obtained from India-Water Resources Information System (WRIS) database (http://www.india-wris.nrsc.gov.in/wris.html). Global soil texture and gradient data were obtained from the Food and Agricultural Organization (FAO) website (http://www.fao.org). SRTM's DEM mentioned above was used for delineation of watersheds in SWAT.

3. Methodology

In this study, linear and non-linear dimensionality reduction techniques were applied to watershed attributes of 30 watersheds of Godavari river basin. The lower dimensionality components were used to classify the watersheds into different groups. The watersheds were also classified using the streamflow indices which were used as reference classification. The classification based on dimensionality reduction techniques were compared to reference classification. Best classification technique was decided based on similarity index. Three regionalization techniques were evaluated to transfer SWAT parameters within the group of best classification technique. A flow chart depicting the proposed methodology is presented in figure 2 and is briefly discussed in the following sections.

3.1 *Reference classification*

A hydrologically similar group (hydrologically homogeneous region) is defined as a group of drainage basins whose hydrologic responses are similar. Three of the six streamflow indices suggested by Sawicz *et al.* (2011), viz., slope of flow duration curve, runoff ratio and streamflow elasticity were used to obtain hydrologically homogenous regions (groups). The streamflow indices data was then normalized to zero mean and unit variance for each index (Z score normalized) and classified into six groups using K-means clustering algorithm (MacQueen 1967). Davies Bouldin index (Davies and Bouldin 1979) was used to determine the optimum number of clusters. The groups of reference classification were made homogenous using the homogeneity and discordance test by Hosking and Wallis (1997). The homogenous groups of reference classification were used to compare the classification results obtained using the Isomap, PCA and watershed attributes techniques.

3.2 *Isomap*

Isomap is a nonlinear dimensionality reduction technique based on classical multidimensional scaling (CMDS) (Tenenbaum *et al.* 2000). It has shown higher efficiency in representing variance of dataset compared to PCA with fewer components. It consists of three steps in which distances between each point is calculated using the k-nearest neighbours in the first step. In the second step, the geodesic distances between all pair of points is determined based on the interpoint distances calculated in the first step. In the final step, CMDS is applied to the geodesic distances to project the nonlinear structure of dataset into low dimensional space. In this way, Isomap preserves geodesic distance using the non-Euclidean metric and retains the nonlinear features in the data (Tenenbaum *et al.* 2000). The application of Isomap for watershed classification is explained in details in Ganvir and Eldho (2017).

3.3 *Principal component analysis (PCA)*

PCA is a linear dimensionality reduction technique in which high dimensional data is projected into lower dimensions called principal components (PCs) (Jackson 1991). It does this by orthogonally transforming the data of possibly correlated variables into linearly uncorrelated variables. All the PCs are orthogonal to each other and such that first PC is in the direction to capture maximum variance of data, second PC is orthogonal to first and captures the next maximum variance in the data, and so on. First few PCs account for maximum variance of the data.

3.4 *Soil and water assessment tool (SWAT) model*

SWAT model is developed by the United States Department of Agriculture–Agricultural Research Service (USDA–ARS) to simulate the effects of various land management and climatic scenarios on hydrologic and water quality response in large complex watersheds with varying soils, land use and management practices (Arnold *et al.* 1998; Neitsch *et al.* 2011). It is a semi-distributed model, which works on daily time step. The model divides watershed into sub-watersheds and further into

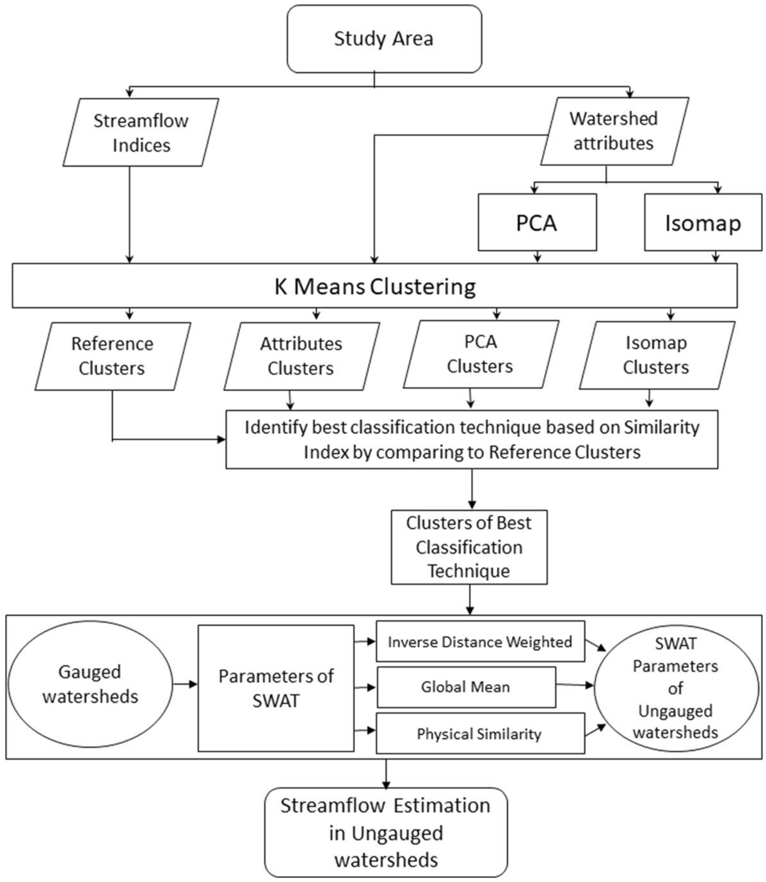

Figure 2. Flowchart of methodology used for streamflow estimation in ungauged basin.

hydrologic response units (HRUs) based on the combination of land use, soil type and slope. The SWAT inputs include data of weather, soils, groundwater, channel, plant water use, plant growth, soil chemistry and water quality parameters. SWAT outputs can be simulated at various spatial scales, i.e., from hydrologic response unit (HRU) level to watershed (Abbaspour *et al.* 2007). The SWAT model uses soil conservation service (SCS) curve number or the Green-Ampt infiltration equation to estimate runoff based on the temporal resolution of input rainfall data. Curve numbers are recalculated daily, based on soil water content on that day. Peak runoff rate predictions are based on a modification of the rational formula.

The driving force of SWAT model is the hydrologic cycle and water balance (Arnold *et al.* 1998) which is explained below

$$SW_t = SW_i + \sum_{i=1}^{t} \left(R_{day} - Q_{surf} - E_a - W_{seep} - Q_{gw}\right), \tag{1}$$

where SW_t is final soil water content (mm); SW_i is initial soil water content on day i (mm); R_{day} is amount of precipitation on day i (mm); Q_{surf} is amount of surface runoff on day i (mm); E_a is amount of evapotranspiration on day i (mm); W_{seep} is amount of water entering the vadose zone from the soil profile on day i (mm); Q_{gw} is amount of return flow on day i (mm).

3.4.1 *Model calibration, validation and sensitivity analysis*

In this study, SWAT-CUP Sequential Uncertainty Fitting Version 2 (SUFI-2) algorithms are applied for calibration, validation, sensitivity and uncertainty analysis of SWAT model parameters. SUFI-2 uncertainty accounts for all sources of uncertainties such as uncertainty in driving variables (e.g., rainfall data, temperature and land use), conceptual model, parameters, and measured data (e.g., surface runoff). In SUFI-2 algorithm, two measures are used for defining model uncertainty namely P-factor and R-factor. The P-factor, the percentage of observed data bracketed by 95% prediction uncertainty (95PPU), is calculated at 2.5% and 97.5% levels of the cumulative distribution of output variables through Latin hypercube sampling method. The R-factor is the average thickness of the 95PPU band divided by the standard deviation of the measured data. Theoretically, a P-factor of 1 and R-factor of 0 indicate that the simulation exactly corresponds to the measured data. For streamflow, a value of P-factor > 0.7 or 0.75 and R-factor around 1 is considered acceptable. Latin hypercube sampling based on one factor at a time (LH-OAT) method which is incorporated within SUFI-2 in SWAT-CUP is used to identify the important parameters that have significance in the model simulation.

The sensitivity analysis evaluates the rate of change in the output of a model with respect to changes in model input. In this study, sensitivity analysis in SUFI-2; one-at-a-time sensitivity analysis and global sensitivity analysis are used for sensitivity analysis. One-at-a-time sensitivity shows the sensitivity of a variable to the changes in a single parameter while other parameters are kept constant at some value. Global sensitivity analysis shows the change in the objective function due to the change in each parameter. Both the sensitivities are measured using two statistical indicators, *p*-stat and *t*-stat (Abbaspour 2015). The absolute values of these indicators help to determine the most sensitive parameters. Lower the absolute value of *p*-stat and higher the value of *t*-stat, more sensitive is the SWAT model parameter.

3.4.2 *Evaluating the performance of SWAT model*

There are different statistical measures available for evaluating the model performance in SWAT-CUP. Model performance was evaluated using three performance evaluation indices, viz., coefficient of determination (R^2), Nash–Sutcliff efficiency and percentage bias (PBIAS) (Moriasi *et al.* 2007). Coefficient of determination (R^2) describes the degree of collinearity between simulated and measured data. The Nash–Sutcliffe efficiency (NSE) is a normalized statistic that determines the relative magnitude of the residual variance (‘noise’) compared to the measured data variance (‘information’) (Nash and Sutcliffe 1970). Percent bias (PBIAS) measures the average tendency of the simulated data to be larger or smaller than their observed counterparts. The equations used for R^2, NSE and PBIAS are given below.

$$R^2 = \left(\frac{\sum_{i=1}^{N} (Y_i - \bar{Y}) \left(\hat{Y}_i - \bar{\hat{Y}} \right)}{\sqrt{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} \sqrt{\sum_{i=1}^{N} \left(\hat{Y}_i - \bar{\hat{Y}} \right)^2}} \right)^2, \quad (2)$$

$$\mathrm{NSE} = 1 - \left[\frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} \right], \quad (3)$$

$$\mathrm{PBIAS} = \left[\frac{\sum_{i=1}^{N} (Y_i - \bar{Y}) * 100}{\sum_{i=1}^{N} Y_i} \right], \quad (4)$$

where Y_i and $\hat{Y}_i$ are the ith observed and predicted streamflow; $\bar{Y}$ and $\bar{\hat{Y}}$ are the mean observed and simulated streamflow, respectively, and N is the total number of observations.

3.5 *Regionalization techniques*

In this study, following regionalization techniques are used.

3.5.1 *Global average*

This is a simple spatial proximity technique in which a parameter value of the ungauged watershed is calculated as the mean of all the donor watersheds in the group (Merz and Blöschl 2004; Parajka *et al.* 2005; Samuel *et al.* 2011).

3.5.2 *Inverse distance weighted (IDW)*

IDW is an interpolation technique based on the inverse spatial distance between watershed centroids. The spatial distance between the watersheds is calculated using the latitude and longitude of the watersheds’ centroids. The IDW equation (Shepard 1968) used in this study to estimate model parameters in ungauged watershed is:

$$P_j = \sum_{i=1}^{n} W_i p_i, \quad (5)$$

where n is the number of gauged watersheds; p_i is the model parameter of gauged watersheds, P_j is the model parameter of an ungauged watershed and W_i is the weight function of each gauged watershed and is calculated as follows:

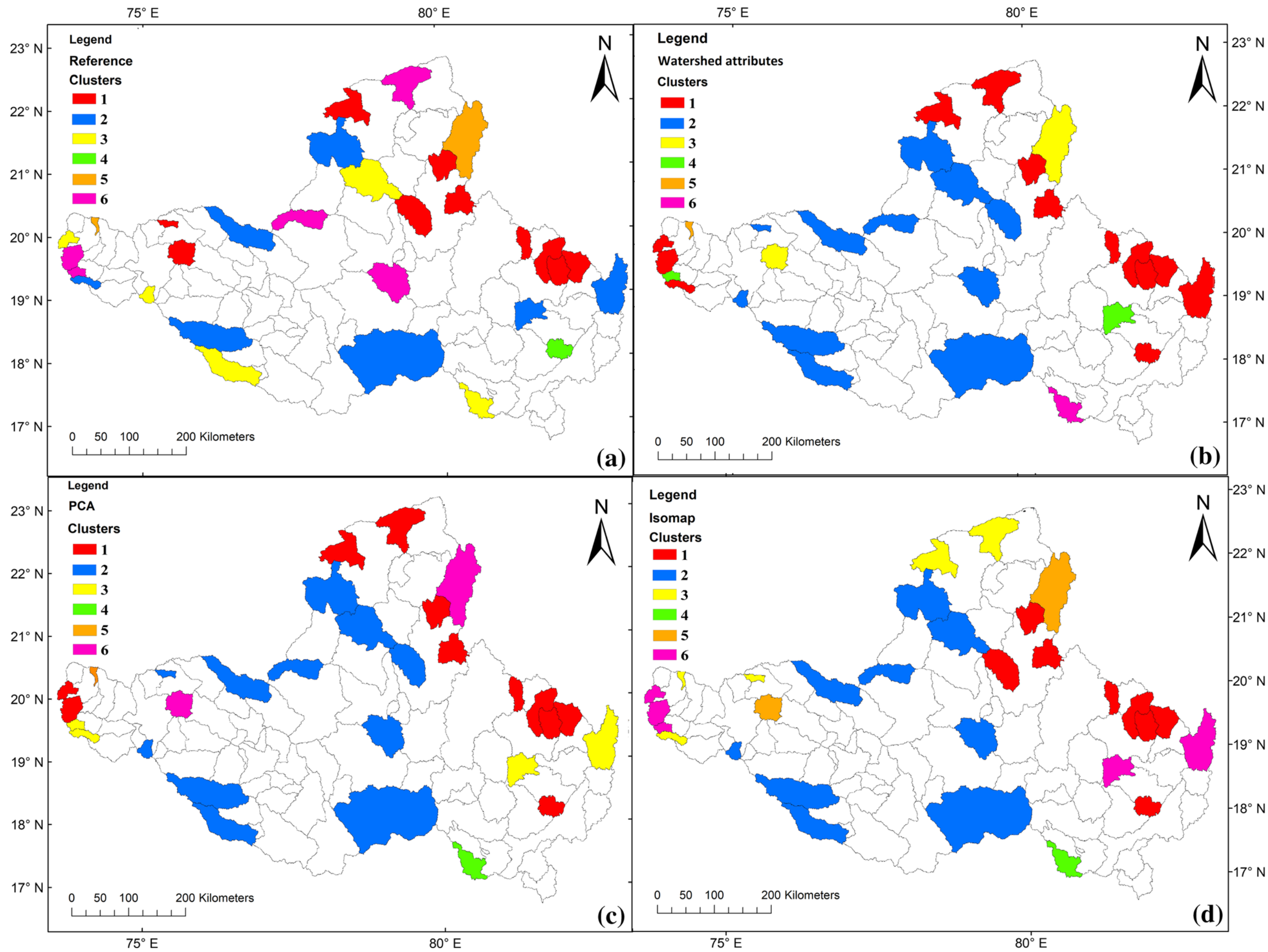

Figure 3. Classification of watersheds based on **(a)** reference groups, **(b)** watershed attributes, **(c)** PCA, and **(d)** isomap.

$$W_i = \frac{\left(d_i^{-2}\right)}{\sum_{i=1}^{n}\left(d_i^{-2}\right)}, \tag{6}$$

where d_i is distance from the centroid of the gauged watersheds to the centroid of the ungauged watershed. In the selection of gauged watersheds, additional criteria such as the NSE value can be considered; however, the number of available gauged watersheds can be a constraint.

3.5.3 *Physical similarity*

In physical similarity method, model parameters from donor watersheds are transferred to the target catchment if they are physically similar to each other (McIntyre *et al.* 2005). Watersheds are said to be physically similar if the catchment attributes that have causative linkage with hydrological behaviour are similar. Physical similarity between two watersheds can be calculated as:

$$d_{t,i} = \sqrt{\sum_{i=1,I} w_i \left(\frac{x_{t,i} - x_{d,i}}{\sigma_{Xi}}\right)^2}, \tag{7}$$

where $x_{t,i}$ and $x_{d,i}$ are the value of each catchment descriptor i ($i = 1, \ldots, I$) for the target and donor catchment, respectively, w_i is the weight associated with the i catchment descriptor, and σ_{Xi} is the standard deviation of the descriptor across the entire set of catchments under study. The smaller the distance, $d_{t,i}$ between the target and donor catchment, the more similar they are. When a single descriptor is considered, w_i for other descriptors is set to zero. When all descriptors are considered to be equally important, w_i is set to one. Once similar catchments are identified, the entire set of model parameters calibrated on the donor catchments can be transferred from the closest donor to the target catchment. Ranked proximity

Table 3. *Number of watersheds in clusters of classification techniques.*

Cluster	Streamflow indices	Watershed attributes	PCA	Isomap
1	10	13	11	9
2	6	11	11	8
3	6	2	4	5
4	5	2	2	5
5	2	1	1	2
6	1	1	1	1

is used if the catchment attributes are many and have different units. In the present study, all the attributes (listed in table 2) were given equal weightages and ranked proximity was calculated for donor watersheds. The model parameters for target watershed were estimated from donor watersheds by giving weightages based on ranked proximities.

4. Results and discussion

4.1 *Classification*

Thirty watersheds of Godavari were classified using (a) streamflow indices, (b) watershed attributes, (c) applying PCA to watershed attributes and (d) applying Isomap to watershed attributes (figure 3). Classifications obtained using (b), (c) and (d) were compared to (a) (reference classification) to decide the best classification technique. Table 3 shows the number of watersheds in each group of the classification techniques.

4.1.1 *Classification using streamflow indices*

Runoff ratio, slope of flow duration curve and streamflow elasticity for the 30 watersheds are calculated using the daily streamflow data from the year 1995–2005. Table 4 shows the summary statistics of the runoff signatures. Runoff ratio is negatively correlated with slope of flow duration curve and streamflow elasticity. All the three streamflow indices show negligible correlation and hence were used for classification. Davies Bouldin's index was least for six number of clusters thus the normalized streamflow indices were classified into six groups using K-means clustering algorithm. The three largest groups contain 22 watersheds out of 30 (table 3). The groups were tested for homogeneity using the discordancy measure. Only one watershed was found discordant with one group which was moved to other groups and discordancy measure was calculated again for all the groups till all the groups were homogenous. The classification of watersheds based on streamflow indices with homogenous groups is used as reference classification.

4.1.2 *Classification using watershed attributes*

Eleven watershed attributes among the spatial, climatic and physiographic parameters that define the hydrological processes were chosen for classification (table 2). The choice was based on earlier studies and availability of data. The attributes of 30 watersheds were normalized and classified into six groups using K-means clustering algorithm.

4.1.3 *Classification using PCA*

PCA was applied to the normalized watershed attributes of the 30 watersheds to obtain principal components. Figure 4 shows the percentage of variance accounted for as a function of the number of dimensions retained using PCA. It was observed that first six reduced components accounted for 90% of variance of the original data. K-means

Table 4. *Summary statistics of runoff signatures for 30 watersheds of Godavari basin.*

	Runoff ratio	Slope of flow duration curve	Streamflow elasticity
Minimum	0.17	1.35	−0.24
1st quartile	0.57	4.98	0.00
Mean	0.86	6.08	0.07
Median	0.73	6.04	0.03
3rd quartile	1.09	7.38	0.11
Maximum	2.79	9.16	0.57
Std. dev.	0.54	1.85	0.13

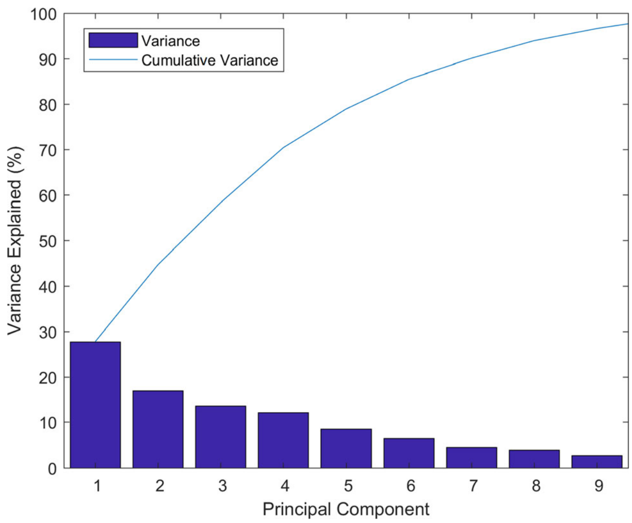

Figure 4. Percentage of total variance explained by reduced components of PCA.

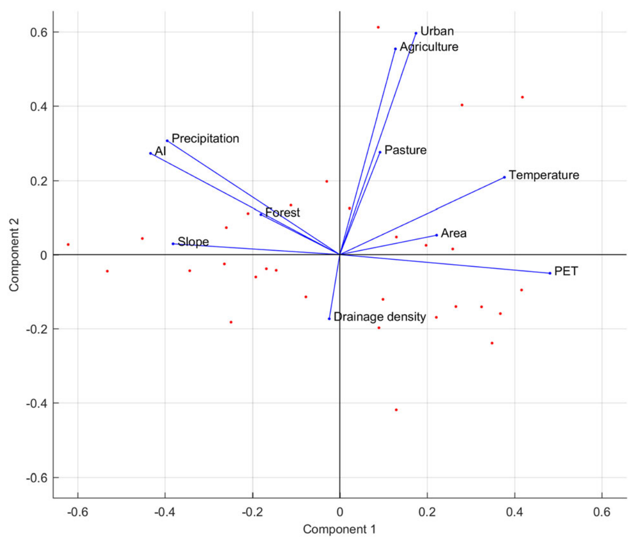

Figure 5. PCA loading plots for the two first principal components.

clustering was carried out using these six components for forming six groups. To visualize the analysis results, both the principal component coefficients (loadings) for each attribute and the principal component scores for each observation (watershed) are presented in a single plot in figure 5. The length of vectors for the attributes shows the weightage given to them in the first two principal components. PET, temperature, urban land use and agricultural land use have more weightages in the first two principal components. The direction of the vectors accounts for the correlation among them. Aridity index and precipitation are highly correlated.

4.1.4 *Classification using Isomap*

The interpoint distance matrix between all the pairs of watershed was calculated based on the selected attributes (dissimilarity matrix). Using this distance matrix, a neighbourhood graph was constructed by selecting neighbors-k from 3 to 12. Then CMDS was performed on the dissimilarity matrix obtained using the k neighbours to get the low dimensional Isomap components. First four components accounted for 90% variance and were retained for classification of watersheds (figure 6). The classification of watersheds obtained using $k = 8$ showed best results when compared to the reference classification. Spearman's rank correlation coefficient between the dimensions and the variables was used to visualize the contribution of attributes in the reduced components of Isomap (figure 7). Aridity index, urban land use and temperature show higher values in first, second and third dimensions, respectively. PET shows lowest values in first and second dimensions and drainage density in the third dimension for Spearman's rank correlation coefficient.

4.1.5 *Evaluation of classification techniques*

Similarity index (SI) is calculated by comparing largest cluster from a classification technique to the largest cluster of reference classification (Ssegane *et al.* 2012). The similarity index obtained in such a manner would give SI for complete cluster. However, if the classification contains clusters with same number of watersheds, it creates ambiguity in selection of group for comparison. Thus, Ganvir and Eldho (2017) developed a new method to

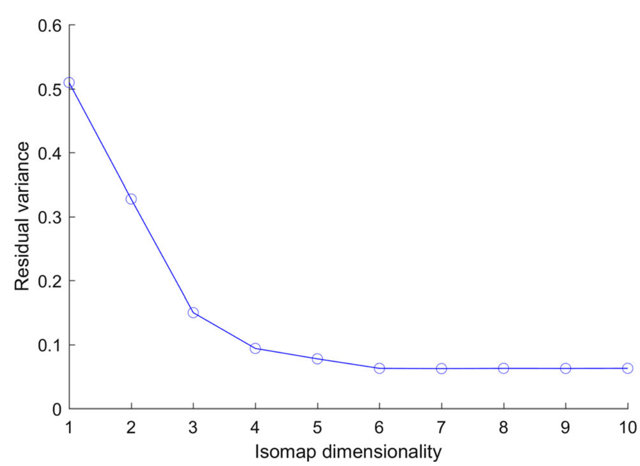

Figure 6. Percentage of residual variane explained by Isomap components.

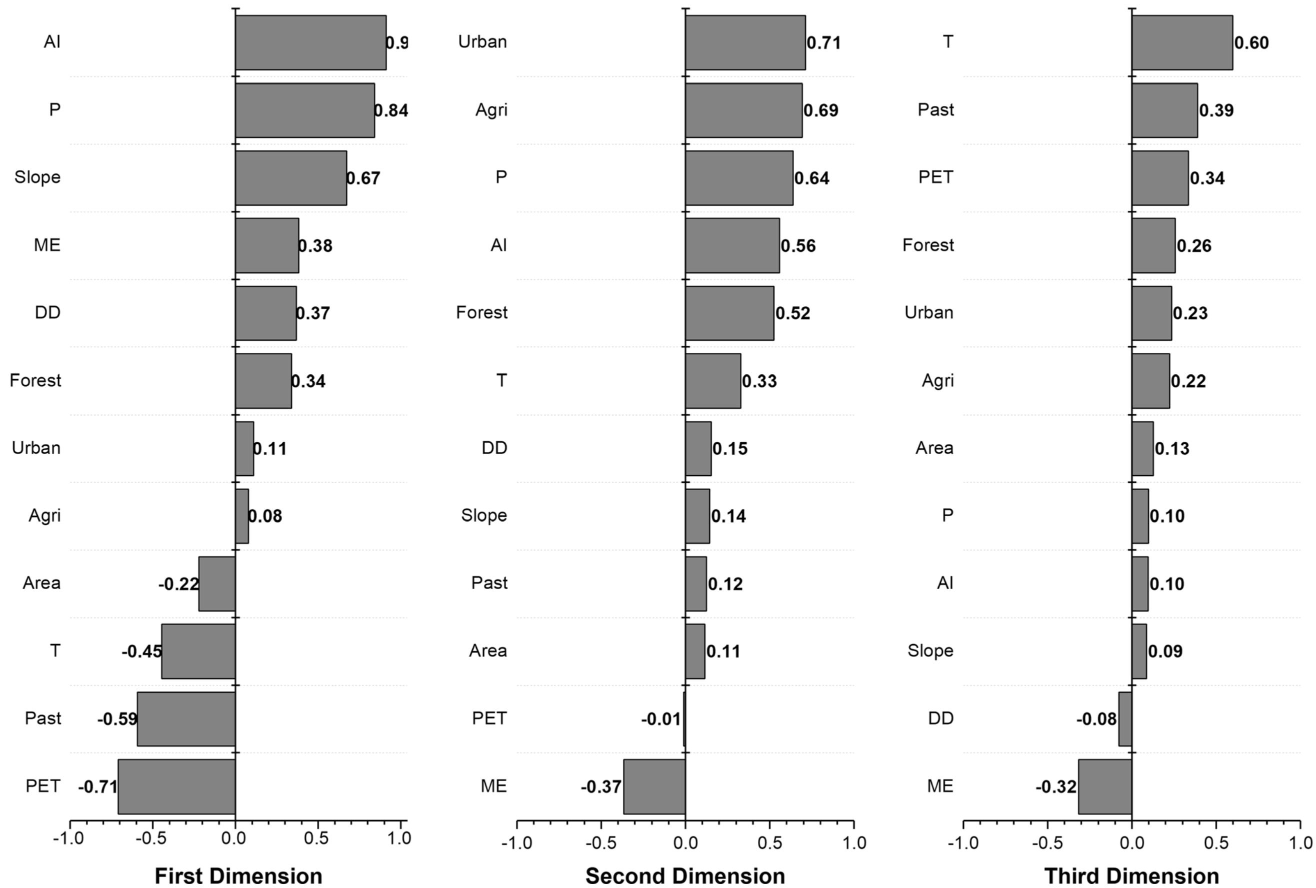

Figure 7. Spearman coefficients (ρ) between the analysed variables and the first three Isomap dimensions (AI= aridity index, P= mean precipitation, ME= mean elevation, DD= drainage density, Agri= agriculture, T= mean temperature, Past= pasture and PET= potential evapotranspiration).

calculate SI to address this limitation. In this method, instead of comparing largest groups, SI is calculated for each watershed by comparing the group of classification technique and reference classification in which the watershed lies. This modified method for calculating SI proposed by Ganvir and Eldho (2017) was used to evaluate classification techniques.

Similarity index obtained for both the classification techniques obtained using PCA and Isomap shows improvement over classification obtained using merely watershed attributes. Among the two dimensionality reduction techniques, Isomap outperformed PCA in classifying the watersheds into hydrologically similar watersheds (table 5). The mean, standard deviation and coefficient of variation of SI for Isomap were 0.448, 0.214 and 1.9, respectively. Although the average SI values for each of the techniques do not differ much, the relative frequency distribution of SI values for each technique suggests that Isomap classifies greater number of watersheds accurately and a smaller number of watersheds inaccurately. Isomap has a 20% likelihood, with SI in the range of 0.7–0.8, whereas PCA and watershed attributes have a 25–35% likelihood, with SI in the range of 0.6 and 0.7, respectively. Isomap is more likely to classify the watersheds with higher SI and less likely to classify those with lower SI (figure 8). Isomap has a 50% likelihood of classifying watersheds with SI more than 0.5. Thus, Isomap accurately classified most of the watersheds than other classification techniques. Hence, the clusters of Isomap classification were used for regionalization of SWAT parameters for streamflow estimation in Godavari basin. The following sections describe the results of calibration, validation, and streamflow estimation for Isomap clusters.

Table 5. *SI distribution for classification techniques.*

Sample statistics	Watershed attributes	PCA	Isomap
Average	0.38	0.43	0.45
Standard deviation	0.18	0.20	0.21
Coefficient of variation	0.47	0.45	0.48

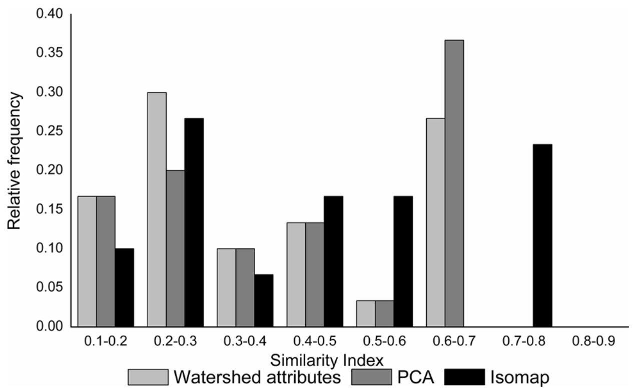

Figure 8. Relative frequency of numbers of watersheds (y-axis) with similarity index (x-axis) for Isomap, PCA and watershed attributes classification.

4.2 *SWAT calibration, validation and sensitivity analysis*

SWAT-CUP Sequential Uncertainty Fitting Version 2 (SUFI-2) was used for calibration, validation and sensitivity analysis of streamflow data during the years 1995–2005. First three years, 1995–1997, were used as warm-up period. SWAT simulation was done from the year 1998 to 2005. Auto-calibration with narrow sensitive range of parameters were selected for calibration of discharge at monthly time scale. Sensitive parameters for streamflow discharge were identified by running 1000 simulations in SWAT-CUP, while calibration using SUFI-2 by LH procedure. Using global sensitivity in SUFI-2 algorithms, most sensitive parameters for the streamflow calibration were chosen. The most sensitive parameter set for streamflow based on the *p*-stat and *t*-stat values include CN2, SOL_AWC and ESCO related to surface runoff; SLSUBBSN, OV_N, HRU_SLP related to HRU characteristics; and GW_REVAP, REVAPMN, GWQMN related to baseflow as given in table 6. As the main objective is to transfer the model parameters from gauged to ungauged, the set of selected sensitive parameters was kept unchanged for all the watersheds.

In the present study, first two groups contain nearly 60% of the watersheds while number of watersheds in remaining four groups were very few. Thus, only first two groups were considered for regionalization. All the watersheds from first two groups were calibrated for the complete period of 1998–2005. Within a group, one watershed was treated as pseudo-ungauged in turn and the remaining watersheds in the group were treated as donor watersheds. However, during calibration some of the watersheds showed poor model performance (table 7). Probably it was due to reservoirs present in those watersheds or the data quality. Such watersheds were excluded from the procedure of regionalization and remaining watersheds were used for the regionalization procedure. Watershed numbers 9, 57 and 92 from first group and watershed number 94 from the second group were excluded. The R^2, NSE and PBIAS values during the calibration procedure are shown in figure 9 using box and whisker plot. The ends of whisker represent maximum and minimum values, the ends of box represent 25th and 75th quartile and the median value is represented by the line inside the box.

R^2 describes the proportion of the variance in measured data explained by the model. R^2 ranges from 0 to 1, with higher values indicating less error variance and typically values > 0.5 are considered acceptable. Henriksen *et al.* (2003) suggest that an

Table 6. *SWAT parameters used for calibration and regionalization.*

Parameters	Description	Process
*r_CN2	Initial SCS CN II value	Runoff
r_SOL_AWC	Available water capacity of the soil layer	Soil
r_ESCO	Soil evaporation compensation factors	Evaporation
*v_SLSUBBSN	Average slope length	HRU
v_OV_N	Manning's 'n' value for overland flow	HRU
r_HRU_SLP	Average slope steepness	HRU
r_GW_REVAP	Groundwater 'revap' coefficient	Groundwater
r_GWQMN	Threshold depth of water in the shallow aquifer required for return flow to occur (mm)	Groundwater
v_REVAPMN	Threshold depth of water in the shallow aquifer for revap to occur (mm H_2O)	Groundwater

*v_ means the existing parameter value is to be replaced by a given value, *a_ stands for a given value is added to the existing parameter value, *r_ represents an existing parameter value is multiplied by (1+ the given value).

Table 7. *Performance of SWAT model for watersheds selected for regionalization during calibration.*

Basin no.	Stream gauging station	Group	Calibrated R^2	Calibrated NS	Calibrated PBIAS
10	Salebardi	1	0.72	0.71	−2.5
14	Wairagarh	1	0.76	0.73	−26.5
19	Rajoli	1	0.83	0.77	−8
34	Cherribeda	1	0.8	0.8	−3
48	Kosarguda	1	0.88	0.75	−34.7
50	Ambabal	1	0.84	0.76	−26.8
52	Sonarpal	1	0.77	0.63	−24.3
94	Potteru	1	0.32	0.21	−106.32
9	Bishnur	2	0.45	0.41	62.31
12	Nandgaon	2	0.77	0.66	2.5
16	Mangrul	2	0.82	0.82	8
25	Karnergaon	2	0.84	0.83	5.5
54	Bhatpalli	2	0.82	0.81	15.7
57	Samangaon M	2	0.38	0.32	49.26
85	Somanpally	2	0.62	0.6	22.8
92	Bhatkheda	2	0.42	0.39	70.33
99	Aurad (Sh)	2	0.72	0.72	−7.7

Table 8. *Overall performance of global mean (GM), Inverse distance weighted (IDW) and physical similarity (PS) based on SWAT model performance.*

Statistical measures	R^2 GM	R^2 IDW	R^2 PS	NSE GM	NSE IDW	NSE PS	PBIAS GM	PBIAS IDW	PBIAS PS
Mean	0.74	0.75	0.75	0.50	0.47	0.56	−33.58	−26.15	−28.22
Median	0.75	0.74	0.72	0.60	0.64	0.60	−36.30	−34.00	−40.00
Standard deviation	0.09	0.07	0.07	0.39	0.50	0.21	39.11	43.50	30.80
Coefficient of variation	0.12	0.09	0.10	0.78	1.05	0.38	−1.16	−1.66	−1.09
Maximum	0.88	0.88	0.88	0.79	0.79	0.79	27.20	33.90	24.70
Minimum	0.59	0.66	0.66	−0.54	−1.10	0.02	−104.00	−113.80	−69.20

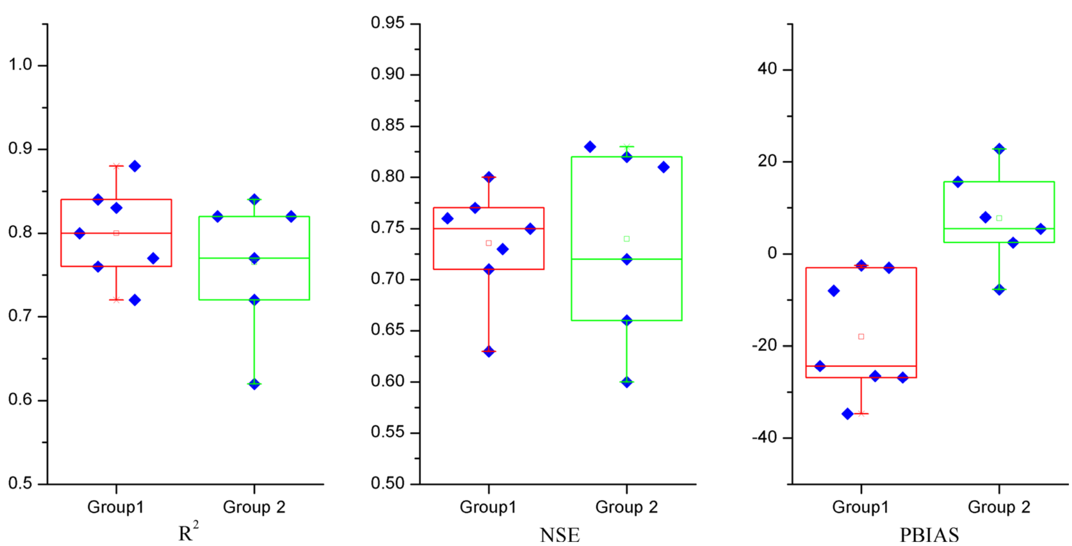

Figure 9. R^2, NSE and PBIAS values for group 1 and 2 of Isomap classification during calibration.

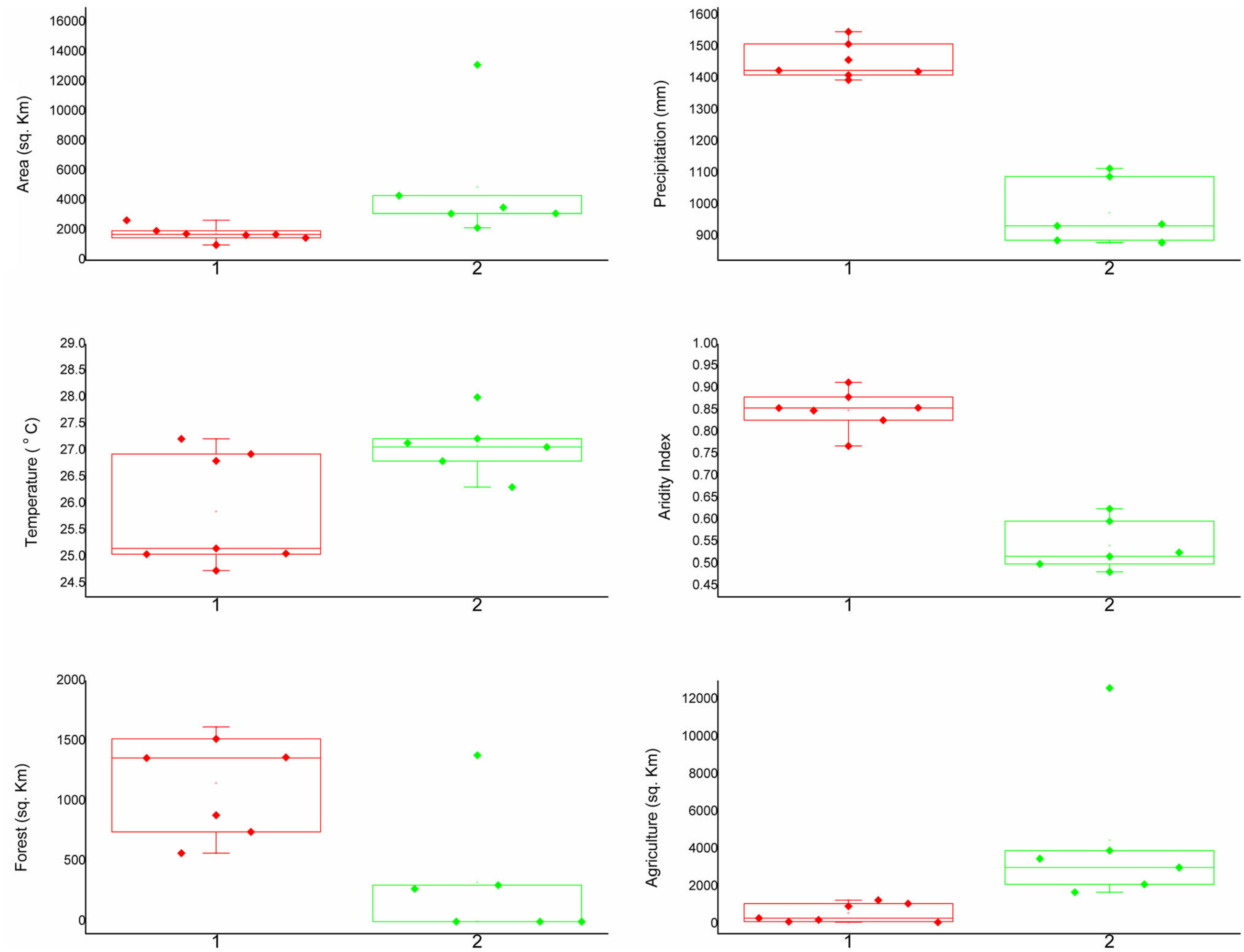

Figure 10. Distribution of watershed attributes for first two largest groups of Isomap classification. Whisker plots show 10th, 25th, 50th (median), 75th and 90th percentile.

R^2 value > 0.85 is excellent for a hydrological model, values between 0.65 and 0.85 are very good, 0.5–0.65 are good, 0.20–0.50 are poor and <0.20 are very poor. R^2 values of group 1 and group 2 are between 0.72–0.88 and 0.62–0.84, respectively, which are in good to very good range.

NSE ranges between $-\infty$ and 1.0 (1 inclusive), with NSE = 1 being the optimal value. Values between 0.0 and 1.0 are generally viewed as acceptable levels of performance, whereas values <0.0 indicate that the mean observed value is a better predictor than the simulated value, which indicates unacceptable performance (Moriasi *et al.* 2007). NSE values of group 1 and group 2 are between 0.63–0.8 and 0.6–0.83, respectively, which are in acceptable range.

The optimal value of PBIAS is 0.0, with low-magnitude values indicating accurate model simulation. Positive values indicate model underestimation bias and negative values indicate model overestimation bias (Gupta *et al.* 1999). PBIAS values of group 1 are between -34.7 and -2.5. Negative values of PBIAS for group 1 suggests that most of the model simulated streamflow is overestimated. PBIAS values for group 2 is between $-$ 7.7 and 22.8, all are positive values except for watershed number 99 which suggests model simulated streamflow for group 2 is underestimated.

4.3 *Regionalization*

The overall summary statistics for the performance of the regionalization methods is shown in table 8. All the three regionalization methods maintain the collinearity between the observed and estimated streamflow data with mean R^2 values of 0.74, 0.75 and 0.75 for global mean, IDW and physical similarity, respectively. Physical similarity performs better than global mean and IDW methods in

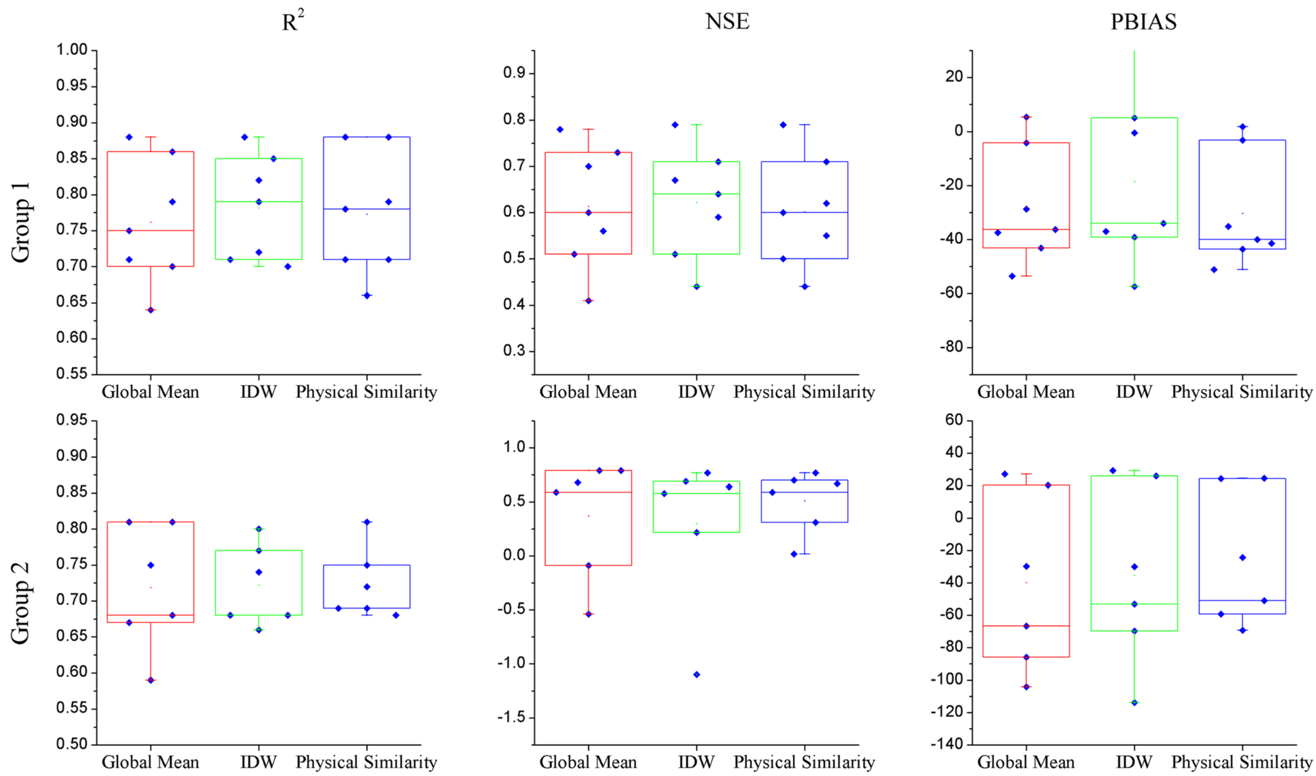

Figure 11. R^2, NSE and PBIAS values for group 1 and 2 for regionalization methods.

terms of NSE with a value of 0.56 and 0.38 for mean and coefficient of variation. PBIAS values suggest that model is overpredicting the streamflow in most of the watersheds in case of all the regionalization methods. All the three methods show poor values of NSE for basin number 12 and 85 (these watersheds showed poor values of NSE during calibration also), both the watersheds are grouped in same group by Isomap. Group 2 consists of watersheds with larger areas than watersheds of group 1. Upon further analysis (see figure 10), it was observed that these two watersheds have larger areas than other watersheds considered in the study (4,335 and 13,141 km^2 for watershed number 12 and 85, respectively). These two watersheds also have the lowest streamflow elasticity values among all the watersheds. These results suggest that these regionalization methods are not suitable to estimate streamflow in watersheds with larger areas and which are least sensitive to change in streamflow with change in precipitation. However, among the three methods, physical similarity showed comparatively good results for these two watersheds. The reason that physical similarity is performing better in these two watersheds is that, the regionalized model parameters for these watersheds were weighted based on physical similarity. As these watersheds are larger in terms of area, more weightage was given to the parameters of larger watersheds while regionalization. However, global mean simply averaged the parameter values and IDW gave weightages based on inverse distance thereby completely neglecting the attribute, 'area'. Figure 10 also shows that the Isomap very well captures the data of watershed attributes to classify them into different groups.

Group-wise results for SWAT model performance for Isomap classification are shown in figure 11. In group 1, the mean values of R^2 for global mean, IDW and physical similarity are 0.75, 0.79 and 0.79, respectively, which suggest that all the three regionalization methods captured the collinearity between the estimated and observed flows. Mean NSE value for IDW is 0.64 which is higher than other two methods which shows that IDW is better at estimating the streamflow in group 1. Most of the PBIAS values of all the three methods are negative, which shows that the streamflow values are over-predicted by all the three methods. However, figure 11 shows higher range as well as interquartile range of NSE for global mean values than other two methods which means that global mean has higher variability in

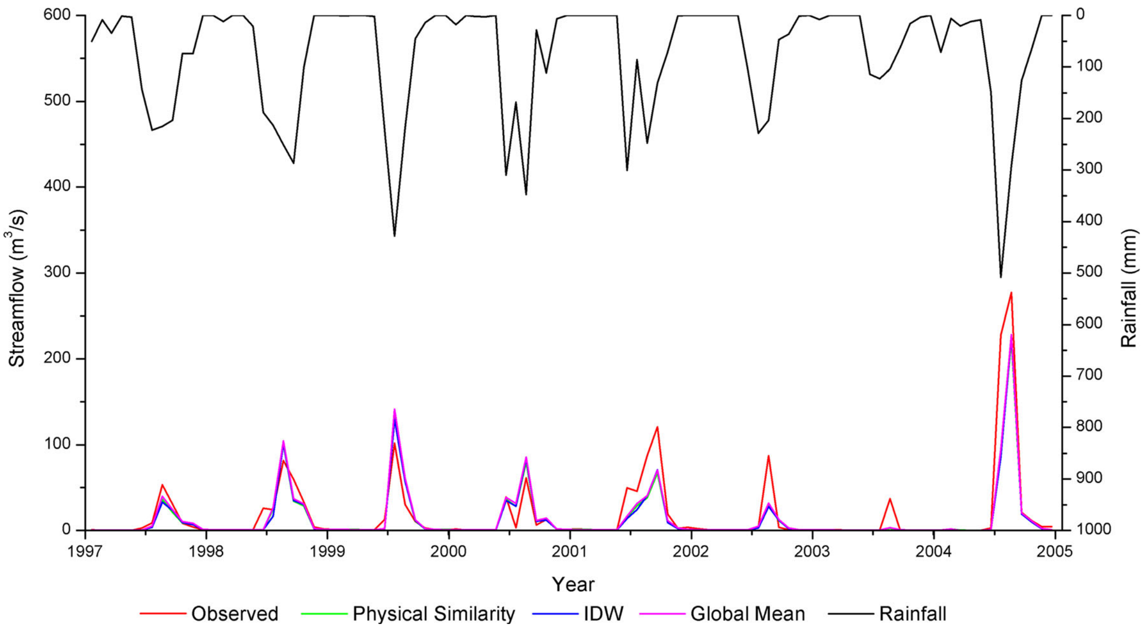

Figure 12. Streamflow estimation for Watershed no. 16 (Mangrul) using regionalization techniques.

terms of estimating streamflow while IDW and physical similarity shows less variability in terms of streamflow estimation.

Even in group 2, all the three methods maintained the collinearity between the observed and estimated streamflow data. Median value for NSE is nearly same for all the three methods. However, IDW and physical similarity show lesser variability in terms of NSE. PBIAS values range from 20 to −104 for global mean, 29 to −119 for IDW and 24.7 to −69.2 for physical similarity. Most of the values for PBIAS are negative which indicates model is overpredicting the streamflow values for the regionalization methods. The results of streamflow estimation for watershed number 16 (Mangrul), for all the regionalization techniques is shown on representative basis out of 30 watersheds used in the study in figure 12. It is evident that regionalization techniques capture the overall streamflow except peak flows.

5. Conclusions

This study proposes a framework for reliable estimation of streamflow in ungauged basins by combining watershed classification techniques with regionalization methods. The developed framework is applied in ungauged watersheds of Godavari river basin in India. Thirty watersheds of Godavari river were first classified by applying Isomap and PCA over the watershed attributes. Further global mean, IDW and physical similarity were used to regionalize the SWAT parameters. R^2, NSE and PBIAS were used to evaluate the regionalization methods based on the predicted streamflow in SWAT.

The results of the present study classify watersheds and allocate watersheds into different groups based on noteworthy watershed attributes. If these properties define the hydrologic response of a watershed, then classification techniques tend to group similar watershed in one group. Isomap outperforms PCA and watershed attributes in these terms. Moreover, classification helps to reduce the glitch of transferring parameters from dissimilar watershed to the targeted watershed. Among the regionalization techniques, physical similarity performs better than global mean and IDW. Global mean simply averages the model parameter values of donor watersheds and IDW gives weightages to the model parameters of donor watersheds based on the distance from the target watersheds in the group. However, if the attributes used for classification have wide range, these techniques seem to be very crude for estimating model parameters of the target watershed. On the other hand, physical similarity overcomes these drawbacks by giving more weightage to the model parameters of the donor watershed which is more similar to the target watershed.

Overall classifying the watershed based on Isomap prior to regionalization using physical similarity for estimating streamflow in SWAT improves the efficiency of estimating streamflow in comparison to other regionalization techniques used in the study. Further, the application of present methodology to a study area that includes more number of watersheds with vide variability in the watershed attributes may improve the reliability of the present methodology. Besides, present study considers stationary scenario for land use land cover (LULC) and climate. Further application of the present methodology for non-stationary scenarios of LULC and climate can be investigated, to understand its effects.

References

ARCmap Version-10.1 2012 (Computer software) ESRI, Redlands, CA.

Abbaspour K C 2015 *SWAT-CUP: SWAT Calibration and Uncertainty Programs – A User Manual*; Swiss Federal Institute of Aquatic Science and Technology, Duebendorf, Switzerland.

Abbaspour K C, Yang J, Maximov I, Siber R, Bogner K, Mieleitner J, Zobrist J and Srinivasan R 2007 Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT; *J. Hydrol.* **333** 413–430, https://doi.org/10.1016/j.jhydrol.2006.09.014.

Arnold J G, Srinivasan R, Muttiah R S and Williams J R 1998 Large area hydrologic modeling and assessment. Part I: Model development; *J. Am. Water Resour. Assoc.* **34** 73–89.

Athira P, Sudheer K P, Cibin R and Chaubey I 2016 Predictions in ungauged basins: An approach for regionalization of hydrological models considering the probability distribution of model parameters; *Stoch. Environ. Res. Risk Assess.* **30** 1131–1149, https://doi.org/10.1007/s00477-015-1190-6.

Blöschl G and Sivapalan M 1995 Scale issues in hydrological modelling: A review; *Hydrol. Process.* **9** 251–290, https://doi.org/10.1002/hyp.3360090305.

Blöschl G, Sivapalan M, Wagener T, Viglione A and Savenije H 2013 *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*; Cambridge University Press, Cambridge, UK.

Chiang S-M, Tsay T-K and Nix S J 2002a Hydrologic regionalization of watersheds. I: Methodology development; *J. Water Resour. Plan Manag.* **128** 3–11.

Chiang S-M, Tsay T-K and Nix S J 2002b Hydrologic regionalization of watersheds. II: Applications; *J. Water Resour. Plan Manag.* **128** 12–20.

Moriasi D N, Arnold J G, Van Liew M W, Bingner R L, Harmel R D and Veith T L 2007 Model evaluation guidelines for systematic quantification of accuracy in watershed simulations; *Trans. ASABE* **50(3)** 885–900, https://doi.org/10.13031/2013.23153.

Davies D L and Bouldin D W 1979 A cluster separation measure; *IEEE Trans. Pattern Anal. Mach. Intell.* **2** 224–227, https://doi.org/10.1109/tpami.1979.4766909.

Di Prinzio M, Castellarin A and Toth E 2011 Data-driven catchment classification: Application to the pub problem; *Hydrol. Earth Syst. Sci.* **15** 1921–1935, https://doi.org/10.5194/hess-15-1921-2011.

Ganvir K and Eldho T I 2017 Watershed classification using isomap technique and hydrometeorological attributes; *J. Hydrol. Eng.* **22** 04017040, https://doi.org/10.1061/(asce)he.1943-5584.0001562.

Gitau M W and Chaubey I 2010 Regionalization of SWAT model parameters for use in ungauged watersheds; *Water* **2** 849–871, https://doi.org/10.3390/w2040849.

Gupta H V, Sorooshian S and Yapo P O 1999 Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration; *J. Hydrol. Eng.* **4(2)** 135–143.

He Y, Bárdossy A and Zehe E 2011 A catchment classification scheme using local variance reduction method; *J. Hydrol.* **411** 140–154, https://doi.org/10.1016/j.jhydrol.2011.09.042.

Henriksen H J, Troldborg L and Nyegaard P *et al.* 2003 Methodology for construction, calibration and validation of a national hydrological model for Denmark; *J. Hydrol.* **280** 52–71, https://doi.org/10.1016/s0022-1694(03)00186-0.

Heuvelmans G, Muys B and Feyen J 2004 Evaluation of hydrological model parameter transferability for simulating the impact of land use on catchment hydrology; *Phys. Chem. Earth* **29** 739–747, https://doi.org/10.1016/j.pce.2004.05.002.

Hijmans R J, Cameron S E, Parra J L *et al.* 2005 Very high resolution interpolated climate surfaces for global land areas; *Int. J. Climatol.* **25** 1965–1978, https://doi.org/10.1002/joc.1276.

Hosking J and Wallis J 1997 *Regional frequency analysis: An approach based on L-moments*; Cambridge University Press, Cambridge, UK.

Jackson J E 1991 *A User's Guide to Principal Components*; Hoboken, NJ, Wiley-Interscience, https://doi.org/10.1002/0471725331.

Jarvis A, Reuter H I, Nelson A and Guevara E 2008 Hole-filled SRTM for the globe Version 4, CGIAR-CSI SRTM 90m Database, http://srtm.csi.cgiar.org.

Kileshye Onema J M, Taigbenu A E and Ndiritu J 2012 Classification and flow prediction in a data-scarce watershed of the equatorial Nile region; *Hydrol. Earth Syst. Sci.* **16** 1435–1443.

MacQueen J 1967 Some methods for classification and analysis of multivariate observations; In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (eds) Le Cam L M and Neyman J, vol. 1, University of California Press, Berkeley, CA, pp. 281–297.

McIntyre N, Lee H, Wheater H, Young A and Wagener T 2005 Ensemble predictions of runoff in ungauged catchments; *Water Resour. Res.* **41(12)** https://doi.org/10.1029/2005wr004289.

Merz R and Blöschl G 2004 Regionalisation of catchment model parameters; *J. Hydrol.* **287** 95–123, https://doi.org/10.1016/j.jhydrol.2003.09.028.

Nash J E and Sutcliffe J V 1970 River flow forecasting through conceptual models. Part I: A discussion of principles; *J. Hydrol.* **10** 282–290, https://doi.org/10.1016/0022-1694(70)90255-6.

Nathan R J and Mcmahon T A 1990 Identification of homogeneous regions for the purposes of regionalisation; *J. Hydrol.* **121(1)** 217–238.

Neitsch S, Arnold J, Kiniry J and Williams J 2011 *Soil and Water Assessment Tool Theoretical Documentation Version 2009*; Texas Water Resource Institute, 647p, https://doi.org/10.1016/j.scitotenv.2015.11.063.

Oudin L, Andréassian V and Perrin C *et al.* 2008 Spatial proximity, physical similarity, regression and ungaged catchments: a comparison of regionalization approaches based on 913 French catchments; *Water Resour. Res.* **44(3)**, https://doi.org/10.1029/2007wr006240.

Parajka J, Merz R and Blöschl G 2005 A comparison of regionalisation methods for catchment model parameters; *Hydrol. Earth Syst. Sci. Discuss.* **2** 509–542, https://doi.org/10.5194/hessd-2-509-2005.

Razavi T and Coulibaly P 2013a Streamflow prediction in ungauged basins: Review of regionalization methods; *J. Hydrol. Eng.* **18** 958–975, https://doi.org/10.1061/(asce)he.1943-5584.0000690.

Razavi T and Coulibaly P 2013b Classification of Ontario watersheds based on physical attributes and streamflow series; *J. Hydrol.* **493** 81–94, https://doi.org/10.1016/j.jhydrol.2013.04.013.

Rees G, Croker K, Zaidman M, Cole G, Kansakar S, Chalise S, Kumar A, Saraf A and Singhal M 2002 Application of the regional flow estimation method in the Himalayan region; *IAHS Publ.* **274** 433–440.

Rees G, Holmes M G R, Young A R and Kansakar S R 2004 Recession-based hydrological models for estimating low flows in ungauged catchments in the Himalayas; *Hydrol. Earth Syst. Sci. Discuss.* **8** 891–902, https://doi.org/10.5194/hess-8-891-2004.

Samuel J, Coulibaly P and Metcalfe R A 2011 Estimation of continuous streamflow in Ontario ungauged basins: Comparison of regionalization methods; *J. Hydrol. Eng.* **16** 447–459, https://doi.org/10.1061/(asce)he.1943-5584.0000338.

Sawicz K, Wagener T and Sivapalan M *et al.* 2011 Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA; *Hydrol. Earth Syst. Sci.* **15** 2895–2911, https://doi.org/10.5194/hess-15-2895-2011.

Sellami H, La Jeunesse I and Benabdallah S *et al.* 2014 Uncertainty analysis in model parameters regionalization: A case study involving the SWAT model in Mediterranean catchments (Southern France); *Hydrol. Earth Syst. Sci.* **18** 2393–2413, https://doi.org/10.5194/hess-18-2393-2014.

Shepard D 1968 A two-dimensional interpolation function for irregularly-spaced data; In: *23rd ACM Natl. Conf.*, pp. 517–524, https://doi.org/10.1145/800186.810616.

Singh R D, Mishra S K and Chowdhary H 2001 Regional flow duration models for large number ungauged Himalayan catchments for planning micro-hydro projects; *J. Hydrol. Eng.* **6(4)** 310–316.

Sivapalan M, Takeuchi K and Franks S W *et al.* 2003 IAHS Decade on predictions in ungauged basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences; *Hydrol. Sci. J.* **48** 857–880, https://doi.org/10.1623/hysj.48.6.857.51421.

Ssegane H, Tollner E W and Mohamoud Y M *et al.* 2012 Advances in variable selection methods. I: Causal selection methods versus stepwise regression and principal component analysis on data of known and unknown functional relationships; *J. Hydrol.* **438–439** 16–25, https://doi.org/10.1016/j.jhydrol.2012.01.008.

Swain J B and Patra K C 2017 Streamflow estimation in ungauged catchments using regionalization techniques; *J. Hydrol.* **554** 420–433, https://doi.org/10.1016/j.jhydrol.2017.08.054.

Tenenbaum J B, de Silva V and Langford J C 2000 A global geometric framework for nonlinear dimensionality reduction; *Science* **209** 2319–2323.

Trabucco A and Zomer R J 2009 Global Aridity Index (Global-Aridity) and Global Potential Evapo-Transpiration (Global-PET) Geospatial Database, http://www.csi.cgiar.org.

Young A R 2006 Stream flow simulation within UK ungauged catchments using a daily rainfall-runoff model; *J. Hydrol.* **320** 155–172, https://doi.org/10.1016/j.jhydrol.2005.07.017.

Corresponding editor: Rajib Maity