Water Resources Management
https://doi.org/10.1007/s11269-023-03678-8

Novel Approaches for Regionalising SWAT Parameters Based on Machine Learning Clustering for Estimating Streamflow in Ungauged Basins

Javier Senent-Aparicio[1] · Patricia Jimeno-Sáez[1] · Raquel Martínez-España[2,3] · Julio Pérez-Sánchez[1,4]

Received: 22 February 2023 / Accepted: 20 November 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract
Streamflow prediction in ungauged basins (PUB) is necessary for effective water resource management, flood assessment, and hydraulic engineering design. Spain is one of the countries in Europe expected to suffer the most from the consequences of climate change, notably an increase in flooding. The authors selected the Miño River basin in the northwest of Spain, which covers an area of 2,168 km^2, to develop a novel approach for predicting streamflow in ungauged basins. This study presents a regionalisation of the soil and water assessment tool (SWAT), a semi-distributed, physically based hydrological model. The regionalisation approach transfers SWAT model parameters based on hydrological similarities between gauged and ungauged subbasins. The authors used *k*-means and expectation–maximisation (EM) machine learning clustering techniques to group 30 subbasins (9 gauged subbasins) into homogeneous, physical, similarity-based clusters. Furthermore, the regionalisation featured physiographic attributes (basin area, elevation, and channel length and slope) and climatic information (precipitation and temperature) for each subbasin. For each homogeneous group, the SWAT model was calibrated and validated for the gauged basins (donor basins), and the calibrated parameters were transferred to the pseudo-ungauged basins (receptor basins) for streamflow prediction. The results of the streamflow prediction in the pseudo-ungauged basins demonstrate satisfactory performance in most of the cases, with average NSE, R^2, RSR, and RMSE values of 0.78, 0.91, 0.42, and 5.10 m^3/s, respectively. The results contribute to water planning and management and flood estimation in the studied region and similar areas.

Keywords Hydrological Model · Streamflow Prediction · Ungauged Basins · Regionalisation · Clustering · SWAT

Extended author information available on the last page of the article

Published online: 28 November 2023

1 Introduction

Streamflow, or runoff, is one of the crucial flows, along with precipitation and evaporation, in the hydrological cycle (Trenberth et al. 2007). The amount of runoff has a direct impact on human life, as it is the main water resource for the different uses in the basins. Increased demand for water due to population growth and improved living standards (Jodar-Abellan et al. 2018), along with decreased water resource quality due to the continuous introduction of undesirable chemicals (Ali et al. 2009; Basheer 2018a, b), underscores the imperative need for accurate streamflow estimation to ensure an adequate quantity of quality water is available. Therefore, streamflow estimation is not only essential for understanding the different hydrological processes in the basin but also for water resource management, planning, flood prediction, and hydraulic engineering design (Guo et al. 2021). Rainfall–runoff models are standard tools for modelling river basin streamflow, and calibrating their parameters is a mandatory task to achieve reliable predictions. The number of monitoring stations worldwide is decreasing, generating a lack of hydrometric data – especially in emerging economies. This decrease is typically due to the high costs of investment in, as well as the operation and maintenance of, traditional hydrometric monitoring systems. In the case of ungauged basins, no observed streamflow data is available for model calibration because it is either of poor quality, inaccessible, or nonexistent. Therefore, streamflow prediction in ungauged basins (PUB) has been and continues to be a significant challenge for the global hydrological community (Darko et al. 2021). The International Association of Hydrological Sciences (IAHS) initiated a 10-year scientific plan to address the PUB problem in 2003 (Sivapalan et al. 2003; Hrachowitz et al. 2013) and recently reiterated its importance by including it among the major unsolved problems in hydrology (Blöschl et al. 2019). A common approach to solving this problem is to take of physically-based hydrological models and regionalise their parameters using basin characteristics (Yadav et al. 2007; Cheng et al. 2021).

Regionalisation consists of transposing the model parameters or general hydrological information of a gauged basin, termed a donor basin, to a similar ungauged basin, termed a receptor basin (Razavi and Coulibaly 2013a). The most commonly used parameter regionalisation approaches are regression-based and similarity-based (Wu et al. 2022). Similarity-based regionalisations can be grouped into: those based on spatial proximity (Beza et al. 2023; Ssegane et al. 2012) and those based on physical similarity (Singh et al. 2009; Mosavi et al. 2021). Hydrological regionalisation based on the basin's physical properties can provide information on how and to what extent climatic and landscape features control the basin's hydrological characteristics (Gao et al. 2018). This regionalisation concept is based on the premise that basins with similar characteristics (e.g. climate, topography, vegetation, and soils) typically have similar streamflow responses (Smakhtin 2001). The physical similarity method involves a cluster analysis of basins to find a donor basin with physical characteristics similar to the target basin (Guo et al. 2021). Clustering has been successfully applied as a machine learning technique in earth science modelling due to its impressive performance in nonlinear relationship processing. Basin clustering can be performed using various clustering algorithms, such as *k*-means (Razavi and Coulibaly 2013b), hierarchical agglomerative clustering (Farsadnia et al. 2014), fuzzy clustering (Mosavi et al. 2021), and hybrid clustering (Ramachandra Rao and Srinivas 2006).

Following regionalisation, an accurate hydrological model should be used to simulate streamflow in ungauged basins. Basin-scale models are prominent hydrological models due to their ability to simulate scenarios and their applicability to developing management policies. Notable among them, is the soil and water assessment tool (SWAT) (Arnold et al. 1998), a semi-distributed, ecohydrological, public domain model. SWAT is the most efficient model for solving various hydrological problems at different scales and in different scenarios (Balha et al. 2023). It is well documented in the literature, and has been used for streamflow simulation in ungauged basins (Srinivasan et al. 2010; Sisay et al. 2017; Mosavi et al. 2021; Singh et al. 2022). The model relies on computational efficiency due to its semi-distributed and aggregated approach, which also makes it applicable to continental domains. Several studies have focused on transferring parameters based on the physical similarity approach to predict streamflow records in ungauged basins using SWAT (Sellami et al. 2014; Swain and Patra 2017; Mosavi et al. 2021; Wu et al. 2022; Gebeyehu et al. 2023). However, few studies have used machine learning, particularly clustering, to regionalise SWAT model parameters (Mosavi et al. 2021).

Strong evidence indicates that Spain is one of the countries most affected by climate change in Europe. An observable rise in the magnitude of rainfall has increased the flooding of rivers and *wadis*, escalating the potential risk to infrastructure and urban areas (Eguibar et al. 2021; Senent-Aparicio et al. 2023). Therefore, estimating streamflow records in ungauged basins is essential. In this study, we selected the headwaters of the Miño River based on the availability and quality of the streamflow data required to perform and validate this research.

The main aim of this study is to evaluate streamflow estimates in a set of ungauged basins of the Miño River in the northwest of Spain using a hydrological SWAT model with parameters obtained using a regionalisation method. The model parameters are transferred from gauged basins to target ungauged basins using a physical similarity criteria. The homogeneous basins are defined using the most widely used clustering techniques, such as *k*-means and expectation–maximisation (EM). To our knowledge, combining these clustering techniques with the SWAT model presents a novel approach which has never been examined for the purpose of modelling ungauged basins. No similar studies on the Miño River region or the rest of Spain exist. Therefore, this methodology is a significant step forward for water resource management in the area.

2 Study Area and Data

2.1 Study Area

The Miño River is the most important river in the province of Galicia, located in the northwest corner of Spain. The study area is the headwater of the Miño River Basin (HMRB; Fig. 1a) within the *Terras do Mino*, declared a biosphere reserve in 2002 by UNESCO and designated a site of community importance (SCI) by the European Union due to its significant ecological value. The HMRB has an area of 2,168 km^2 and a mean elevation of 503 m, ranging from 363 m in the river valley to 1,028 m in the mountainous areas (Fig. 1b).

The climate is mild and rainy, influenced by the Atlantic Ocean (Di Blasi et al. 2013). Furthermore, the mean annual precipitation in the study basin is 1,200 mm, and the aver-

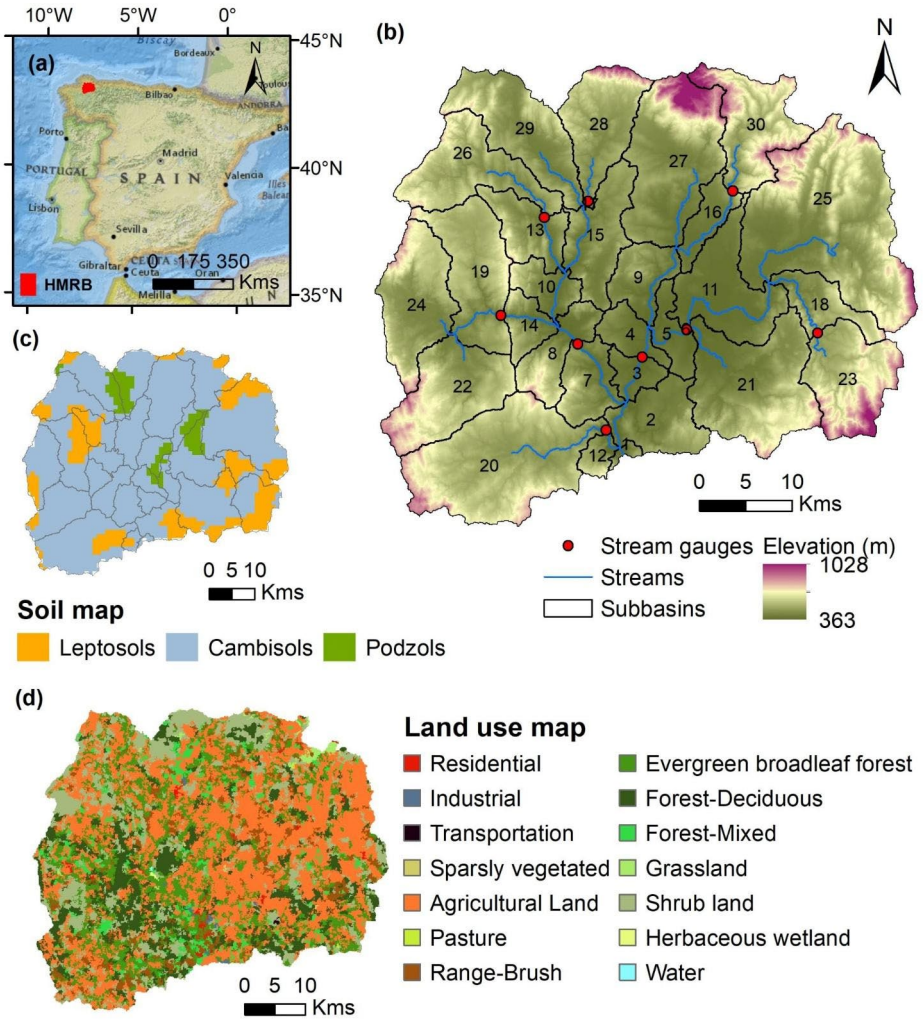

Fig. 1 (**a**) Location of the HMRB in Spain; (**b**) digital elevation model (DEM), stream gauges, and sub-basin boundaries of the basin; (**c**) soil; and (**d**) land use map of the basin

age annual temperature is 11.4 ºC. Persistent Atlantic fronts from the west produce abundant flows from autumn to spring that gradually decrease until the summer, which is the driest season (Jimeno-Sáez et al. 2018). In total, 77% of the basin has a predominantly Humic Cambisol soil type, the dominant soil type throughout the region (Fig. 1c), with a loam texture (41% sand, 36% silt, and 23% clay). The basin includes a mixture of natural and agricultural land interspersed with small rural zones (Senent-Aparicio et al. 2019). The dominant land use is agriculture (41% of the area), followed by forest-type land cover (36%) and brush, grassland, and pasture areas (23%), as shown in Fig. 1d.

2.2 Data

Seven parameters were selected for subbasin clustering based on a literature review (Razavi and Coulibaly 2013a; Barbarossa et al. 2017; Swain and Patra 2017; Mosavi et al. 2021) and data availability: area, maximum elevation, minimum elevation, channel length, channel slope, mean annual precipitation, and mean annual temperature. In general, these characteristics are most frequently used by researchers in streamflow regionalisation (Razavi and Coulibaly 2013a). Figure 2 illustrates the spatial distribution of these subbasin characteristics. The areas of the subbasins range from 1 to 265 km^2. The subbasins with the highest elevations are located in the eastern part of the basin (Fig. 2b). The valley zone, corresponding to the central part of the basin, contains the subbasins with lower elevations (Fig. 2b and c), lower channel slopes (Fig. 2e), lower precipitation (Fig. 2f), and higher mean temperatures (Fig. 2g).

The physical parameters of each subbasin were obtained from the information generated by the SWAT model in QGIS software. The SWAT model required a digital elevation model (DEM) 25×25 m provided by the Spanish National Geographic Institute (IGN: https://www.ign.es/web/ign/portal/cbg-area-cartografia; Fig. 1b), a 1 km resolution soil map implemented from the Harmonized World Soil Database (Nachtergaele et al. 2008; Fig. 1c), and a land use map extracted from Corine Land Cover (2012; Fig. 1d). This data has been used for the same area in previous studies (Jimeno-Sáez et al. 2018; Senent-Aparicio et al. 2019). The basin comprises 30 subbasins, nine of which are gauged subbasins due to the presence of stream gauges at their outlets (Figs. 1b and 2a). Monthly streamflow data available for the period 2011–2018 was collected from the Centre for Hydrographic Studies of CEDEX website (http://ceh-flumen64.cedex.es/anuarioaforos/default.asp) for all hydrological stations. Daily precipitation and temperature data was downloaded from the SWAT website (https://swat.tamu.edu/data/spain/). This climate data from the Spanish National Meteorological Service (AEMET), in a ready-to-use format for entry into the hydrological SWAT model, is available from 1951 to 2019 with a resolution of 5 km for all of Spain (Senent-Aparicio et al. 2021).

3 Methodology

The research objective was to develop a technique for estimating streamflow in the ungauged basins using a hydrological regionalisation method. Figure 3 provides a flowchart of the steps involved. A detailed description of the workflow is given in Appendix A. The three steps of the methodology are described in the following sections.

3.1 Hydrological Modelling

We used the hydrological SWAT model for subbasin delineation, hydrological modelling, and streamflow estimation. SWAT is a semi-distributed, physically based hydrological model operated in QGIS through the QSWAT extension. SWAT is a globally employed and efficient tool for quantifying various hydrological components, such as surface runoff, sediments, and pollution (Al-Khafaji et al. 2020). The flowchart in Fig. 3 shows that the first step of the SWAT model was to delineate the subbasins from the DEM and define the outlets

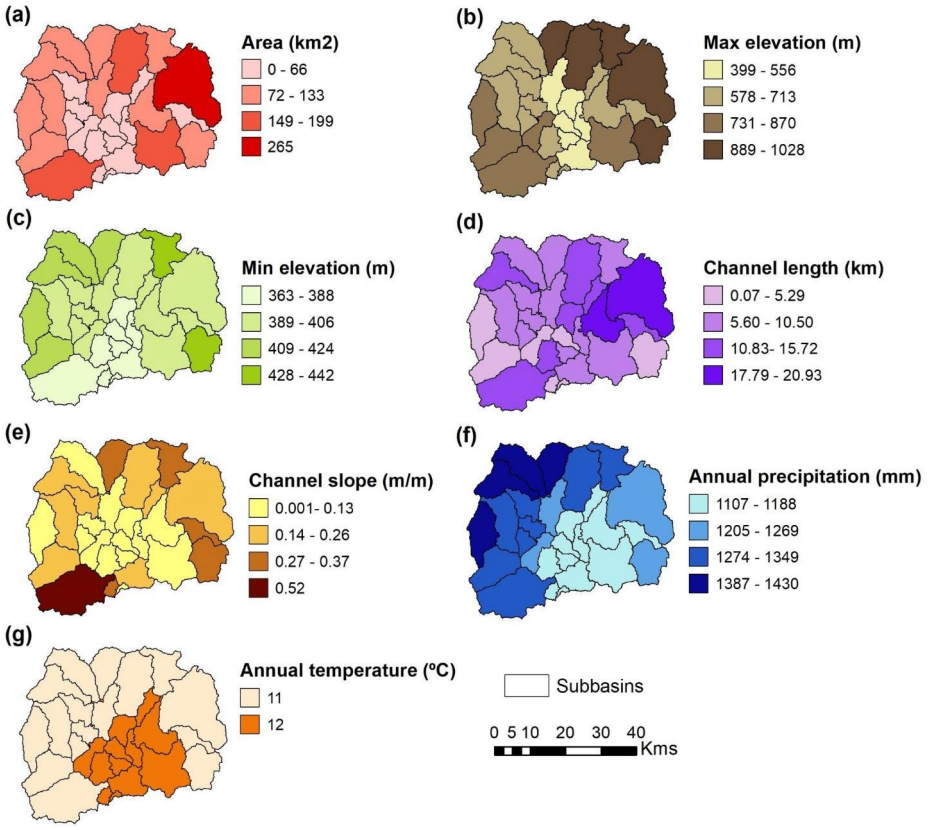

Fig. 2 Spatial distributions of subbasin characteristics in HMRB

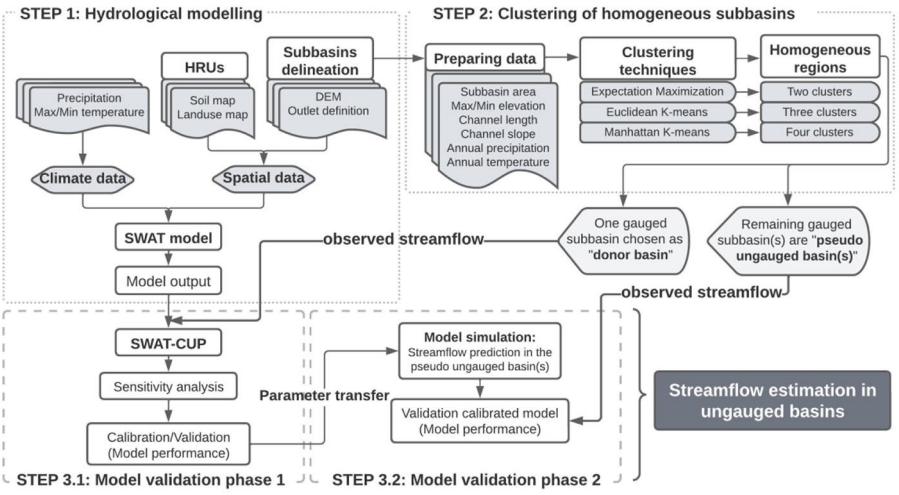

Fig. 3 Methodology flowchart

of several subbasins. The model then used the slope, land use, and soil type information to define hydrological response units (HRUs). The final step was to enter the meteorological data into the model and update its database. We selected the Hargreaves-Samani approach, requiring daily maximum and minimum temperature data to estimate potential evapotranspiration. The SWAT model was then run, and the streamflow of the subbasins was simulated as the model's output based on a water balance that considers the soil water content, precipitation, surface runoff, evapotranspiration, percolation, and baseflow quantity. Furthermore, the SWAT model simulated the various parameters using different mathematical equations and empirical formulas (Neitsch et al. 2011; Arnold et al. 2012;).

3.2 Clustering of Homogeneous Basins

Regionalisation by physical similarity involves classifying basins to find at least one donor basin with physical characteristics similar to the target basin (Guo et al. 2021). We grouped homogeneous basins by testing different clustering techniques, including grouping items with similar characteristics that are frequently used, among other applications, to detect weather patterns (Aytaç 2020). We created homogeneous groups of basins using two of the best-known and most widely used techniques: *k*-means (Mokdad and Haddad 2017; Aytaç 2020; Lou et al. 2021) and EM (Di et al. 2019a, b; Asante-Okyere et al. 2020). Both techniques must predefine the number of clusters and are explained in Appendix A.

3.3 Hydrological Model Validation

In this study, we evaluated the SWAT model on a monthly basis following a two-phase strategy. In the first phase, we calibrated and validated the model in the donor basin of each cluster. The SWAT model simulation did not always produce satisfactory results by default, requiring calibration and validation using observed streamflow data (Swain et al. 2022). In this study, we performed an automatic calibration using SWAT calibration and uncertainty procedures (SWAT-CUP) software (Abbaspour 2012). In particular, we used the SUFI2 algorithm for calibration to identify the most influential parameters (i.e. sensitivity analysis) in the hydrological process and their optimum values. According to Arsenault et al. (2015), it is possible to reduce the parameters used in the model, decreasing its complexity with little or no loss in the regionalised model's performance. The period 2009–2010 was used for model warm-up, 2011–2015 for calibration, and 2016–2018 for validation. Once the model was calibrated and validated for each donor basin, we tested it in a second phase for the pseudo-ungauged basins of each cluster, predicting the streamflow for the period 2011–2018. The model's performance was evaluated in the two validation phases; we calculated the statistics listed in Table B1 (in Appendix B) by comparing the simulated streamflow with the streamflow observed in the stream gauges.

4 Results and Discussion

4.1 Clustering Homogeneous Subbasins

The parameters to be defined for the two clustering algorithms were the maximum number of iterations until there were no changes in the techniques and the number of optimal clusters (k). The number of iterations established was 10. Although we initially performed tests with values of 5, 7, 10, and 12, we determined that after 10 iterations, no variations or changes of instances occurred. We estimated the optimal value of k using two methods: dendrograms and the elbow method (Jain and Dubes 1988; Liu and Deng 2021; Aksan et al. 2021). A dendrogram is a plot derived from a hierarchical cluster which represents the data in the form of a tree that organises the data into subgroups that are divided until the desired level of detail is reached. The plot is created by forming clusters of observations, along with their levels of similarity, at each step. The level of similarity is measured on the horizontal axis, and the different observations are specified on the vertical axis. Fig. C1 in Appendix C gives the dendrogram for the available datasets.

The y-axis shows the instances, and the x-axis displays the clusters. Analysing the x-axis allows four levels with two, three, or four clusters to be differentiated. The levels with two clusters were the most evident. We confirmed this resultusing the elbow method, the results of which are displayed in Fig. C2. As seen in this figure, it was necessary to identify where the change of tendency in the curve is to determine the best number of groups. However, the options of three and four groups could also be valid solutions since the change in trend is gradual rather than abrupt, indicating no major differences between groups.

After we optimised the number of clusters and observed that the best version to use was groups of two clusters (although recognising that groups of three and four clusters were also valid) we performed the EM and k-means techniques using the Euclidean and Manhattan distances for the dataset. After 10 iterations, the solutions obtained were the same 70% of the time for any of the methods or distances. Figure 4 illustrates the results and makes it possible to identify the variation in clustering. Figure 4a shows the performance of the EM

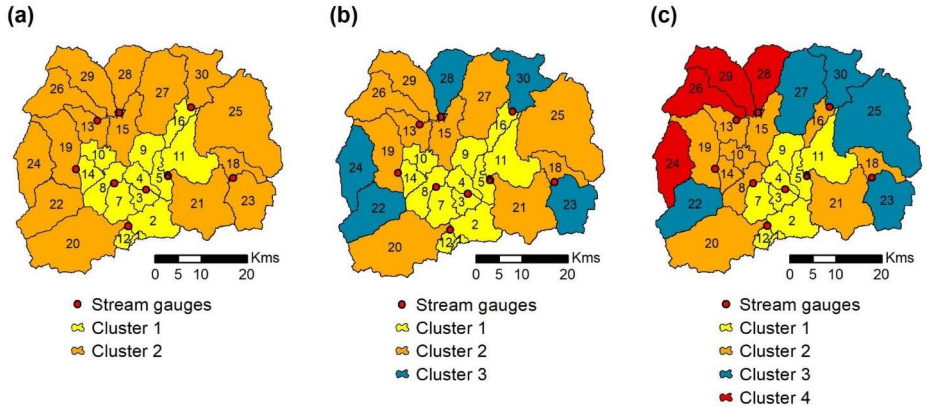

Fig. 4 Homogeneous zones obtained using different clustering techniques: (**a**) two clusters obtained using EM, (**b**) three clusters obtained using Euclidean k-means, and (**c**) four clusters obtained using Manhattan distance k-means. The points denote the location of the stream gauges

technique, and Fig. 4b and c present the results of the *k*-means technique using the Euclidean and Manhattan distances, respectively.

The subbasins were classified into different homogeneous regions or clusters, meaning that each cluster grouped the subbasins with homogeneous characteristics. The results were the same regardless of the technique or distance used in seven of the 10 runs. This finding indicates that the groups were sufficiently separated, with the core group (Cluster 1) remaining almost intact when running two, three, or four clusters. Cluster 1 obtained using the EM and Cluster 1 obtained using the Euclidean *k*-means contained the same subbasins. Cluster 1 obtained using *k*-means and the Manhattan distance were very similar, but Subbasins 8, 10, 14, and 16 became part of Cluster 2. This result indicates that the basins's characteristics are similar and that the clustering was satisfactory. Table B2 (see Appendix B) presents the mean values of the characteristics for each of the homogeneous regions identified using each clustering technique.

Each Cluster 1 includes subbasins with a smaller area, lower elevation, shallower streams, lower annual precipitation, and slightly higher mean annual temperatures than the other clusters. Cluster 3, obtained using Euclidean *k*-means, contains the most upstream subbasins (i.e. the basins at the highest altitude), in which precipitation is higher and the channels have a steeper gradient than in the other subbasins. Cluster 4, obtained using Manhattan distance *k*-means, contains the northwesternmost subbasins with the highest annual precipitation. The clustering results demonstrate that each cluster contains at least two gauged subbasins (see Table B3 in Appendix B).

4.2 SWAT Validation in the Donor Basins

Once the subbasins were clustered, we selected one subbasin in each cluster to be a donor, and the remaining subbasins became receptor basins. In this study, we analysed all possible combinations; that is, all of the gauged subbasins were donor basins at some point in time. Therefore, nine SWAT models, one for each donor subbasin, were developed, calibrated, and validated.

Before calibrating the SWAT model, we performed 500 model runs and selected sensitive parameters by performing a sensitivity analysis on 10 widely used parameters (Jimenez-Navarro et al. 2021; Jimeno-Sáez et al. 2021; Castellanos-Osorio et al. 2023) that can influence river streamflow (see Table B4 in Appendix B). In particular, we selected the SWAT parameters that obtained a *p*-value of less than 0.05 for each donor basin – the lower the *p*-value, the more sensitive the parameter. Table B4 lists the sensitive parameters selected for each subbasin. Each model was also designated using the acronym SB, followed by the subbasin number. The sensitive parameters identified were not the same in all nine cases. For example, the CN2 parameter, which determines the volume of surface runoff contributing to the total streamflow and depends on several factors related to the type of soils and their uses, was the only sensitive parameter for all donor subbasins. The parameters related to groundwater (i.e. GWQMN, GW_DELAY, RCHRG_DP, and GW_REVAP) were sensitive for eight of the nine models. These parameters have frequently been found to be the most sensitive in various studies conducted in different areas, as demonstrated by da Silva et al. (2018) in Brazil and Guo and Su (2019) on precipitation inputs from multiple sources in a Chinese basin. The SOL_AWC parameter was not used because it obtained a *p*-value

higher than 0.05 in all cases. This ranking of parameter sensitivity is supported by Raposo et al. (2013), Jimeno-Sáez et al. (2018), and Senent-Aparicio et al. (2019) in this region.

Automatic calibration was subsequently performed through 1,000 simulations to determine the optimal values of the sensitive parameters using the NSE statistic as the objective function. The optimal parameter values for each SWAT model are listed in Table 1. The small values of the percolation fraction (RCHRG_DP) reflect the fact that no aquifers in the area favour significant and lasting water storage. The ALPHA_BF values were less than 0.3, indicating soils with a slow response to recharge (Arnold et al. 2012). GW_DELAY, which is quantified in days, determines the recharge delay for shallow groundwater systems. Increasing this delay factor slows the recharge rate. The default GW_DELAY value of 31 days was reduced during calibration, indicating a rapid response of runoff to rainfall, as in Guse et al. (2014). In all of the subbasins, the ESCO value was high, which is expected for the Atlantic climate (Glavan et al. 2011; Ouallali et al. 2020). These optimum values are similar to those obtained in other studies performed in the area (Jimeno-Sáez et al. 2018; Senent-Aparicio et al. 2019).

Table 2 indicates the performance of the SWAT model during the calibration (2011–2015) and validation (2016–2018) periods in the nine donor basins. The SWAT performances indicate that the calibrated models satisfactorily simulated the monthly streamflow in donor basins. The R^2 statistic obtained values greater than 0.87 in all models. According to the criteria established by Moriasi et al. (2007), all models, except SB4 and SB6, were impressive, with NSE values above 0.75 and RSR values below 0.5. The SB4 and SB6 models obtained lower but nevertheless perfectly acceptable statistics.

As can be seen in Fig. C3 and Fig. C4 in Appendix C, the models were also validated graphically by comparing the simulated results with the observed streamflow. All models except SB6 tended to slightly underestimate peak flow events in the calibration and validation phases. The ineffectiveness of the SWAT model in achieving peak flows has previously been noted in multiple studies (Kim et al. 2015; Makwana and Tiwari 2017; Jimeno-Sáez et al. 2018; Blanco-Gómez et al. 2019). The hydrographs illustrate that the models calibrated in the donor basins simulated the streamflow very well, except for SB4 and SB6, where, although acceptable, the estimated streamflow differed significantly from the observed streamflow.

Table 1 Optimal values of SWAT parameters in donor basin models

Parameter	SB4	SB6	SB8	SB19	SB20	SB23	SB26	SB28	SB30
GWQMN	-319.25	948.25	288.25	57.25	295.75	-	-400.25	-311.75	-118.25
GW_DELAY	6.44	-	2.29	15.11	10.69	7.66	2.11	21.76	13.83
RCHRG_DP	0.04	-	0.08	0.13	0.18	0.21	0.001	0.18	0.06
ALPHA_BF	0.08	-	0.08	0.23	0.19	-	0.04	-	0.19
GW_REVAP	0.02	0.10	0.10	0.09	0.10	-	0.04	0.03	0.06
REVAPMN	-4.75	-192.20	-	-	-	-	-	-124.25	499.50
CN2	2.02	11.50	-1.54	3.98	-1.00	-3.10	1.00	12.70	-11.58
ESCO	0.99	0.78	1.00	-	-	0.69	0.99	0.99	0.93
EPCO	-	0.93	-	-	-	0.76	-	-	-

Table 2 Performance of the SWAT model in donor basins during the calibration and validation periods

Donor basin model	Period	R^2	NSE	RSR	RMSE
SB4	Calibration	0.94	0.71	0.54	23.19
	Validation	0.93	0.75	0.50	22.41
SB6	Calibration	0.88	0.72	0.53	5.13
	Validation	0.87	0.56	0.67	7.03
SB8	Calibration	0.95	0.92	0.28	7.30
	Validation	0.94	0.93	0.26	7.07
SB19	Calibration	0.95	0.95	0.23	1.98
	Validation	0.95	0.95	0.22	2.15
SB20	Calibration	0.94	0.94	0.25	1.20
	Validation	0.96	0.96	0.20	0.97
SB23	Calibration	0.90	0.84	0.40	0.86
	Validation	0.88	0.86	0.37	0.94
SB26	Calibration	0.96	0.96	0.20	0.58
	Validation	0.97	0.97	0.18	0.53
SB28	Calibration	0.90	0.89	0.34	0.83
	Validation	0.92	0.87	0.36	1.05
SB30	Calibration	0.93	0.82	0.42	1.47
	Validation	0.92	0.79	0.46	1.88

4.3 SWAT Validation in the Pseudo-ungauged Basins

After we calibrated and validated the SWAT models for the donor basins, we ensured that the models would operate in the study area and transferred the calibrated parameters to estimate streamflow in the pseudo-ungauged basins. The results of the streamflow simulation in the pseudo-ungauged basins for 2011–2018 are summarised in the following figures. The monthly hydrographs in Figs. 5 and 6 present the estimated streamflow in each of the pseudo-ungauged subbasins estimated using all of the models tested. The bar graphs display the errors obtained by comparing them with the observed streamflow.

The results are generally satisfactory. As can be seen, the R^2 values obtained in the streamflow estimations for all pseudo-ungauged basins are significantly high (i.e. above 0.82), indicating a high degree of collinearity between the estimated and observed streamflow and a low error variance. Based on the NSE statistic and according to the criteria established by Moriasi et al. (2007), all the models used to estimate streamflow in the pseudo-ungauged subbasins are excellent (NSE > 0.75) except in seven cases, of which only two experiments performed unsatisfactorily.

The RMSE values are low in all of the models, even in SB4, which obtained an RMSE value of 30.12 m^3/s and a monthly flow ranging between 2.3 and 157 m^3/s. For a more accurate measure of performance, RSR, which normalises the RMSE using the standard deviation of the observations, was also used. According to Moriasi et al.'s (2007) criteria, RSR values below 0.5 are excellent. Only six models obtained RSR values higher than 0.5, of which only two were unsatisfactory, as was the case with the NSE statistic. Hence, most of the models could be classified as excellent when the parameters were transferred to the pseudo-ungauged basins.

In all three Cluster 1s (see Table B3), the model calibrated with Donor Subbasin 4 (SB4) was classed as unsatisfactory (NSE < 0.5 and RSR > 0.70) when estimating streamflow in

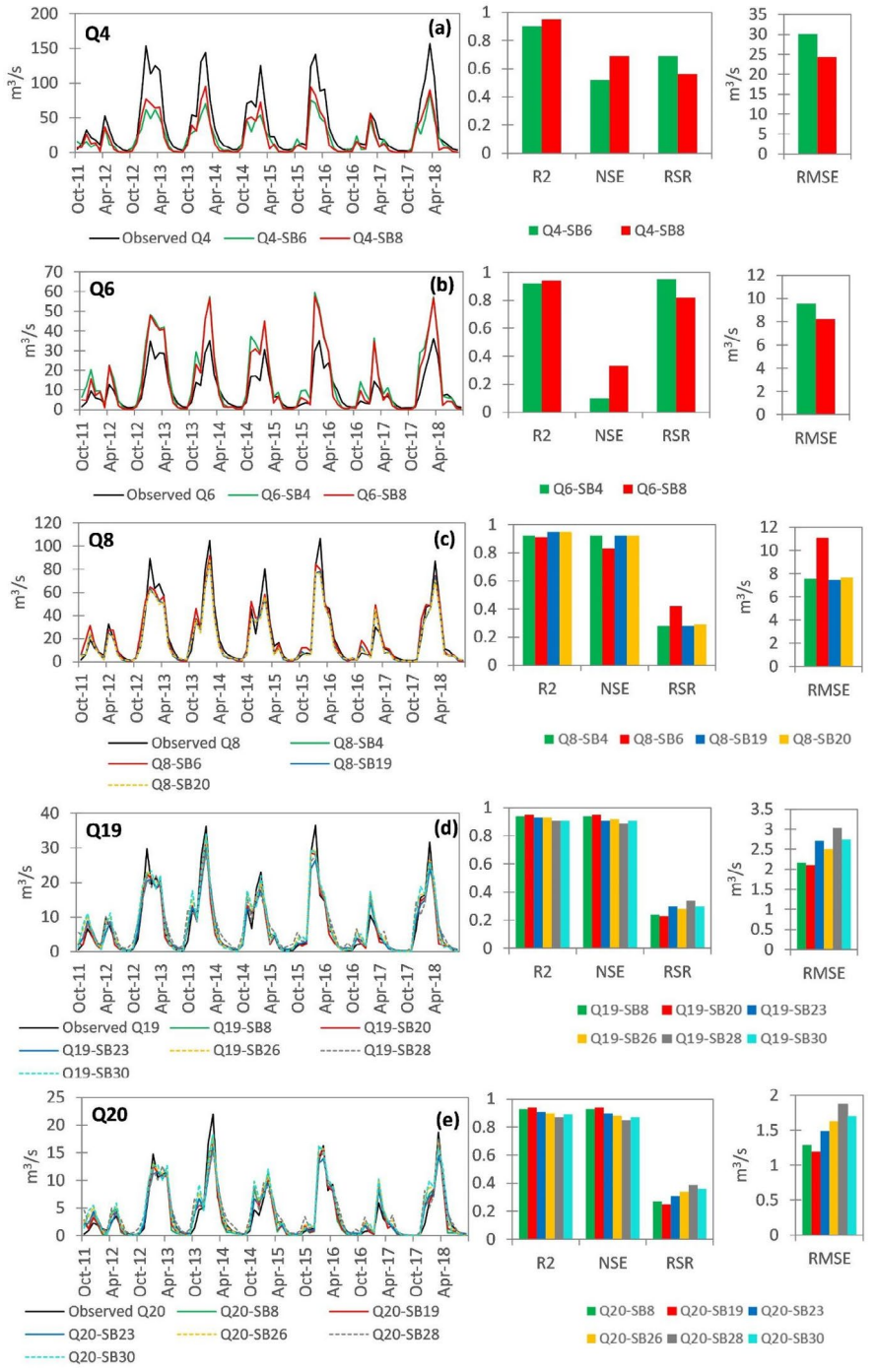

Fig. 5 Observed and simulated monthly streamflow and error graphs for pseudo-ungauged basins 4, 6, 8, 19, and 20 during the entire test period (2011–2018)

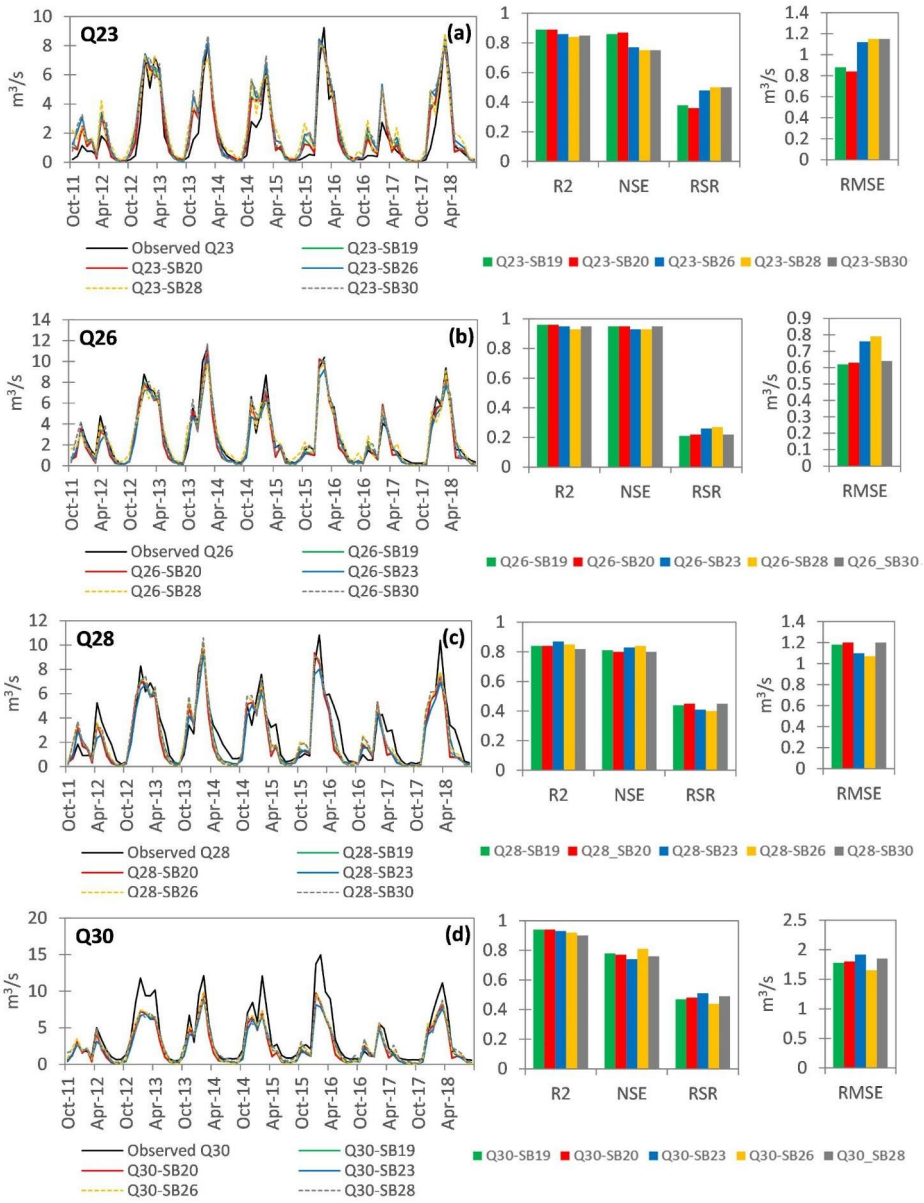

Fig. 6 Observed and simulated monthly streamflow and error graphs for pseudo-ungauged basins 23, 26, 28, and 30 during the entire test period (2011–2018)

Subbasin 6. In addition, in the Cluster 1 obtained using EM and Euclidean *k*-means, the model of Donor Subbasin 8 (SB8) was also unsatisfactory in estimating streamflow in Subbasin 6. The weakest results were related to stream gauges located in Subbasins 4, 6, and 8, which are located in the most downstream subbasins. Therefore, the observed data could be somewhat anthropised and differ more from the modelled data.

The results indicate that transferring the parameters calibrated in the donor basins to the pseudo-ungauged basins with homogeneous characteristics allowed for a highly accurate estimation of streamflow. Therefore, the semi-distributed SWAT model successfully estimated stremflow in the pseudo-ungauged basins, and it can be concluded that the clustering method was useful for estimating streamflow in the ungauged basins. Similar results have been obtained in other studies, such as those by Sellami et al. (2014), Swain and Patra (2017), Choubin et al. (2019), and Mosavi et al. (2021), who also found that the ungauged basins exhibited similar hydrological behaviour to the gauged basins within the same cluster and obtained impressive results in streamflow estimation by regionalisation combining clustering techniques with the SWAT model.

5 Conclusions

The results demonstrate the efficiency of transferring calibrated parameters to similar basins to estimate streamflow. The methodology is flexible and can be adapted to various clustering techniques. In addition, future research could explore additional hydrology-related features to potentially improve the methodology. The absence of a universal regionalisation method is attributed to differences in basin characteristics.

The main limitations of this study are data availability and quality, as well as the availability and length of observed streamflow data. Using a common measurement period can limit the study, and the methodology is only applicable when there is at least one gauged basin in the cluster of ungauged basins for data transfer.

Flow estimation in river basins is vital for various practical applications, such as the design of hydraulic structures, operation of hydroelectric power plants, water allocation for irrigation, industrial, and domestic use, flood and drought prediction, and environmental flow estimation. Therefore, the results of this study are significant for water resource management and planning, not only in the basin studied but also in other regions where a similar approach can be applied. This methodology also constitutes a valuable and effective tool for risk and environmental managers, as it contributes to decision-making in the management of ungauged basins lacking observed hydrological data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11269-023-03678-8.

Acknowledgements The authors acknowledge Scribbr editing services for proofreading the text.

Author Contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by P.J-S., R.M-E. and J.S-A. The first draft of the manuscript was written by P.J-S. and J.S-A and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the Spanish Ministry of Science and Innovation, under grants RTC-2017-6389-5 and the European Union´s Horizon 2020 research and innovation programme within the framework of the project SMARTLAGOON, grant agreement number 101017861.

Data Availability Data are available from the corresponding author upon requests.

Declarations

Ethical Approval The authors state that this work complies with the journal guidelines on ethical issues.

Consent to Participate All the authors have given explicit consent to participate in the manuscript.

Consent to Publish All the authors have given explicit consent to publish this manuscript.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Abbaspour KC (2012) SWAT calibration and uncertainty Program—A user manual; SWAT-CUP-2012, 2012th edn. Swiss Federal Institute of Aquatic Science and Technology, Dubendorf, Switzerland

Aksan F, Jasiński M, Sikorski T et al (2021) Clustering methods for power quality measurements in Virtual Power Plant. Energies 14:5902. https://doi.org/10.3390/en14185902

Al-Khafaji M, Saeed FH, Al-Ansari N (2020) The interactive impact of Land Cover and DEM Resolution on the Accuracy of computed streamflow using the SWAT model. Water Air Soil Pollut 231:416. https://doi.org/10.1007/s11270-020-04770-0

Ali I, Singh P, Aboul-Enein HY, Sharma B (2009) Chiral analysis of ibuprofen residues in water and sediment. Anal Lett 42(12):1747–1760. https://doi.org/10.1080/00032710903060

Arnold JG, Srinivasan R, Muttiah RS, Williams JR (1998) Large area hydrologic modeling and assessment part I: model development. J Am Water Resour Assoc 34:73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x

Arnold JG, Kiniry JR, Srinivasan R et al (2012) SWAT 2012 Input/Output Documentation

Arsenault R, Poissant D, Brissette F (2015) Parameter dimensionality reduction of a conceptual model for streamflow prediction in Canadian, snowmelt dominated ungauged basins. Adv Water Resour 85:27–44. https://doi.org/10.1016/j.advwatres.2015.08.014

Asante-Okyere S, Shen C, Ziggah YY et al (2020) A novel hybrid technique of integrating gradient-boosted machine and clustering algorithms for Lithology classification. Nat Resour Res 29:2257–2273. https://doi.org/10.1007/s11053-019-09576-4

Aytaç E (2020) Unsupervised learning approach in defining the similarity of catchments: hydrological response unit based k-means clustering, a demonstration on Western Black Sea Region of Turkey. Int Soil Water Conserv Res 8:321–331. https://doi.org/10.1016/j.iswcr.2020.05.002

Balha A, Singh A, Pandey S. et al (2023) Assessing the impact of land-use dynamics to predict the changes in hydrological variables using effective impervious area (EIA). Water Resour Manage 37:3999–4014. https://doi.org/10.1007/s11269-023-03536-7

Barbarossa V, Huijbregts MAJ, Hendriks AJ, et al (2017) Developing and testing a global-scale regression model to quantify mean annual streamflow. J Hydrol 544:479–487. https://doi.org/10.1016/j.jhydrol.2016.11.053

Basheer AA (2018a) Chemical chiral pollution: impact on the society and science and need of the regulations in the 21st century. Chirality 30(4):402–406. https://doi.org/10.1002/chir.22808

Basheer AA (2018b) New generation nano-adsorbents for the removal of emerging contaminants in water. J Mol Liq 261:583-593

Beza M, Hailu H, Teferi, G (2023) Modeling and Assessing Surface Water Potential Using Combined SWAT Model and Spatial Proximity Regionalization Technique for Ungauged Subwatershed of Jewuha Watershed, Awash Basin, Ethiopia. Adv Civ Eng 2023. https://doi.org/10.1155/2023/9972801

Blanco-Gómez P, Jimeno-Sáez P, Senent-Aparicio J, Pérez-Sánchez J (2019) Impact of Climate Change on Water Balance Components and droughts in the Guajoyo River Basin (El Salvador). Water 11:2360. https://doi.org/10.3390/w11112360

Blöschl G, Bierkens MFP, Chambel A et al (2019) Twenty-three unsolved problems in hydrology (UPH) – a community perspective. Hydrol Sci J 64:1141–1158. https://doi.org/10.1080/02626667.2019.1620507

Castellanos-Osorio G, López-Ballesteros A, Pérez-Sánchez J, Senent-Aparicio J (2023) Disaggregated monthly SWAT + model versus daily SWAT + model for estimating environmental flows in Peninsular Spain. J Hydrol 623:129837. https://doi.org/10.1016/j.jhydrol.2023.129837

Cheng X, Ma X, Wang W et al (2021) Application of HEC-HMS parameter regionalization in small watershed of hilly area. Water Resour Manag 35:1961–1976. https://doi.org/10.1007/s11269-021-02823-5

Choubin B, Solaimani K, Rezanezhad F et al (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. https://doi.org/10.1016/j.catena.2019.104128

da Silva RM, Dantas JC, Beltrão JDA, Santos CA (2018) Hydrological simulation in a tropical humid basin in the Cerrado biome using the SWAT model. Hydrol Res 49:908–923

Darko S, Adjei KA, Gyamfi C et al (2021) Evaluation of RFE Satellite Precipitation and its use in Streamflow Simulation in Poorly Gauged basins. Environ Process 8:691–712. https://doi.org/10.1007/s40710-021-00495-2

Di Z, Chang M, Guo P et al (2019a) Using real-Time Data and Unsupervised Machine Learning techniques to study large-scale spatio–temporal characteristics of Wastewater discharges and their influence on Surface Water Quality in the Yangtze River Basin. Water 11:1268. https://doi.org/10.3390/w11061268

Di Blasi JIP, Martínez Torres J, García Nieto PJ et al (2013) Analysis and detection of outliers in water quality parameters from different automated monitoring stations in the Miño river basin (NW Spain). Ecol Eng 60:60–66. https://doi.org/10.1016/j.ecoleng.2013.07.054

Eguibar MÁ, Porta-García R, Torrijo FJ, Garzón-Roca J (2021) Flood hazards in flat Coastal areas of the Eastern Iberian Peninsula: a Case Study in Oliva (Valencia, Spain). Water 13:2975. https://doi.org/10.3390/w13212975

Farsadnia F, Rostami Kamrood M, Moghaddam Nia A et al (2014) Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps. J Hydrol 509:387–397. https://doi.org/10.1016/j.jhydrol.2013.11.050

Gao M, Chen X, Liu J, Zhang Z (2018) Regionalization of annual runoff characteristics and its indication of co-dependence among hydro-climate–landscape factors in Jinghe River Basin, China. Stoch Environ Res Risk Assess 32:1613–1630. https://doi.org/10.1007/s00477-017-1494-9

Gebeyehu BM, Tegegne G, Melesse AM (2023) Reliability-weighted approach for streamflow prediction at ungauged catchments. J Hydrol 624:129935. https://doi.org/10.1016/j.jhydrol.2023.129935

Glavan M, White S, Holman IP (2011) Evaluation of river water quality simulations at a daily time step–experience with SWAT in the Axe Catchment, UK. Clean–Soil Air Water 39(1):43–54

Guo J, Su X (2019) Parameter sensitivity analysis of SWAT model for streamflow simulation with multi-source precipitation datasets. Hydrol Res 50(3):861–877

Guo Y, Zhang Y, Zhang L, Wang Z (2021) Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: a comprehensive review. Wiley Interdiscip Rev-Water 8. https://doi.org/10.1002/wat2.1487

Guse B, Reusser DE, Fohrer N (2014) How to improve the representation of hydrological processes in SWAT for a lowland catchment–temporal analysis of parameter sensitivity and model performance. Hydrol Process 28(4):2651–2670

Hrachowitz M, Savenije HHG, Blöschl G et al (2013) A decade of predictions in Ungauged basins (PUB)—a review. Hydrol Sci J 58:1198–1255. https://doi.org/10.1080/02626667.2013.803183

Jain AK, Dubes RC (1988) Algorithms for Clustering Data. Prentice Hall

Jiménez-Navarro IC, Jimeno-Sáez P, López-Ballesteros A, Pérez-Sánchez J, Senent-Aparicio J (2021) Impact of Climate Change on the hydrology of the forested Watershed that drains to Lake Erken in Sweden: an analysis using SWAT + and CMIP6 scenarios. Forests 12:1803. https://doi.org/10.3390/f12121803

Jimeno-Sáez P, Senent-Aparicio J, Pérez-Sánchez J, Pulido-Velazquez D (2018) A comparison of SWAT and ANN Models for Daily Runoff Simulation in different climatic zones of Peninsular Spain. Water 10:192. https://doi.org/10.3390/w10020192

Jimeno-Sáez P, Blanco-Gómez P, Pérez-Sánchez J, Cecilia JM, Senent-Aparicio J (2021) Impact Assessment of Gridded Precipitation products on Streamflow Simulations over a poorly gauged Basin in El Salvador. Water 13:2497. https://doi.org/10.3390/w13182497

Jodar-Abellan A, Ruiz M, Melgarejo J (2018) Climate change impact assessment on a hydrologic basin under natural regime (SE, Spain) using a SWAT model. Revista Mexicana De Ciencias Geológicas 35(3):240–253. https://doi.org/10.22201/cgeo.20072902e.2018.3.564

Kim M, Baek S, Ligaray M et al (2015) Comparative studies of different imputation methods for recovering Streamflow Observation. Water 7:6847–6860. https://doi.org/10.3390/w7126663

Liu F, Deng Y (2021) Determine the number of unknown targets in Open World based on Elbow Method. IEEE Trans Fuzzy Syst 29:986–995. https://doi.org/10.1109/TFUZZ.2020.2966182

Lou D, Yang M, Shi D et al (2021) K-Means and C4.5 decision Tree Based Prediction of Long-Term Precipitation variability in the Poyang Lake Basin, China. Atmosphere 12:834. https://doi.org/10.3390/atmos12070834

Makwana JJ, Tiwari MK (2017) Hydrological stream flow modelling using soil and water assessment tool (SWAT) and neural networks (NNs) for the Limkheda watershed, Gujarat, India. Model Earth Syst Environ 3:635–645. https://doi.org/10.1007/s40808-017-0323-y

Mokdad F, Haddad B (2017) Improved infrared precipitation estimation approaches based on k-means clustering: application to north Algeria using MSG-SEVIRI satellite data. Adv Space Res 59:2880–2900. https://doi.org/10.1016/j.asr.2017.03.027

Moriasi DN, Arnold JG, Liew MWV et al (2007) Model evaluation guidelines for systematic quantification of Accuracy in Watershed simulations. Trans ASABE 50:885–900. https://doi.org/10.13031/2013.23153

Mosavi A, Golshan M, Choubin B et al (2021) Fuzzy clustering and distributed model for streamflow estimation in ungauged watersheds. Sci Rep 11:8243. https://doi.org/10.1038/s41598-021-87691-0

Nachtergaele FO, van Velthuizen H, Verelst L et al (2008) Harmonized world soil database. Food and Agriculture Organization of the United Nations, Rome, Italy

Neitsch SL, Arnold JG, Kiniry JR, Williams JR (2011) SWAT Theoretical Documentation

Ouallali A, Briak H, Aassoumi H et al (2020) Hydrological foretelling uncertainty evaluation of water balance components and sediments yield using a multi-variable optimization approach in an external Rif's catchment. Morocco Alex Eng J 59(2):775–789

Ramachandra Rao A, Srinivas VV (2006) Regionalization of watersheds by hybrid-cluster analysis. J Hydrol 318:37–56. https://doi.org/10.1016/j.jhydrol.2005.06.003

Raposo JR, Dafonte J, Molinero J (2013) Assessing the impact of future climate change on groundwater recharge in Galicia-Costa, Spain. Hydrogeol J 21:459–479

Razavi T, Coulibaly P (2013a) Streamflow Prediction in Ungauged basins: review of regionalization methods. J Hydrol Eng 18:958–975

Razavi T, 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

Sellami H, La Jeunesse I, 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

Senent-Aparicio J, Jimeno-Sáez P, Bueno-Crespo A et al (2019) Coupling machine-learning techniques with SWAT model for instantaneous peak flow prediction. Biosyst Eng 177:67–77. https://doi.org/10.1016/j.biosystemseng.2018.04.022

Senent-Aparicio J, Jimeno-Sáez P, López-Ballesteros A et al (2021) Impacts of swat weather generator statistics from high-resolution datasets on monthly streamflow simulation over Peninsular Spain. J Hydrol-Reg Stud 35:100826. https://doi.org/10.1016/j.ejrh.2021.100826

Senent-Aparicio J, López-Ballesteros A, Jimeno-Sáez P, Pérez-Sánchez J (2023) Recent precipitation trends in Peninsular Spain and implications for water infrastructure design. J Hydrol-Reg Stud 45:101308. https://doi.org/10.1016/j.ejrh.2022.101308

Singh PK, Kumar V, Purohit RC et al (2009) Application of principal component analysis in Grouping Geomorphic parameters for Hydrologic modeling. Water Resour Manage 23:325–339. https://doi.org/10.1007/s11269-008-9277-1

Singh L, Mishra PK, Pingale SM et al (2022) Streamflow regionalisation of an ungauged catchment with machine learning approaches. Hydrol Sci J 67:886–897. https://doi.org/10.1080/02626667.2022.2049271

Sisay E, Halefom A, Khare D et al (2017) Hydrological modelling of ungauged urban watershed using SWAT model. Model Earth Syst Environ 3:693–702. https://doi.org/10.1007/s40808-017-0328-6

Sivapalan M, Takeuchi K, Franks SW 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

Smakhtin VU (2001) Low flow hydrology: a review. J Hydrol 240:147–186. https://doi.org/10.1016/S0022-1694(00)00340-1

Srinivasan R, Zhang X, Arnold J (2010) SWAT ungauged: Hydrological Budget and Crop yield predictions in the Upper Mississippi River Basin. Trans ASABE 53:1533–1546. https://doi.org/10.13031/2013.34903

Ssegane H, Tollner EW, Mohamoud YM et al (2012) Advances in variable selection methods II: effect of variable selection method on classification of hydrologically similar watersheds in three Mid-atlantic ecoregions. J Hydrol 438–439:26–38. https://doi.org/10.1016/j.jhydrol.2012.01.035

Swain JB, Patra KC (2017) Streamflow estimation in ungauged catchments using regionalization techniques. J Hydrol 554:420–433. https://doi.org/10.1016/j.jhydrol.2017.08.054

Swain S, Mishra SK, Pandey A et al (2022) Hydrological modelling through SWAT over a himalayan catchment using high-resolution geospatial inputs. Environ Challenges 8:100579. https://doi.org/10.1016/j.envc.2022.100579

Trenberth KE, Smith L, Qian T et al (2007) Estimates of the Global Water Budget and its annual cycle using Observational and Model Data. J Hydrometeorol 8:758–769. https://doi.org/10.1175/JHM600.1

Wu H, Zhang J, Bao Z et al (2022) Runoff modeling in Ungauged catchments using machine learning algorithm-based model parameters regionalization methodology. https://doi.org/10.1016/j.eng.2021.12.014. Engineering S2095809922000613

Yadav M, Wagener T, Gupta H (2007) Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. Adv Water Resour 30:1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Javier Senent-Aparicio[1] · Patricia Jimeno-Sáez[1] · Raquel Martínez-España[2,3] · Julio Pérez-Sánchez[1,4]

✉ Patricia Jimeno-Sáez
pjimeno@ucam.edu

Javier Senent-Aparicio
jsenent@ucam.edu

Raquel Martínez-España
raquel.m.e@um.es

Julio Pérez-Sánchez
julio.sanchez@ulpgc.es

1 Department of Civil Engineering, Universidad Católica San Antonio de Murcia, Campus de Los Jerónimos s/n, Guadalupe, Murcia 30107, Spain

2 Department of Computer Engineering, Universidad Católica San Antonio de Murcia, Campus de Los Jerónimos s/n, Guadalupe, Murcia 30107, Spain

3 Department of Information and Communication Engineering, Universidad de Murcia, Espinardo, Murcia 30100, Spain

4 Department of Civil Engineering, Universidad de Las Palmas de Gran Canaria, Campus de Tafira, Las Palmas de Gran Canaria 35017, Spain