Modeling Earth Systems and Environment (2021) 7:341–350
https://doi.org/10.1007/s40808-020-01033-z

ORIGINAL ARTICLE

Regionalization of mean annual flow for ungauged catchments in case of Abbay River Basin, Ethiopia

Habtamu Nega[1,2] · Yilma Seleshi[1]

Received: 15 August 2020 / Accepted: 17 October 2020 / Published online: 31 October 2020

Abstract

Streamflow measurement is one of the features of Ethiopia and other developing countries suffer to estimate. Climate variable (rainfall) and physiographic variables (catchment area, land slope, and elevation) were selected to develop a regression model equation that can be used to estimate the mean annual flow for ungauged catchments in the Abbay river basin. 70% of 27 hydrometric stations were used to train the regression model, and the remaining 30% were used for validation. Furthermore, statistical tests were used for selecting the best trustworthy model. The result shows that both catchment area (A) as the only predictor with both coefficients of determination (R^2) and Nash–Sutcliffe efficiency (NSE) value of 0.93 and catchment area (A), mean annual rainfall (MAR), and the average elevation (AE) as predictors with both coefficients of determination (R^2) and Nash–Sutcliffe efficiency (NSE) value of 0.96 regression equations were the trustworthy models. Moreover, the validation analysis was performed for nine stations distributed in the study area and resulting from a statistical test and graphical visualization the second regression model equation was a trustworthy model. The model parameters (catchment characteristics) are easily accessible for practitioners that want to estimate the mean annual flow in the study area.

Keywords Mean annual flow · Regionalization · Regression analysis · Ungauged catchments

Introduction

One of the world's most important and vital natural resources is water (Zigaf and Eriksson 2018) Water is one of human being's basic need that has its involvement in day-to-day activities all over the world. Water resource planning and projects have been implemented for centuries to accomplish the need of the people in the way of electric power generation, domestic water use, irrigation, recreation, as well as transportation. Moreover, the water resource in the world is its effort on life and the existence of wildlife and ecosystem. Eventually, water has a huge devastating effect that may damage property and the loss of the irreplaceable life of humans by flood, drought, and waves in the oceans. This usefulness and consequence of water in the world to some degree are controlled by the design and management of different water resource schemes.

The design and management of hydraulic structures are done by the help of hydrometeorological gauges which record different hydrological information in the basins such as streamflow, precipitation, and water level. However, coverage of hydrometeorological gauging stations are a vast problem all over the world, especially in developing countries like Ethiopia, which is the tower of water resources in the East Africa composed of 12 river basins. Additionally, insufficient record period is also a problem even in gauged catchments. According to Tamalew and Kemal (2016), the situation in Ethiopia is problematic, as there are unevenly distributed hydrometric stations, and large areas lack gauging stations compounded by the lack of sufficient data.

Razavi et al. (2013) and Sivapalan et al. (2003) stated that ungauged catchment or poorly gauged catchment definition depends on the variable from the catchment we want to practice. Javeed and Apoorva (2015) indicated that streamflow measurements are important to obtain climate

✉ Habtamu Nega
Habtamushafi@gmail.com

Yilma Seleshi
Yilma.Seleshi@gmail.com

[1] School of Civil and Environmental Engineering, Addis Ababa Institute of Technology, Addis Ababa University, Addis Ababa, Ethiopia

[2] Department of Hydraulic and Water Resources Engineering, Faculty of Technology, Debre Tabor University, Debre Tabor, Ethiopia

and/or land-use change for future assessment of the hydrological behaviour of river basins. According to Peter and Woo (1982) spatial and temporal variability of mean annual flow is prerequisite for the use and development of water resources. Additionally, Vogel et al. (1999) stated that to obtain responsible management, estimates of annual watershed runoff volumes are required.

Planning and design of water resource projects in an ungauged catchment are solved by regional analysis (Yu et al. 2005). Regional analysis is an appropriate approaches for the prediction of ungauged basins (PUB) in the International Association of Hydrological Sciences (IAHS) (Sivapalan et al. 2003). Regionalization approaches generally can be divided into hydrologic-dependent model and hydrologic-independent model. The first one is about the transferring of rainfall-runoff model parameters between basins. However, for the second approach, the equation and its parameters are transferred (Razavi et al. 2013). Among the hydrologic-independent model methods for streamflow regionalization (regression-based, time-series model, and scaling relationships), regression equations (including linear and non-linear) developed between the desired hydrologic responses and catchment attributes are the most commonly used ones. One of the advantages of hydrologic model-independent methods is the lower data requirement and the simplicity of their structure that does not require special knowledge and expertise of hydrological modelling.

The regression model equation relates streamflow and catchment characteristics in an explanatory way that helps to understand general hydrological patterns and processes adopted across different scales. Global-scale (Barbarossa et al. 2017; Burgers et al. 2013) and regional-scale (Tran et al. 2015; Papamichail et al. 2002; Vogel et al. 1999; Sedighi 2008; Zigaf and Eriksson 2018) regression model approaches for mean annual flow using catchments characteristics were developed. However, in Ethiopia, the regionalization study is focused on flood frequency analysis (Abebe et al. 2013; Gebeyehu 1989) using a rainfall-runoff model (Wale et al. 2009; Berhane 2013; Tamalew and Kemal 2016; Tesfaye 2011).

The study aims to develop a regional empirical equation for ungauged catchments using gauged catchments in the Abbay river basin of Ethiopia using regression models that consider different characteristics of catchments.

Methodology

Description of study area

Abbay river basin is located in the northwestern region of Ethiopia between 7° 40′ N and 12° 51′ N latitude, and 34°25′ E and 39° 49′ E longitude. It originates from the centre of the catchment (Sekela, West Gojam), flows to the north into Lake Tana as Gilgel Abbay. Then, it drains out from the southeast of Lake Tana and flows through a deep gorge. Throughout the course, this river receives from Beshilo, Welaka, Jemma, Muger, Fincha, Guder, Didessa, and Dabus from the East and South; and the Abeya, Suha, Chemoga, Birr, Fettam, Dura, and Beles from the North and the West. The Dinder and Rahad rise to the West of the Lake Tana and flow westwards joining the Abbay River after crossing the border. It is the second largest basin next to the Wabisheble River Basin which has a catchment area of (199,812 km^2), covering parts of Amhara, Oromia, and Benishangul-Gumuz regional states. It has the sixteen sub-basins viz., Anger, Beles, Beshilo, Dabus, Didesa, Dindir, Fincha, Guder, Jemma, Lake Tana, Muger, North Gojam, Rahad, South Gojam, Welaka, and Wonbera as shown in Fig. 1.

Data collection

In this study, mainly three datasets were collected. The rainfall data were collected from the National Meteorological Agency (NMA). A monthly record of 32 rainfall stations was employed in this study. A monthly streamflow data of 27 hydrometric stations were collected from the Ministry of water, irrigation, and energy (MoWIE), Hydrology Directorate. Table 1 and Fig. 2 summarize the spatial distributions of hydrometric stations used in the study. These stations were split into two datasets. The first one was for the development of equations or training datasets, and the second one was for validation of developed equations or test datasets.

The splitting processes were done in R studio in which 70% of the data were used for training and the remaining 30% for validation. Quality assessment was done for both hydrometric and meteorological datasets. Catchment physiographic data (Area, Slope, and Elevation) were extracted from 30 m Digital Elevation Model (DEM) using Arc GIS 10.4 software.

Method of analysis

Regression analysis estimates a relationship between the dependent and one or more independent variables. Sedighi (2008) stated that “a model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation” and if the model is passed a different test, then the model can be used to forecast dependent variable for a value of different independent variables. Therefore, the analysis was started by choosing independent variables or predictors based on the influence on runoff generation documented in several studies (Barbarossa et al. 2017; Burgers et al. 2013; Gebeyehu 1989; Razavi et al. 2013; Zigaf and Eriksson 2018).

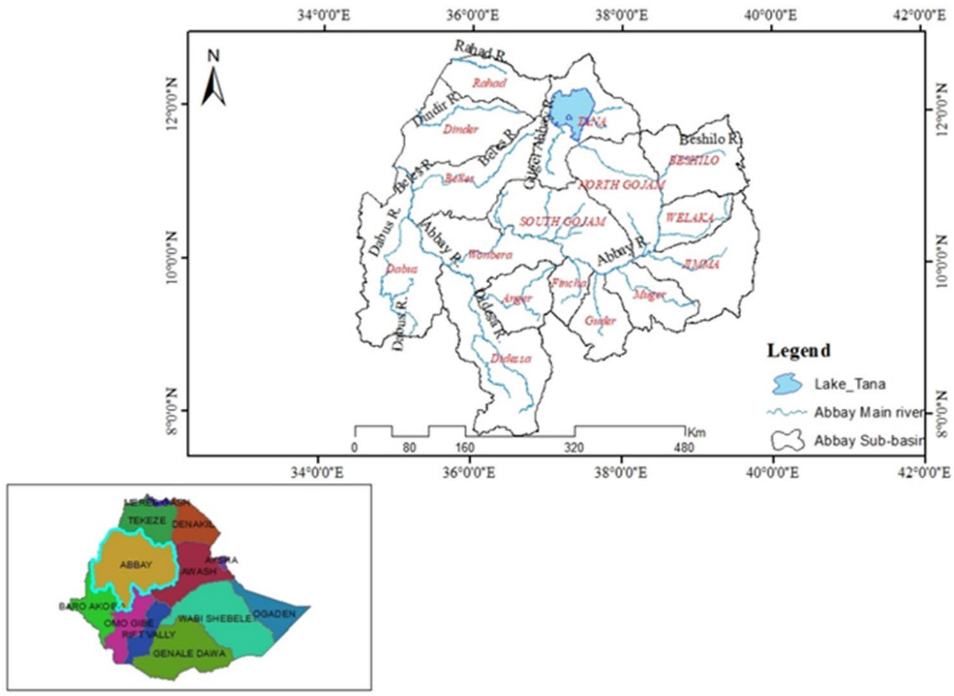

Fig. 1 Location of the study area

Table 1 Hydrometric stations used for the analysis

No.	Hydrometric station	Lat	Long	No.	Hydrometric station	Lat	Long
1	Abbay Nr. Kessie	11.07	38.18	15	Gula at Dembecha	10.55	37.5
2	Abbay at Bahir Dar	11.6	37.4	16	Gumara Nr. Bahir Dar	11.83	37.63
3	Andassa Nr. Bahir Dar	11.5	37.48	17	Hoha Nr. Assosa	10.15	34.63
4	Angar Nr. Nekemete	9.43	36.52	18	Jedeb Nr. Amanuel	10.4	37.57
5	Ardy Nr. Metekel	10.95	36.52	19	Koga at Merawi	11.37	37.05
6	Azuari Nr. Mota	10.97	38.02	20	Lower Fettam at Galibed	10.48	37.02
7	Belo Nr. Guder	8.87	37.67	21	Megech Nr. Azezo	12.48	37.45
8	Beressa Nr. Debre Berehan	9.67	39.52	22	Muger Nr. Chancho	9.3	38.73
9	Birr Nr. Jiga	10.65	37.38	23	Neshi Nr. Shambu	9.75	37.25
10	Chemoga Nr. Debre Markos	10.3	37.73	24	Robigumer Nr. Lemi	9.75	39
11	Didessa Nr. Arjo	8.68	36.42	25	Shina Nr. Adiet	11.25	37.5
12	Dura Nr. Metekel	10.98	36.48	26	Temcha Nr. Dembecha	10.53	37.5
13	Gelgel Abbay Nr. Merawi	11.37	37.03	27	Wenka Nr. Estie	11.62	38.07
14	Guder at Guder	8.95	37.75	28			

Climatic variable (mean annual rainfall) and three physiographic variables (catchment area, average slope, and average elevation) were selected for performing the regression analysis.

Once the choice of predictor variables is accomplished, the next steps in regression analysis are the determination of any linear relationships between selected independent variables. The reason behind this is that, in the case of multiple linear regression analysis, it will be difficult to know which variable affects the model if there is a relationship (high correlation coefficient) between predictors and the output model may even be biased if there is significant correlation between the independent variables (Gebeyehu 1989). As shown in Table 2 and Fig. 3, there is no high correlation between predictors.

Model development

Regression analysis is divided into two as simple and multiple linear regression analyses based on the number of predictor variables. In this study, the regression model was developed by initially taking area (A) as the only independent variable and then adding mean annual rainfall (MAR) as a predictor. Finally, a stepwise regression analysis was

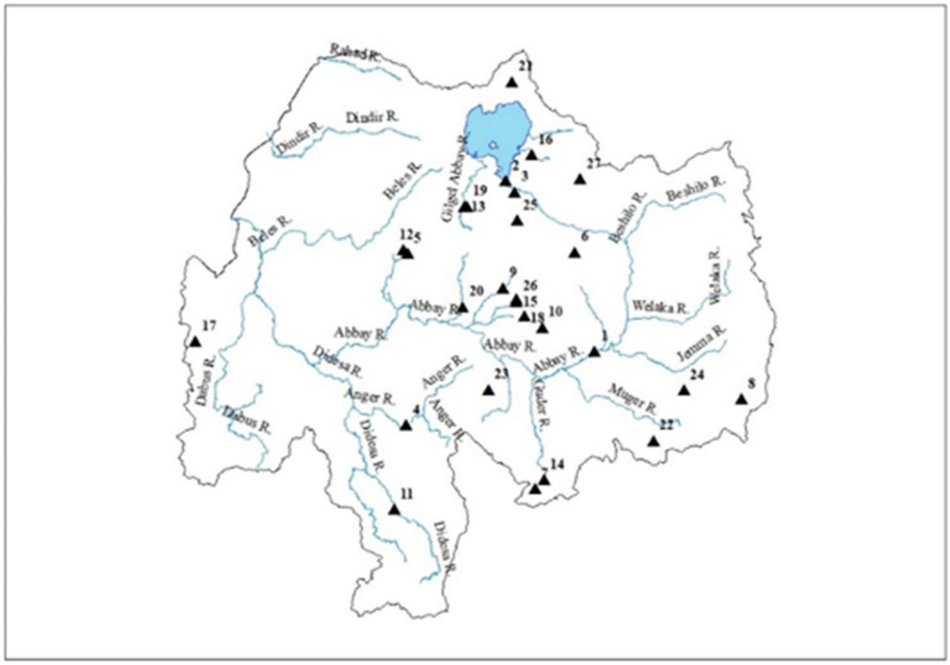

Fig. 2 Location and spatial distributions of selected hydrometric stations

Table 2 Correlation matrix of variables

	Area	MAR	AE	AS
Area	1			
MAR	0.096	1		
AS	0.085	0.27	1	
AE	0.28	0.53	0.11	1

MAR is mean annual rainfall, AE is average elevation, and AS is the average slope

performed until a good result was achieved. Hence, this model was developed by considering runoff as a function of combinations of independent variables whichever gives good results.

$$MAF = f\{A, \text{MAR}, \text{AS} \,\&\, \text{AE}\}$$

$$\text{MAF} = \alpha A^{\beta} * (\text{MAR})^{\gamma} * (\text{AS})^{\delta} * (\text{AE})^{\mu}$$

where MAF is mean annual flow (Mm^3), A is catchment area (Km^2), MAR is mean annual rainfall (mm), AS is average slope (percent), AE is the average elevation (m), and α, β, γ, δ, and μ are coefficients. The above equation can be changed to linear by taking logarithm in both directions, and it can be re-written as follows:

$$\log\text{MAF} = \log\alpha + \beta * \log A + \gamma * \log\text{MAR} + \delta * \log\text{AS} + \mu * \log\text{AE}.$$

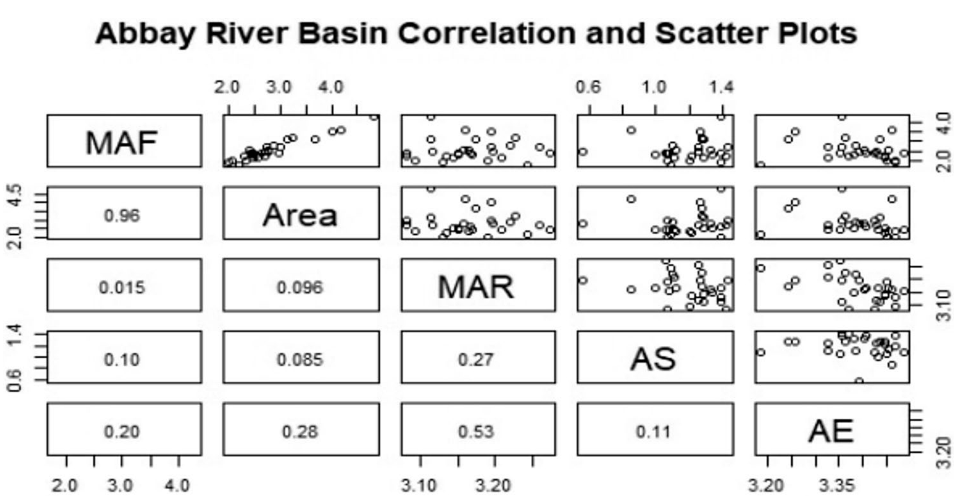

Fig. 3 Abbay basin scatter plot with correlation coefficients of variables

Model fitting

After model development, the coefficient needs to be determined. This task was done by selecting the most common ordinary least squares (OLS) as model fit, because it results in an explicit equation, which facilitates interpretation and comparison with other studies (Barbarossa et al. 2017).

Model criticism and selection

Next to the coefficient determination of the model parameters, the adequacy of the model was checked by performing subsequent statistical tests.

Hypothesis tests Hypothesis testing is a scientific process of testing whether or not the hypothesis is believable. The significance of the regression equation was checked by testing the hypothesis of the null value of slope $\beta 1$ for simple regression and $\beta 1$ to βn for multiple regression analyses. The steps of a hypothesis test are the development of null and alternative hypothesis followed by setting the significance level α (0.05 in this study) and calculating test statistics and probability value (p-value). Finally, they make decision based on the calculated and tabulated tests and conclude it. So, in this study, the hypothesis made to check the significance of the model is.

H0: $\beta 1 = 0$ for simple regression or $\beta 1 = \ldots\ldots = \beta n = 0$ for multiple regression;

Ha: $\beta 1 \neq 0$ for simple regression or $\beta k \neq 0$ at least one k; for multiple regression.

For simple linear regression, the test statistics used to check the significance of the regression equation is a t test. The test statistic t is constructed by comparing the slope of the estimated and true regression line, as normalized by the standard error on the estimate of $\beta 1$ (Mauro Naghettini et al. 2017).

The calculated t value was compared with the critical t value using two-tailed ($\alpha/2$) and n-2 degree of freedom. The null hypothesis was rejected for $|t| > t_{\frac{\alpha}{2},n-2}$.

Alternatively, the significance of the regression equation was evaluated using an analysis of variance (ANOVA). For multiple linear regression, the significance of the regression equation was tested using the overall F test. In this situation, the main difference with simple linear regression F test is that due to the number of independent variables increment, the degree of freedom of the sum of squares due to regression and residual sum of squares are j and $n - j - 1$, respectively, where j is the number of independent variables. If $|F| > F_{\alpha,j,n-j-1}$, the null hypothesis was rejected. However, in this study for multiple linear regression analysis, the decision is not only made by F test, but also the probability value (p-value) of each incorporated independent variable with a significant level (α) was compared, and if the p-value is greater than α, the hypothesis as well as the combination of the independent variable is accepted.

Coefficient of determination (R^2) including adjusted (R^2_{adj}) One of the important tests for examining the quality of the model is with the coefficient of determination also called goodness-of-fit index or as the most commonly used designation R^2. However, in multiple linear regression, the addition of the independent variable will increase the value of R^2 which can be leading to the wrong conclusion. To overcome this problem, the modified or adjusted R^2 formula only increases if the independent variable added improves the model.

Nash–Sutcliffe efficiency (NSE) The Nash–Sutcliffe efficiency (NSE) is a normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance (Nash and Sutcliffe 1970). Nash–Sutcliffe efficiency indicates how well the plot of observed versus simulated data fits the 1:1 line. NSE = 1, corresponds to a perfect match of the model to the observed data. NSE = 0, indicates that the model predictions are as accurate as of the mean of the observed data, $-\infty < \text{NSE} < 0$, indicates that the observed mean is a better predictor than the model.

Model validation

Validation of the selected regression model in this study was performed using 9 hydrometric stations (30% of the total stations) as ungauged (pseudo-ungauged) catchments.

As shown in Table 3 last column, the large, medium, and small size catchments were incorporated and the spatial distribution of the stations is shown in Fig. 4.

Table 3 Hydrometric station used for validation as Pseudo-ungauged catchments

No	River	Site	Catchment area (Km^2)
1	Abbay	@ Bahir Dar	15,321
2	Angar	Nr. Nekemte	4674
3	Ardy	Nr. Metekel	219
4	Beressa	Nr. Debre Berhan	211
5	Didessa	Nr. Arjo	9981
6	Gumara	Nr. Bahir Dar	1394
7	Lower Fettam	@ Galibed	757
8	Megech	Nr. Azezo	462
9	Robigumer	Nr. Lemi	887

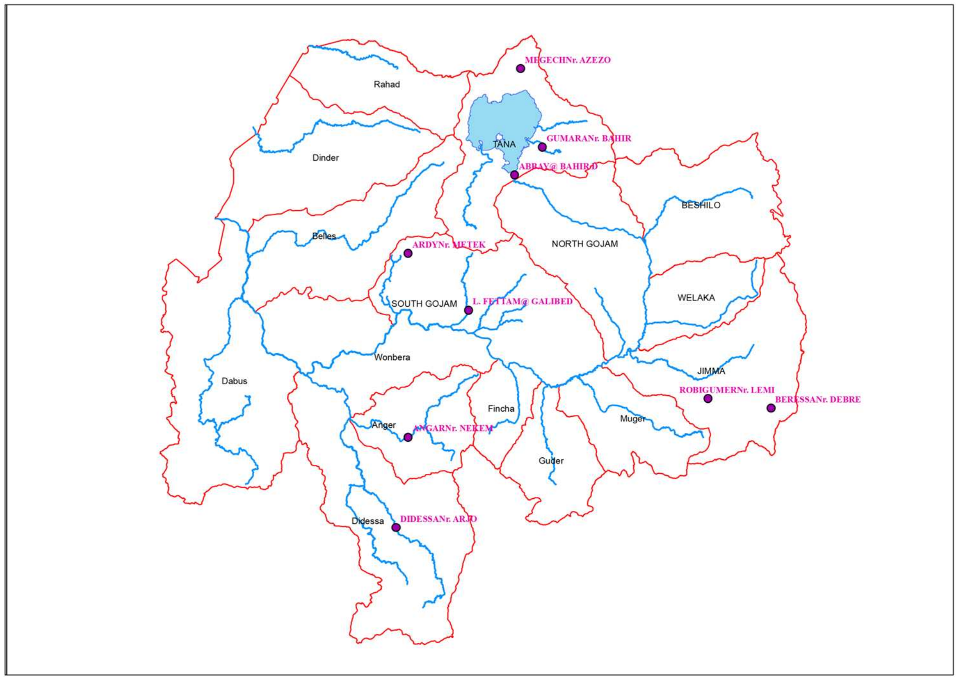

Fig. 4 Spatial distribution of validation stations

Table 4 Regression analysis result for each model

Predictors	Multiple R^2	R^2	R^2_{adj}	NSE	Standard error
A	0.96	0.93	–	0.93	0.1577
A & MAR	0.96	0.93	0.92	0.93	0.1623
All	0.98	0.97	0.96	0.97	0.1130
A, MAR & AE	0.98	0.96	0.96	0.96	0.1178

Result

The regression model developed as stated in the model criticism and selection section performance of the models was assessed graphically and statistically. At last, the seemly regression models were validated with selected stations. The subsequent four subheadings address each regression model result.

Regression analysis area (A) as the only predictor

The first regression model used only area as the predictor variable that estimates the response variable mean annual flow. The logarithmic transformation form of the simple linear regression equation is

$$\log \text{MAF} = \log\alpha + \beta * \log A$$

Based on Table 4, the relationship between the dependent and independent variables is significant. Table 5 summarizes the coefficients and hypothesis tests of this model. As it can be seen that the *t* test of slope β is 14.37 which is greater than $t_{(0.025,17)}$ which is read as 2.11 from the student's *t* table,

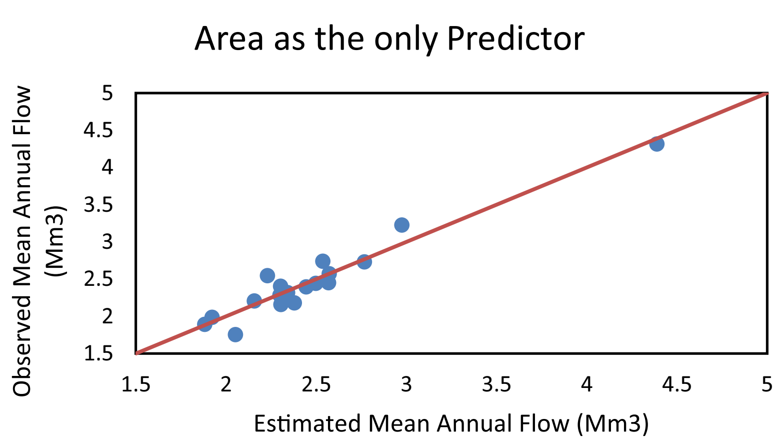

Fig. 5 Comparison of estimated versus observed MAF using area as the only predictor

Table 5 Coefficient table for area as the only predictor variable

	Coefficients	Standard error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	0.0957	0.1699	0.5635	0.5809	− 0.2644	0.4559
Area	0.8930	0.0621	14.3697	1.5E−10	0.7612	1.0247

and the corresponding *p*-value was less than the adapted significance level 0.05.

In addition, *F* test in the ANOVA table (Table 6) shows that it is much greater than $F_{(0.05,1,17)}$ which is read as 4.45 from the *F*-distribution table and the corresponding *p* value was less than the adapted significance level 0.05. Furthermore, graphically a comparison was made by plotting on estimated and observed values as shown in Fig. 5.

Regression analysis using area and mean annual rainfall as a predictor

The second regression model was a multiple linear regression analysis that was started by adding mean annual rainfall to the first regression model area as the only predictor variable. The log-transformed form of this regression analysis is

$$\mathrm{MAF} = \log\alpha + \beta\log(A) + \gamma\log(\mathrm{MAR}).$$

In this model, as shown in Table 4, there was no increment or decrement in coefficients of determination (R^2) and Nash Sutcliffe (NSE) but an increment in standard error was noticed. In addition, in this model, adjusted coefficients of determination (R^2_{adj}) were considered (Fig. 5).

However, from ANOVA (Table 7), F test shows that null hypothesis test is rejected and the coefficient table (Table 8) exhibits that the added explanatory variable MAR is not significant because the corresponding *t* test value and *p* values were less than $t_{(0.025,16)} = 2.12$ and significance level 0.05, respectively.

Table 6 Analysis of variance (ANOVA) table for area as the only predictor variable

	df	SS	MS	*F*	Significance *F*
Regression	1	5.1378	5.1378	206.49	1.45E−10
Residual	16	0.3981	0.0249		
Total	17	5.5359			

Table 7 ANOVA table for area and mean annual rainfall as the predictor variables

	df	SS	MS	*F*	Significance *F*
Regression	2	5.1409	2.5705	97.6263	2.51E−09
Residual	15	0.3950	0.0264		
Total	17	5.5359			

Stepwise regression analysis

The last regression analysis was an attempt to find the most appropriate predictor variables for the basin. Backward elimination stepwise regression analysis was used to find the best-fit model. Therefore, all independent variables were initially used and if there is a violation of the hypothesis test in *p* value or *t* test, the corresponding independent variable that violates was subsequently eliminated from the analysis, and using the remaining independent variables, the analysis was repeated until the best fit achieved.

Therefore, in this analysis, using all predictor variables initially satisfies *t* test, *F* test, and *p* value except average slope (bold) as shown in Tables 9 and 10.

Table 8 Coefficient table for area and mean annual rainfall as predictor variables

	Coefficients	Standard error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	− 0.9638	3.0620	− 0.3148	0.7573	− 7.4902	5.5626
Area	0.8968	0.0649	13.8217	6.1E−10	0.7585	1.0351
MAR	0.3307	0.9541	**0.3466**	**0.7337**	− 1.7030	2.3644

Table 9 Coefficient table of all independent variables

	Coefficients	Standard error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	− 19.6736	4.9082	− 4.01	0.0015	− 30.2771	− 9.0701
Area	0.9713	0.0504	19.26	6.1E−11	0.8623	1.0803
MAR	3.5216	1.0054	3.50	0.0039	1.3496	5.6936
AS	0.2144	0.1440	**1.49**	**0.1604**	− 0.0967	0.5255
AE	2.3947	0.5754	4.16	0.0011	1.1515	3.6378

Table 10 ANOVA table for all independent variables as the predictor

	df	SS	MS	*F*	Significance *F*
Regression	4	5.3699	1.3425	105.12	9.20E−10
Residual	13	0.1660	0.0128		
Total	17	5.5359			

In the next stage, the average slope eliminated, and regression analysis was performed again with the remaining independent variable. Table 4 shows a minor dropping of R^2 and Nash Sutcliff and increment of the standard error of the model compared with previous analysis.

This combination of catchment characteristics is statistically significant (t test > t-cr, p value < 0.05, and F test > F-cr) as shown in Tables 11 and 12. Comparison of observed and estimated mean annual flow using area, mean annual rainfall, and the average elevation is shown in Fig. 6.

Validation

The validation analysis was made only on two regression models that pass statistical tests (Table 13); these are the simple linear regression that area as the only predictor and multiple regression area, mean annual rainfall, and average elevation as predictors. Using the coefficients in Table 13, the observed and estimated mean annual flow for each equation was estimated. The first equation has well estimated in the Abbay @ BahirDar and Gumara stations compared to the second equation. However, for the remaining seven stations, the second equation estimates excellently well with the first equation.

The coefficient of determination and Nash–Sutcliffe efficiency was calculated for each model, and the first model gives more acceptable results than the second one. The estimation of the second model for Abbay @ BahirDar was almost three times the observed one. The first model exhibits that the two catchments that have the same area in the basin have the same mean annual rainfall but the Abbay river basin rainfall distribution is not uniform. Therefore, this model will be correct at least if both catchments are in the same rainfall distribution parts of the basin. Therefore,

Table 11 Coefficient table of area, MAR, and AE as predictors

	Coefficients	Standard error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	− 17.5316	4.8922	− 3.5836	0.0030	− 28.024	− 7.0388
Area	0.9815	0.0521	18.8359	2.4E−11	0.8697	1.0932
MAR	3.0777	1.0010	3.0746	0.0082	0.9308	5.2246
AE	2.2463	0.5909	3.8018	0.0019	0.9791	3.5136

Table 12 ANOVA table for area and mean annual rainfall as the predictor variables

	df	SS	MS	*F*	Significance *F*
Regression	3	5.3415	1.7805	128.28	2.03E−10
Residual	14	0.1943	0.0139		
Total	17	5.5359			

Table 13 Summarized coefficients of regression analysis

Predictor	α	β	γ	μ
Area	$10^{0.096}$	0.89		
A, MAR and AE	$10^{-17.53}$	0.98	3.08	2.25

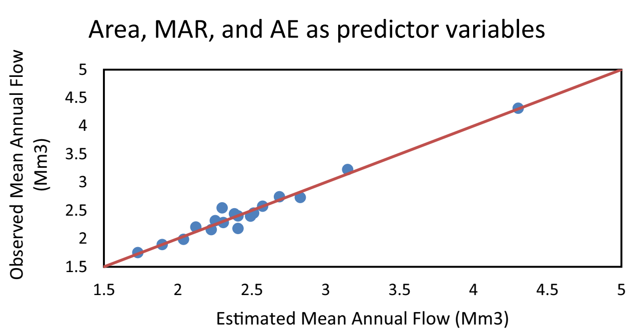

Fig. 6 Estimated versus observed MAF of Area, MAR, and AE as a predictor

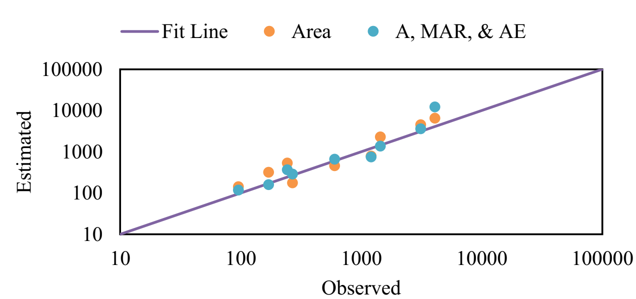

Fig. 7 Validation plots of the observed and estimated mean annual flow

Table 14 R^2 and Ns of validation results

Predictor	area	*A*, MAR, & AE with Abbay @ BDR	A, MAR, & AE without Abbay @ BDR
R^2	0.97	0.80	0.96
NSE	0.80	0.49	0.96

considering these reasons and seeing Fig. 7, the second model (A, MAR, & AE as predictors) was selected as a validated model. The validation statistics are shown in Table 14. The comparison was made on how much the estimated and observed values on each station are related in terms of magnitude and graphically (Fig. 7).

Discussion

Ordinary least-square regression model of mean annual flow (MAF) with four independent variables (viz., area, mean annual rainfall, average slope, and average elevation) was performed in this study. This analysis started with the area as the only predictor and gives 0.93 in both coefficients of determination (R^2) and Nash–Sutcliffe efficiency (NSE) with a standard error of 0.1577. The analysis done was found statistically significant as shown in coefficient table (Table 5) and ANOVA (Table 6).

In the second model, the mean annual rainfall was chosen as an independent variable but did not show any significant improvement of the model rather than the increment of standard error (0.1623). Moreover, this model failed to satisfy the hypothesis test. As we show in Table 8, the exponent of the area does not change but the sign of the intercept is opposite with area as the only predictor, and the exponent of mean annual rainfall is less than 0.5.

Lastly, the most suitable model was selected using a stepwise regression model with backward elimination by starting from taking of all independent variables as predictors. In this time, R^2, R^2adj, and NSE increase 0.97, 0.97, and 0.96 respectively, and the standard error decreased (0.1130). However, in Table 9, average slope did not satisfy the hypothesis test ($p > \alpha$). Next, the analysis was proceeded by eliminating the average slope and it showed that the area, mean annual rainfall, and average elevation are statistically significant with 0.96 in R^2, R^2adj, and NSE. In this regression model, the exponent of predictors showed their relationships with the mean annual flow. For the area, the exponent is almost one (0.98); hence, the relationship between the size of the catchment and mean annual flow is linear, or in other words, unit increase in the area contributes a unit volume of water to the river channel. Mean annual rainfall had exponents greater than one (3.08) which reflects non-linear relationships between them. Also, the exponent of average elevation is greater than one that tells there were no linear relationships between them.

The comparison of previous study results based on the coefficient of determination and the number of independent variables was made. For instance Vogel et al. (1999), for area as the only predictor R^2 obtained 0.91, and with all the variables as predictors, R^2 increased to 0.99. Also, the number of geomorphological and climate variables was more associated with this study, and they used a maximum of five predictors to develop the most suitable mean annual flow that in almost all regions, catchment area was inclusive. Papamichail et al. (2002) selected the best-fitted model that includes area, annual rain, and length of a stream with R^2 of 0.78. Hence, their result is not much greater than this study even though one predictor is not the same. Moreover, Burgers et al. (2013) founded R^2 of 0.40 for area as the only predictors and incorporating mean annual rainfall they obtained an increment of R^2 of 0.56, while, in this study, the inclusive of mean annual rainfall did not improve the model rather than the increment of standard error. The exponent of the area looks similar to this study of 0.86. The above-mentioned studies almost used similar climate as well as physiographic variables with this study; however, Tran et al. (2015) include geomorphologic and anthropogenic variables for 533 catchments. Barbarossa et al. (2017) developed a global regression model using catchment area, mean annual rainfall, mean annual temperature, mean elevation, and mean slope as the predictor for 1885 catchments. They obtained R^2 of 0.89 which is very less compared with this study.

As mentioned earlier, the catchment area, the mean annual rainfall, and the average elevations are statistically significant next to the area as the only predictor. However, the validation analysis showed the multiple regression equation that includes the three catchment characteristics and is the best model capable of estimating the mean annual flow in ungauged catchments in the basin. Therefore, the model looks like

$$\mathrm{MAF} = 10^{-17.53} A^{0.98} * \mathrm{MAR}^{3.08} \mathrm{AE}^{2.25}$$

where *A* is catchment area (Km^2), MAR is mean annual rainfall (mm), and AE is the average elevation in (m).

The selected regression model is one of the best models for estimation of mean annual flow for ungauged catchments. Here, the model parameters (catchment characteristics) are easily accessible for practitioners that want to estimate the mean annual flow in the study area. The limitations of the present studies naturally include exclusive of the independent variables which might influence the formation of runoff through the catchment. These are the soil characteristics, land use/land cover, shape of the catchment, and temperature. It suffers from the same limitations as other local studies associated with the quality and number of hydrometeorological data.

Conclusions

This study aimed at the development of empirical equations that can be used for ungauged catchments in the Abbay river basin using the regression model. Climate and physiographic characteristics were used as predictors to develop the empirical equation that satisfies statistical tests. The developed regression model that is capable to estimate the mean annual flow for ungauged catchments includes catchment area (Km^2), mean annual rainfall (mm), and average elevation (m) as predictors that are statistically significant with values of 0.96 for all R^2, R^2_{adj}, and NSE. The validation of the model was tested using nine hydrometric stations, and the result strengthens the model capability to estimate the mean annual flow for ungauged catchments in the Abbay river basin concerning R^2 and NSE of 0.96 and using validation plot (Fig. 7). Therefore, it is possible to use the developed empirical equation in the water resources development project for ungauged catchments at any geographical space with the same environmental conditions. This model can be used even at regional and continental scale with some modifications.

References

Abebe SG, Semu AM, Seleshi BA (2013) Basin regionalization for the purpose of water resource development in a limited data situation: case of Blue Nile River Basin, Ethiopia. J Hydrol Eng. https://doi.org/10.1061/(ASCE)HE.1943-5584

Barbarossa V, Huijbregts MAJ, Hendriks AJ, Beusen AHW, Clavreul J, King H, Schipper AM (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

Berhane M (2013) Estimation of monthly flow for ungauged catchment (Case Study Baro—Akobo basin). MSc Thesis Addis Ababa University, Addis Ababa

Burgers HER, Schipper AM, Hendriks AJ (2013) Size relationships of water discharge in rivers : scaling of discharge with catchment area, main-stem length and precipitation. Hydrol Process 28:5769–5775. https://doi.org/10.1002/hyp.10087

Gebeyehu A (1989) Regional frequency analysis. Dissertation, The Royal Institute of Technology

Javeed Y, Apoorva KV (2015) Flow regionalization under limited data availability—application of IHACRES in the Western Ghats. Aquat Procedia 4(lcwrcoe):933–941. https://doi.org/10.1016/j.aqpro.2015.02.117

Naghettini M, Silva AT, Costa V, Fernandes W, de Andrade Pinto EJ (2017) Fundamentals of statistical hydrology. In: Naghettini M (ed), Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-43561-9

Nash JE, Sutcliffe IV (1970) River flow forcasting through conceptual models part I—a discussion of principles. J Hydrol 10:282–290

Papamichail DM, Georgiou PE, Diamantopoulou MJ (2002) Multiple regression models for predicting annual runoff in ungauged catchments. Conference Paper

Razavi T, Coulibaly P, Asce M (2013) Streamflow prediction in ungauged basins : review of regionalization methods streamflow prediction in ungauged basins: review of regionalization methods. J Hydrol Eng 18(8):958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

Sedighi A (2008) A linear regression model for predicting stream response time in Karst watersheds using Dems. Dissertation, University of Florida

Sivapalan M, Takeuchi K, Franks SW, Gupta VK, Karambiri H, Lakshmi V (2003) IAHS decade on predictions in ungauged basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. Hydrol Sci J. https://doi.org/10.1623/hysj.48.6.857.51421

Tamalew C, Kemal A (2016) Estimation of discharge for ungauged catchments using rainfall-runoff model in Didessa Sub-basin: the case of Blue Nile River. Int J Innov Eng Res Technol [IJIERT] 3(9):62–72

Tesfaye B (2011) Predicting discharge at ungauged catchments using rainfall-runoff model (case study: Omo-Gibe River Basin). Addis Ababa University, Addis Ababa

Tran LT, O'Neill RV, Bruins RJF, Smith ER, Harden C (2015) Linking land use/land cover with climatic and geomorphologic factors in regional mean annual streamflow models with geospatial regression approach. Prog Phys Geogr. https://doi.org/10.1177/0309133314562441

Vogel RM, Wilson I, Daley C (1999) Regional regression models of annual streamflow for The United States. J Irrig Drain Eng 125(3):148–157

Wale A, Rientjes THM, Gieske ASM, Getachew HA (2009) Ungauged catchment contributions to Lake Tana's water balance. Hydrol Process 23:3682–3692. https://doi.org/10.1002/hyp.7284

Waylen P, Woo M-K (1982) Prediction of mean annual flows in the Fraser River catchment, British Coumbia. J Am Water Resour Assoc. https://doi.org/10.1111/j.1752-1688.1982.tb00055.x

Yu PS, Tsai HP, Chen ST, Wang YC (2005) Estimation of design flow in ungauged basins by regionalization. Department of Hydraulic and Ocean Engineering, National Cheng Kung University, Taiwan, 2494–2499

Zigaf A, Eriksson A (2018) Regional regression models of mean annual streamflow and design flow in a tropical region of Colombia. [Lund University]. www.tvrl.lth.se