Jordan Journal of Civil Engineering, Volume **16**, No. **1**, **2022**

Soil Erosion Prediction Using Gene Expression Programming Based on Physically, Fully-Distributed Hydrologic Model Outputs

Anas Mahmood Al-Juboori [1)]

[1)] Dams and Water Resources Research Center, University of Mosul, Mosul, Iraq. E-Mail: anasmmr@uomosul.edu.iq

ABSTRACT

The present research aims to develop a regional equation for estimating soil erosion for ungauged basins in Nineveh province in Iraq. The research aims to use Gene Expression Programming (GEP) to develop a regional equation to estimate the soil erosion for the study area based on the results of the hydrological model outputs. The fully distributed hydrological model "Gridded Surface Sub-surface Hydrological Analysis (GSSHA)" was used to simulate the soil erosion for the study area. The morphometric characteristics of the basin (area, length and slope) in addition to daily rainfall were used as input variables, while soil erosion resulting from the GSSHA model was used as a target variable in the proposed GEP model. The results proved the efficiency of the proposed model for developing a regional equation to predict soil erosion using regional morphometric properties for the study area. The results show that the GSSHA model has a good performance in simulating soil erosion. The observed amount of soil erosion is 634 m^3, while the simulated amount of the valley with observation data is 682 m^3.

KEYWORDS: Soil erosion, GSSHA model, GEP, Rainfall-runoff model.

INTRODUCTION

The issue of estimating eroded soils from the catchments is one of the complex issues facing engineers, especially when designing hydraulic structures. Soil erosion mechanism is a very complex topic, as it is the outcome of a set of overlapping hydrological and geomorphological mechanisms. Most of the valleys and river basins in the world, especially in developing countries, are characterized by the lack of gauging stations to monitor sediment load. Therefore, mathematical models are the best solutions for estimating the amount of sediment. Several mathematical models exist to simulate soil erosion. GSSHA model is considered one of the most important hydrological models for simulating soil erosion (Downer and Ogden, 2006). GSSHA model is a fully distributed physical hydrological model that uses the finite difference method to simulate soil erosion. GSSHA model was derived from the CASC2D hydrologic model.

GSSHA model was used to simulate soil erosion in Muddy Brook watershed in the USA. The results showed the accuracy of GSSHA model for streamflow modeling (Downer and Ogden, 2004). GSSHA model was applied to the urbanizing arid catchment in Saudi Arabia to simulate flood (Sharif et al., 2017). GSSHA and SWAT (Soil and Water Assessment Tool) were used to simulate sediment erosion in a small agricultural catchment in Japan. The results showed that GSSHA model performance was better than that of SWAT model, especially for estimating sediment concentration (Sith and Nadaoka, 2017). SWAT model was applied to simulate sediment load in a karst watershed of the semi-arid Mediterranean basin (Martínez-Salvador and Conesa-García, 2020). The Soil Moisture Balance/Budgeting (SMB) method was applied to simulate sediment yield and runoff (Gupta et al., 2019). Arc-View Soil and Water Assessment Tool (AVSWAT) model was applied to simulate sediment inflow to Mujib dam catchment area in Jordan (Ijam and Al-Mahamid, 2012).

GEP algorithm is one of the most revolutionary

Received on 28/4/2021.
Accepted for Publication on 3/10/2021.

algorithms that are widely used in hydrological modeling. GEP was developed by Ferreira (2001) based on Dron's theory of evolution. GEP was applied in preparing hydrological models for flood prediction (Zorn and Shamseldin, 2015; Seckin and Guven, 2012). GEP was used to prepare mathematical models to the regionalization of flow duration curve at ungauged sites (Al-Juboori and Guven, 2016; Hashmi and Shamseldin, 2014; Booker and Snelder, 2012). GEP algorithm was used to model stage-discharge curve for of the Pahang River (Azamathulla et al., 2011). GEP was used to simulate the monthly water level data of Van lake in Turkey (Aytek and Guven, 2014). GEP was used to solve the complex problems in the preparation of hydrological rainfall-runoff models (Shoaib et al., 2015; Fernando et al., 2012). Suspended sediment load was estimated to the Middle Euphrates Basin in Turkey using GEP algorithm (Guven and Talu, 2010).

In the current research, GSSHA model will be used to estimate soil erosion to ungauged valleys in Nineveh governorate in Iraq with limited observed data. The results of the hydrological model will be used with morphological characteristics of these valleys to develop a regional mathematical equation to predict soil erosion in the study area using GEP algorithm.

Study Area and Data Used

The study area represents all the valleys located on the course of Tigris River between the site of Mosul dam and the south of Mosul city. The Digital Elevation Model (DEM) with an accuracy of 90 * 90 meters was used to delineate the studied basins. The results showed that there are six valleys flowing into the right bank of the Tigris River and seven valleys flowing into the left bank of the Tigris River. Figure (1) shows the valleys of the study site, while Table 1 shows the morphometric characteristics of those valleys. The symbol R was used to symbolize the valleys located on the right bank of the Tigris River and the symbol L was used to symbolize the valleys located on the left bank of the Tigris River. The study area includes thirteen rainfall stations to measure the depth of rain. The boundaries of the Thiessen polygons to these stations have been determined to estimate the rainfall rate for each valley, as shown in Figure (1). Daily rainfall data for the period from 1/11/2016 to 31/10/2019 was collected to these stations. The classification and distribution of soil samples in the study area are 82% silty loam and 18% silty clay loam. The valleys of the study area are located in the undulating region of Iraq within the range of the intermediate folds. The region is also characterized by the presence of flood plains, especially at the confluence of valleys with the Tigris River.

Table 1. Morphometric characteristics of the studied basins

Basin name	Basin area (km^2)	Basin slope (m/m)	Basin length (km)	Basin perimeter (km)
R1	77.6	0.0203	20.74	67.9
R2	2485.4	0.0195	72.96	341.9
R3	391.2	0.0291	32.18	122.9
R4	42.8	0.0346	10.77	41.7
R5	108.8	0.0424	18.42	74.1
R6	119.8	0.0377	13.85	62.2
L1	118.9	0.0837	17.72	67.6
L2	152.0	0.142	28.25	97.9
L3	323.3	0.0354	33.32	126.7
L4	95.8	0.0274	16.71	59.2
L5	840.9	0.0548	52.15	206.0
L6	358.2	0.0255	31.79	117.1
L7	44.9	0.0216	10.91	41.7

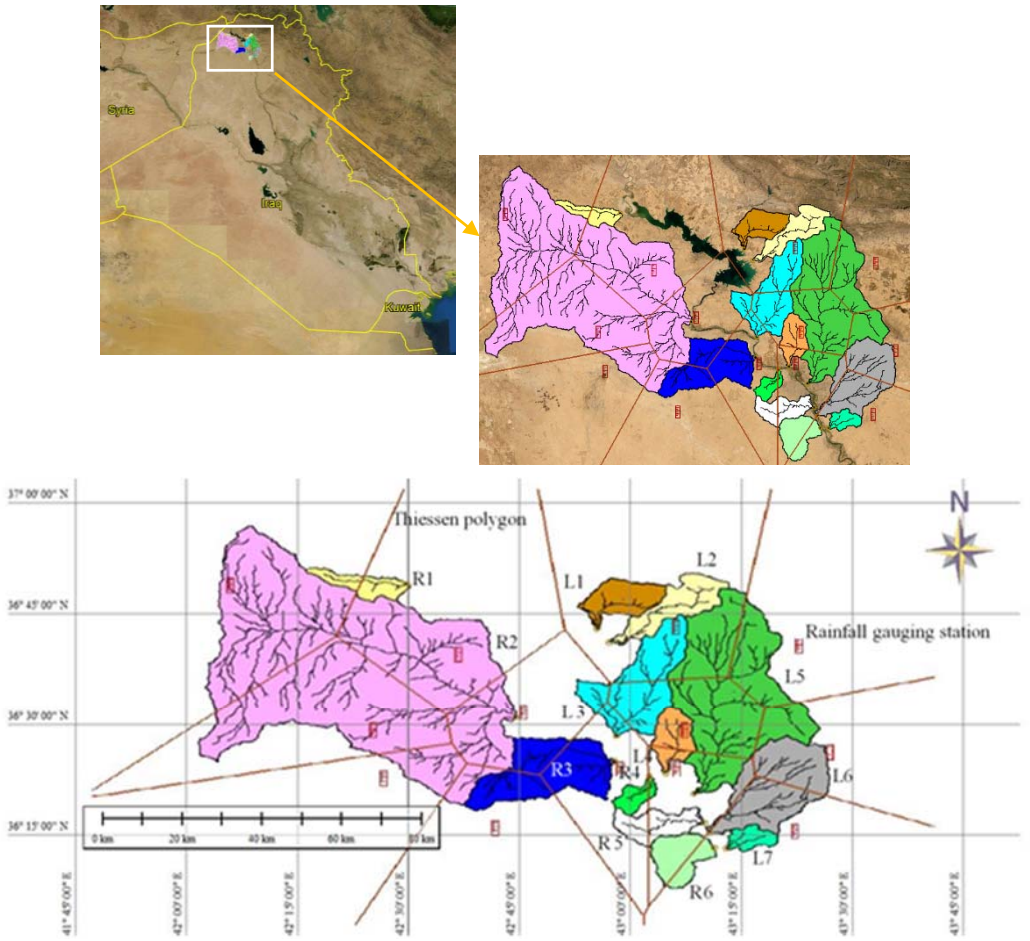

Figure (1): Study area

GSSHA Model Overview

The theoretical basis of the GSSHA model for simulating surface runoff and soil erosion depends on the principle of two-dimensional finite difference method by dividing the study area into a small square cell network similar to a fishing net (Downer and Ogden, 2006). GSSHA model was derived from the CASC2D hydrologic model. The network is fed with all information by using a set of layers that are created by geographic information system (GIS) and then converted into index maps that describe the physical characteristics of the study area.

Three processes of hydrological modeling are required for GSSHA model setting, including the simulation of overland flow, infiltration and soil erosion. The overland flow is modeled using the two-dimensional finite difference method. The grid in GSSHA model is square in shape ($\Delta x = \Delta y$). The overland flow is estimated for each grid using Manning equation in the x and y directions. The important mathematical equations used to estimate overland flow in the model are described as (Downer and Ogden, 2006):

$$p_{ij}^{t} = \frac{1}{n}\left(d_{ij}^{t}\right)^{\frac{5}{3}}\left(S_{fx}^{t}\right)^{\frac{1}{2}} \quad (1)$$

$$q_{ij}^{t} = \frac{1}{n}\left(d_{ij}^{t}\right)^{\frac{5}{3}}\left(S_{fy}^{t}\right)^{\frac{1}{2}} \quad (2)$$

$$d_{ij}^{t+1} = d_{ij}^{t} + \frac{\Delta t}{\Delta x}\left(p_{i-1,j}^{t} + q_{i,j-1}^{t} - p_{ij}^{t} - q_{ij}^{t}\right) \quad (3)$$

where, p and q are the overland flow in the x and y directions, d is the flow depth, n is the Manning coefficient, t is time, S is the grid slope and *d*ij is the flow depth in each grid at any time interval. The Green and Ampt method is used to simulate infiltration loss (Downer and Ogden, 2006). Calculating infiltration using the Green and Ampt method requires many more variables than other methods of calculating infiltration, such as the Darcy method. It is a function of soil suction head, porosity, hydraulic conductivity and time. The application of the Green and Ampt method requires four soil hydraulic parameters; soil saturated hydraulic conductivity, soil capillary suction head parameter, effective porosity and initial soil moisture content. The Green and Ampt equation is described as:

$$f(t) = k\left[\frac{\psi\Delta\theta}{F(t)} + 1\right] \quad (4)$$

where, ψ is wetting front soil section head, θ is water content, k is hydraulic conductivity and $F(t)$ is the cumulative depth of infiltration. The GSSHA model uses the Kilinc and Richardson equation that is modified by Julien in order to calculate the soil erosion of the basin (Downer and Ogden, 2006). The improved Julian equation is characterized by being more stable in the case of irregular flow under different conditions of land use and soil type. The potential sediment discharge in each grid cell is calculated using the following equations in the x and y flow directions:

$$q_{sx_{ij}} = 25500p_{ij}^{2.035}\left|S_{fx_{ij}}\right|^{1.644}\frac{S_{fx_{ij}}}{\left|S_{fx_{ij}}\right|}\frac{K*C*Y}{0.15} \quad (5)$$

$$q_{sy_{ij}} = 25500p_{ij}^{2.035}\left|S_{fy_{ij}}\right|^{1.644}\frac{S_{fy_{ij}}}{\left|S_{fy_{ij}}\right|}\frac{K*C*Y}{0.15} \quad (6)$$

where, K is soil erodibility factor, with values ranging from 0 to 1, C is soil cropping factor (0-1) and P is conservation factor (0-1).

Gene Expression Programming (GEP) Overview

GEP is the process of imitating biological evolution to create computer programs to simulate certain phenomena by developing a symbolic regression model to solve a problem. GEP was developed by Cândida Ferreira (2001). GEP model is classified as a non-parametric model because the form of the model is unknown and the goal of the GEP algorithm is to find the best function to fit the data. The encoding process in GEP is one of the most important steps in developing a model. GEP model is developed using the notation called the Karva Language. The expressions encoded by Karva Language are called k- expressions. Considering simple mathematical expressions (d*e+f), this expression is encoded to expression tree as shown in Figure 2.

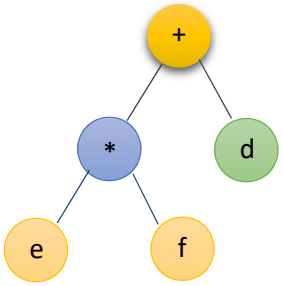

Figure (2): Expression tree for (d*e+f)

GEP model is constructed through five major steps; defining a set of independent variables to be used in individual programs, defining a set of mathematical functions and operations, selecting fitness function, selecting the head length, number of genes and linking function and selecting genetic factors. The definitions of genetic factors are briefly described as follows: *Inversion*: The codes in a section of the gene are reflected in order. *Mutation*: The symbols in genes are replaced by alternate symbols. *Transposition*: It selects a set of codes and transmits the codes to a different position within the same gene. *Recombination*: Two chromosomes are randomly selected and then the genetic material between them is exchanged to produce new chromosomes. A chromosome is composed of one or more genes. Genes are linked using a suitable linking function to generate the final expression, as shown in Figure 3. The GEP algorithm flowchart is shown in Figure 4.

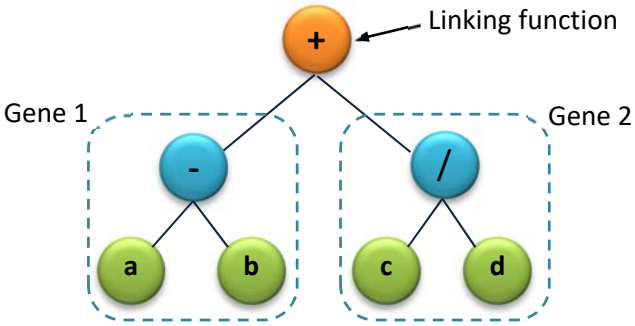

Figure (3): Linking function example

Performance Evaluation

In this study, two different statistical indicators were chosen to evaluate the proposed models; namely, the coefficient of determination (R^2) and the Root Mean Square Error (RMSE).

$$R^2 = \left[\frac{\sum_{i=1}^{n}(E_o - \overline{E_o})(E_p - \overline{E_p})_i}{\sqrt{\sum_{i=1}^{n}(E_o - \overline{E_o})^2(E_p - \overline{E_p})_i^2}}\right]^2 \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_o - E_P)^2} \quad (8)$$

where E_o and E_p are the observed and predicted data values, respectively, $\overline{E_o}$ and $\overline{E_p}$ are the mean values of the observed and predicted data and n is the length of the data. The best model is that having the lowest RMSE value and R^2 value closest to one.

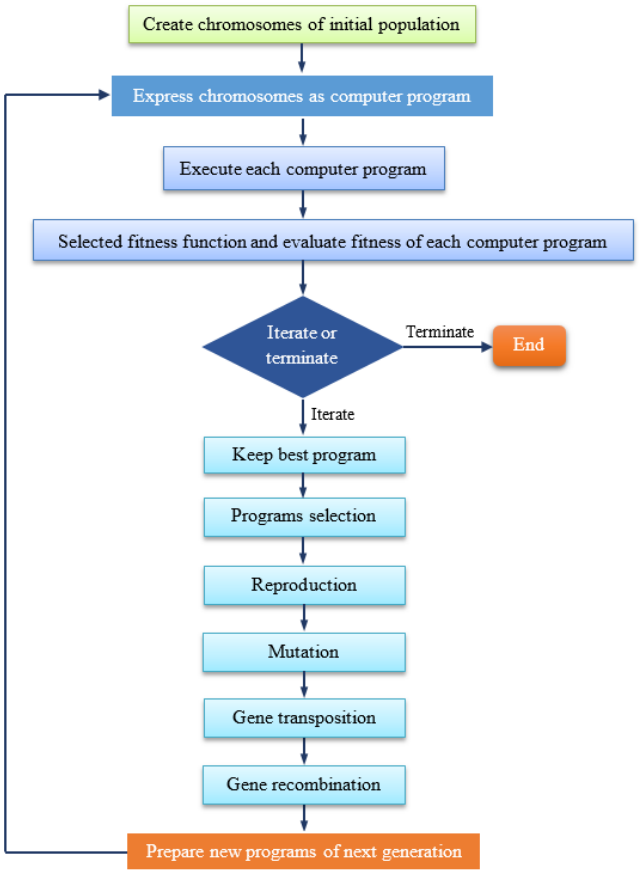

Figure (4): Flowchart of the GEP algorithm (Ferreira, 2001)

RESULTS AND DISCUSSION

The basins of the study area are classified as ungauged basins due to the lack of discharge gauging stations, except for limited monitoring of one basin, which is located on the left bank of the Tigris River in Mosul city with symbol L9 in Figure 1. The simulation of surface runoff and soil erosion using the GSSHA model requires the determination of infiltration parameters and parameters that affect soil erosion. The Green and Ampt method is used in GSSHA model to estimate the infiltration loss. The infiltration parameters included in Green and Ampt method are: "hydraulic conductivity, capillary head, porosity, pore distribution index, residual saturation, field capacity and wilting point". The Kilinc and Richardson equation modified by Julien is used to simulate the soil erosion. The soil erosion parameters included in the modified Kilinc and Richardson equation are: coefficient of detachment by rainfall, rill erodibility coefficient, rill erodibility exponent, critical rill detachment and erodibility coefficient.

The GSSHA model of L9 basin was calibrated using the observed data available on February 19, 2003 (Mohammad, 2005). The daily rainfall for this storm was 19 mm and the observed sediment volume was 634 m^3. The simulated soil erosion volume was 682 m^3. The calibrated infiltration and soil erosion parameters are shown in Table 2. The model was tested using the observed data available on January 22, 2004 (Mohammad, 2005) and the calibrated infiltration and soil erosion parameters in Table 2. The daily rainfall for this storm was 27 mm and the observed sediment volume was 1106 m^3. The simulated soil erosion volume was 1182 m^3. The results showed a good performance of GSSHA model to simulate soil erosion for the study area. The percent of error for the L9 basin model is 7.5% and 6.9%, respectively, for the February and January data.

The high similarity in the geological and soil characteristics in the study area and the infiltration and soil erosion parameters obtained from the calibration of the hydrological model of L9 basin will be adopted as initial parameters for all basins of the study area.

Table 2. Calibrated infiltration and soil erosion parameters

Infiltration		**Soil erosion**	
Parameter	**Value**	**Parameter**	**Value**
Hydraulic conductivity	0.21	Coefficient of detachment	16.3
Capillary head	27.5	Rill erodibility coefficient	0.0004
Porosity	0.46	Rill erodibility exponent	0.67
Pore distribution	0.178	Critical rill detachment	0.1
Residual saturation	0.040	Erodibility coefficient	0.28
Field capacity	0.371		
Wilting point	0.212		

The GSSHA model was applied to all basins in the study area and the results of soil erosion were collocated to develop the GEP models to predict the soil erosion model for the studying area. The GEP models were developed using the morphometric characteristics for the basins and daily rainfall depth as predictor variables, while the soil erosion resulting from GSSHA model was considered as the target variable. Four different combinations of input variables were proposed to develop the GEP models. The proposed model combinations are described as:

$$G1: E = f(A, P) \quad (9)$$

$$G2: E = f(L, P) \quad (10)$$

$$G3: E = f(A, L, P) \quad (11)$$

$$G4: E = f(A, L, S, P) \quad (12)$$

where, *G1-4* is model symbol, E is soil erosion, A is basin area, L is basin length, S is basin slope and P is daily rainfall depth. The proposed input combinations were used to develop the GEP models to predict the soil

erosion for the studying area. The data was divided into two parts, 80% of it was used to train the model, while the other 20% was used to validate the models. The data that will be used to validate the model was randomly selected to represent all the characteristics of the basins in the study area. The results of the GEP models for the four different proposed input combinations are illustrated in Table 3. We can see that the models developed with more input predicators give better results than those developed with fewer input predicators. The R^2 values in training phase are 0.51, 0.53, 0.71 and 0.88 for G1, G2, G3 and G4 models, respectively. The R^2 values in the validating phase are 0.58, 0.57, 0.69 and 0.91 for G1, G2, G3 and G4 models, respectively. The results of RMSE are 1604, 1584, 1274 and 656m^3 in the training phase and 1629, 1666, 1399 and 810m^3 in the validating phase for G1, G2, G3 and G4 models, respectively. It is clear that the G4 model is the best model for predicting soil erosion for the study area. The GEP symbolic expression for the G4 model combination is shown in Equation 13.

Table 3. Results of GEP model

Model symbol	Training		Validation	
	R^2	RMSE (m^3)	R^2	RMSE (m^3)
G1	0.51	1604	0.58	1629
G2	0.53	1584	0.57	1666
G3	0.71	1274	0.69	1399
G4	0.88	656	0.91	810

$$E = LP + S - A + \frac{\sqrt{A}}{S} + \frac{8}{S} - 200 \qquad (13)$$

where, E is soil erosion in (m^3), L is basin length in (km), P is daily rainfall in (mm), A is basin area in (km^2) and S is basin slope in (m/m). The scatter plots in Figures 5 and 6 show the comparison between the results of soil erosion using the GSSHA model and the GEP model in the training and validating phases, respectively. The results in Figures 5 and 6 proved the efficiency of the GEP model to predict soil erosion in the study area. The R^2 value is 0.89 in the training phase and 0.86 in the validating phase. Using Equation 13 to calculate soil erosion for basin L9, it was found that the amount of eroded soil for the months of February and January was 625 m^3 and 1042 m^3, respectively. In comparison with the observed values of 634 m^3 and 1106 m^3 for the months of February and January, respectively, it is clear that Equation 13 is accurate in estimating soil erosion for the studied valleys.

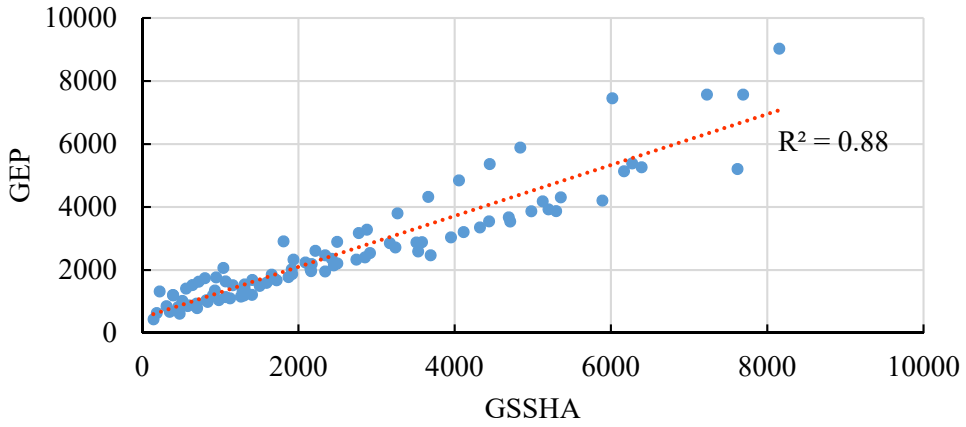

Figure (5): GSSHA-GEP soil erosion comparison in the training phase

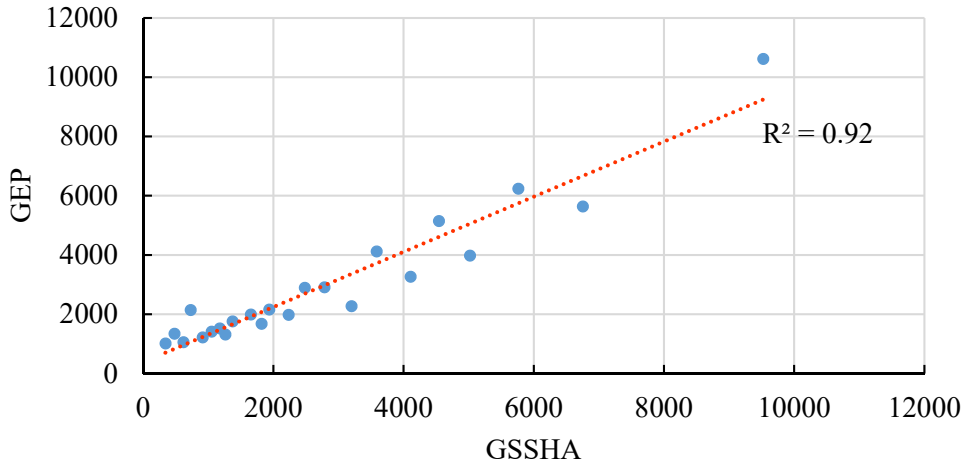

Figure (6): GSSHA-GEP soil erosion comparison in the validating phase

CONCLUSION

The aim of this paper was to evaluate the performance of GSSHA model to simulate the soil erosion in semi-arid regions with minimum observed data. The results showed a good performance of GSSHA model to simulate soil erosion. The observed soil erosion volume was 634 m^3, while the simulated volume was 682 m^3 for the valley that has observed data. The results showed a good accuracy of the proposed GEP model to develop a mathematical expression for predicting soil erosion in ungauged basins using morphometric properties to these basins. The proposed methodology in this study may help in solving the obstacles and problems facing designers and engineers when calculating the volume of eroded soil in the ungauged valleys that constitute the majority of the world regions, especially the Middle East. The researcher suggests expanding the study area to include other areas of northern, central and southern Iraq and then developing a new equation to estimate the amount of sediments to any valley in Iraq.

REFERENCES

AL-Juboori, A.M., and Guven, A., (2016). "Hydropower plant site assessment by integrated hydrological modeling, gene expression programming and visual basic programming". Wat. Res. Manage., 30 (7), 2517-2530.

Aytek, A., Kisi O., and Guven, A. (2014). 'A genetic programming technique for lake level modeling". Hydrol. Res., 45 (4-5), 529-539.

Azamathulla, H. M., Ghani, A.A., Leow, C.S., Chang C.K., and Zakaria N.A. (2011). "Gene- expression programming for the development of a stage discharge curve of the Pahang river". Wat. Resour. Manage., 25 (11), 2901-2916.

Booker, D., and Snelder, T. (2012). "Comparing methods for estimating flow duration curves at ungauged sites". Journal of Hydrology, 434, 78-94.

Downer, C.W., and Ogden, F.L. (2004). "GSSHA: Model to simulate diverse stream flow producing processes". Journal of Hydrologic Engineering, 9 (3), 161-174.

Downer, C.W., and Ogden, F.L. (2006). "Gridded surface sub-surface hydrologic analysis (GSSHA) user's manual". US Army Corps of Engineers. Engineering Research and Development Center.

Fernando, A., Shamseldin, A., and Abrahart, R. (2012). "Use of gene expression programming for multi-model combination of rainfall–runoff models". J. Hydrol. Eng., 17 (9), 975-985.

Ferreira, C. (2001). "Gene expression programming: A new adaptive algorithm for solving problems". Complex Syst., 13 (2), 87-129.

Gupta, S.K., Tyagi, J., Sharma, G. et al. (2019). "An event-based sediment yield and runoff modeling using soil moisture balance/budgeting (SMB) method". Water Resour Manage., 33, 3721-3741.

Guven, A., and Talu, N.E. (2010). "Gene expression programming for estimating suspended sediment yield in Middle Euphrates basin, Turkey". Clean-Soil, Air, Water, 38 (12), 1159-1168.

Hashmi, M., and Shamseldin, A. (2014). "Use of gene expression programming in regionalization of flow duration curve". Advances in Water Resources, 68, 1-12.

Ijam, A.Z., and Al-Mahamid, M.H. (2012). "Predicting sedimentation at Mujib dam reservoir in Jordan.' Jordan Journal of Civil Engineering, 6 (4), 448-463.

Martínez-Salvador, A., and Conesa-García, C. (2020). 'Suitability of the SWAT model for simulating water discharge and sediment load in a karst watershed of the semi-arid Mediterranean basin". Water Resour. Manage., 34, 785-802.

Mohammad, M.E. (2005). "A conceptual model for flow and sediment routing for a watershed, northern Iraq". Ph.D. Thesis, University of Mosul, Iraq.

Seckin, N., and Guven, A. (2012). "Estimation of peak flood discharges at ungauged sites across Turkey". Water Resour. Manage., 26 (9), 2569-2581.

Sharif, H.O., Al-Zahrani, M., and El Hassan, A. (2017). "Physically, fully-distributed hydrologic simulations driven by GPM satellite rainfall over an urbanizing arid catchment in Saudi Arabia". Water, 9 (163), 1-19.

Shoaib, M., Shamseldin, Y.A., Melville, W.B., and Khan, M.M. (2015). "Runoff forecasting using hybrid wavelet gene expression programming (WGEP) model". J. Hydrol., 527, 326-344.

Sith R., and Nadaoka K. (2017). "Comparison of SWAT and GSSHA for high time resolution prediction of stream flow and sediment concentration in a small agricultural watershed". Hydrology, 27 (4), 1-15.

Zorn, C.R., and Shamseldin, A.Y. (2015). "Peak flood estimation using gene expression programming". Journal of Hydrology, 531 (3), 1122-1128.