

Record Length Needed to Predict Suspended Sediment Load from Stream Discharge Data Using Linear Regression

Isam E. Amin * and Alan M. Jacobs

Department of Geological and Environmental Sciences,
Youngstown State University, One University Plaza, Youngstown, Ohio 44555, USA

*Email: ieamin@ysu.edu

Email: amjacobs@ysu.edu

ABSTRACT

Although it has been established that suspended sediment load of streams can be derived from stream water discharge data, it is assumed that the longer the record length, the better the correlation. Using linear regression analysis, we applied this technique to four diverse US rivers, with sediment record lengths of 30 to 54 years. High correlation coefficients (0.72 to 0.97) showed high accuracy whereas predicted minus observed suspended sediment loads for each river re-calculated for records in increasing increments of 5 to 10 years in length showed high precision regardless of the record length. Consequently, record lengths more than 5 to 10 years were considered unnecessary.

Keywords: Linear regression analysis, suspended sediment load, river discharge, correlation coefficients, predicted vs. observed sediment load, accuracy vs. precision

Mathematics Subject Classification: 62J05

1. INTRODUCTION

Excessive suspended sediment load of streams, present in 25 percent of stream lengths in the United States, impacts the environment negatively (US EPA 2006, Botkin and Keller 2005). Suspended sediment load includes only insoluble clasts that a stream carries above the channel bed. It excludes dissolved load (solutes) and bed load (clasts that are dragged or skip along the bottom of the channel). Erosion of river and stream banks often accounts for about 50% of the transported sediment (Svendsen et al. 1995). The critical role played by sediment transporting non-point source pollution to stream channels is a growing concern (Russell et al 2001).

Prediction of sediment loads from stream-water discharge has been accomplished by regression analysis, that is, matching directly measured sediment loads with discharge and developing mathematical relationships that can be applied to streams in general (e.g., Jain 2001, Walling 1977, Linsely et al 1981, Asselman 2000, Horowitz 2003).

Researchers applied various order regression analyses (linear, logarithmic, polynomial, etc.) to develop sediment rating curves. They compared frequency of data collection using daily, weekly, monthly, and annual data sets. At the low end (daily) loads fluctuate too broadly from occasional storms and at the

higher end (annual) loads are too much influenced by wet versus dry years. Some conclusions by Horowitz (2003) were that:

- Sediment rating curves tend to under-predict/ estimate high and over-predict/ estimate suspended low sediment concentrations
- Using daily and weekly concentrations have larger percent deviations (difference between observed and predicted suspended sediment loads) than quarterly or annual concentrations.
- To limit errors to under ± 20 percent, predictions should be made using quarterly or annual concentrations (for all rivers).

This study determined the minimum number of years of suspended sediment load in rivers and streams to successfully maintain accuracy and precision of the results. An intuitive assumption, incorrect in these cases, is that the greater the number of years of data collection the better the results. This study addressed the question: what is the optimum span of years?

Although accuracy of prediction has been the goal, references cited here indicated that there were streams with higher accuracies than others – i.e., with higher and lower correlation coefficients obtained from the regression analyses. An additional important statistic in this study aside from accuracy of predictions was the precision of predictions (repeating results of low accuracy or high accuracy predictions). Even with lower correlation coefficients, high precision indicates that the regression analysis method is good, but that other influences are in play. This study allowed us to pinpoint other influences rather than criticize the method.

Therefore, this study investigated the predictive correlation between suspended sediment load and stream discharge extant in a river or a stream using linear regression analysis. The objective of this study is: to evaluate the accuracy and precision of this method of prediction of various spans of time representing different sediment record lengths.

The disadvantages of using linear regression analysis are as follows:

- A change in sediment input can offset the interrelation between discharge and suspended sediment (Araujo et al. 2012).
- The daily peak water discharge may not always coincide with the daily peak sediment discharge and the sediment load can change significantly for the same water discharge (Bhowmik et al. 1980).
- Data associated with water quality conditions can be challenging and inaccuracy and discontinuous measurements can be problematic (Araujo et al. 2012).
- Regression analyses assume that the suspended load comes only from the erosion of the river channel by the channel flow. However,

sediments from the interfluvies and floodplains can be added to or removed from the channel with the same measured channel discharge (Amin and Jacobs, 2007).

2. STUDY SITES

The suspended sediment loads and stream discharge in four American rivers were investigated in this study: Maumee, Sacramento, Cuyahoga, and Mississippi. These data were from the following gauging stations: Waterville, Ohio (Maumee River), Red Bluff, California (Sacramento River), Independence, Ohio (Cuyahoga River), and McGregor, Iowa (Mississippi River). A brief description of the four rivers is given in Table 1 and a summary follows.

River	Maumee	Sacramento	Cuyahoga	Mississippi
Location	NE Indiana to NW Ohio	Northern to Central California	NE Ohio	Central U.S.
Mouth	Lake Erie at Toledo, Ohio	San Francisco Bay, California	Lake Erie at Cleveland, Ohio	Gulf of Mexico near New Orleans, Louisiana
Sampling Point (USGS Gaging Station)	Waterville, Ohio	Red Bluff, California	Independence, Ohio	McGregor, Iowa
Perennial/Ephemeral	Perennial	Ephemeral	Perennial	Perennial
Sediment Record Length (yrs)	54	54	37	30

Table 1. Rivers investigated in this study

The Maumee River is a perennial river located in northwestern Ohio. The river channel that bears the name Maumee begins at Fort Wayne, Indiana, and flows into Lake Erie at Toledo, Ohio. It is about 130 miles (209 km) long and its watershed covers approximately 5,024 square miles (13,000 sq. km) in Ohio. The Maumee River has about 4,000 miles (6,440 km) of streams, creeks, and rivers emptying into it, making it the largest watershed of any river flowing into any of the Great Lakes. The watershed is mostly made up of agricultural lands with urban development to an extent, hay and pasture lands, and forests (Ohio History Central – Maumee River, n.d; Ohio Environmental Protection Agency- Maumee River Watershed, n.d). The Maumee River watershed contributes about 50% of sediment to Lake Erie. The average precipitation in the study area is approximately 33 inches (85 cm). Rainfall is evenly distributed throughout the year (Waterville Ohio Weather 2015). Water discharge and suspended sediment record length is 54 years at the Waterville gauging station.

The Sacramento River channel with its drainage basin (area of 27,000 square miles or 70,000 sq km) is the largest in California. It is an ephemeral river in northern and central California, which flows

only after rainfall events. The river channel that bears the name Sacramento begins in northern California and ends at the San Francisco Bay. It is about 447 miles (719 km) long. Approximately one third of California's total surface runoff ends up in this river. This watershed provides drinking water, irrigation water, and water for numerous wildlife species (Mount, n.d.). Over 4/5 of the precipitation falls from November to April. The mean annual precipitation can be as high as or even above 100 inches (250 cm) and as low as and even less than 10 inches (25 cm) (Jones et al. 1972). At the gauging station used in this study (Red Bluff) the average precipitation was 24.5 inches (62 cm) (US Climate Data 2015). The annual sediment load in the river varies greatly due to the variations in streamflow (Jones et al 1972). The record length of sediment and discharge data at Red Bluff was 54 years.

The Cuyahoga River is a perennial river in northeastern Ohio. The river channel that bears the name Cuyahoga begins from Geauga County and ends in Lake Erie at Cleveland, Ohio. The river channel flows about 100 miles (161 km) in a U-shape and has a drainage area of 813 square miles (2,100 sq. km). The Cuyahoga River has largely recovered from past industrial pollution that caused it to catch fire several times in the last century. Mean yearly precipitation in the watershed is about 35 inches (90 cm), including snowfall. Approximately 20 inches (50 cm) of the precipitation normally fall between April and September. The river contributes minimal sediment load to Lake Erie (Cuyahoga Valley National Park Ohio 2013).. The record length of sediment and discharge data at Independence was 37 years.

The Mississippi River is a perennial river located in central U.S. The channel that bears the name Mississippi is 2,340 miles (3,770 km) long, second in length in North America only to the Missouri River (National Park Service, 2017). Its headwaters are in Lake Itasca, Minnesota and its mouth is at its delta as it enters the Gulf of Mexico. It drains an area of approximately 1,234,700 square miles (3,200,000 sq. km), which is about 40 percent of the area of the United States (American Rivers 2014). The Mississippi River serves as a primary drinking water source for millions of Americans and also supports a wide variety of fish and wildlife (American Rivers 2014). The watershed has an average annual precipitation of 33 inches (85 cm). Rainfall is equally distributed throughout the year. August is the wettest month of the year with an average rainfall of 4.6 inches (12 cm) (McGregor Weather 2015).

3. DATA AND METHODS

The United States Geological Survey (USGS) sediment portal website (USGS portal website) provided data sets for the four investigated rivers' gauging stations. The data sets included daily sediment load and streamflow of the rivers. Monthly and annual data were obtained as summations of daily data.

In the linear regression analysis, the sediment load was the dependent variable and the discharge was the independent variable. The analysis was performed using Microsoft Excel®. The regression equation usually takes the form:

$$S_s = a X^m \quad (\text{Equation 1})$$

where: S_s = suspended sediment load
 X = river water discharge
 a = a constant
 m = slope of the regression line

Expressed logarithmically Equation 1 can be expressed as:

$$\log S_s = m \log X + \log a \quad (\text{Equation 2}), \quad \text{which is in the form } y = mx + b,$$

where: $b = \log a$ = intercept and
 m = slope of the regression line.

Because the regression analysis used in this study involved transformation into logarithmic functions as shown by equation (2), the final suspended load values had to be retransformed back into engineering units. Unfortunately, this retransformation inserted a bias that needed to be corrected. Cohn and Gilroy (1991) proposed two methods for bias correction: (1) the “minimum variance unbiased estimator” and (2) the “smearing estimator.” Number (2) was used in this study because it is assumed that the residuals are independent, identically distributed, and can follow any distribution.

4. RESULTS

Equation (1) was used to predict the suspended sediment loads from annual, monthly, and daily data. The higher the correlation coefficient the higher the accuracy of the prediction. The correlation coefficients of the daily, monthly, and annual relationships of the four rivers investigated in this study are summarized in Table 2. In section 5.0 it will be explained that the suspended sediment load can best be predicted by using the monthly regression equation. Table 3 lists the monthly regression equations for the four studied rivers. Figures 1 through 4 show the plots of the monthly regression relationships for the four rivers.

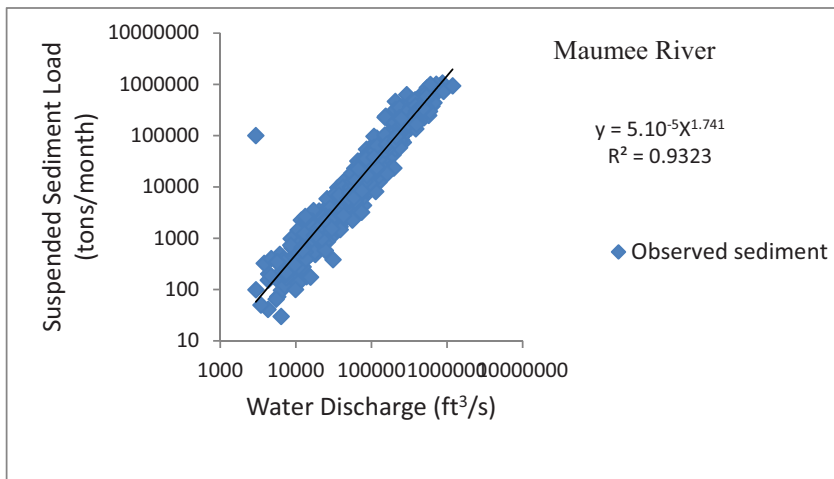


Figure 1 Regression relationship between the monthly suspended sediment load and water discharge of the Maumee River.

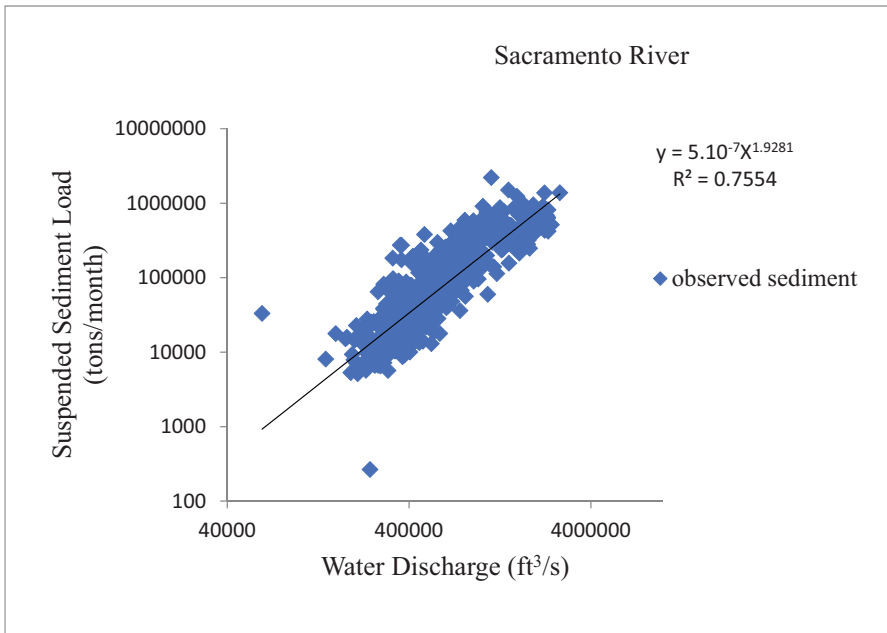


Figure 2 Regression relationship between the monthly suspended sediment load and water discharge of the Sacramento River.

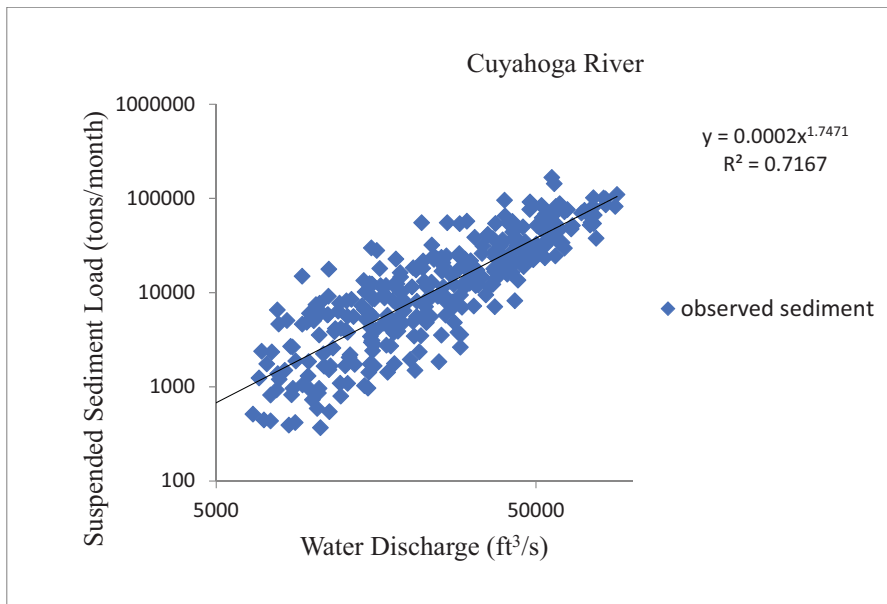


Figure 3 Regression relationship between the monthly suspended sediment load and water discharge of the Cuyahoga River.

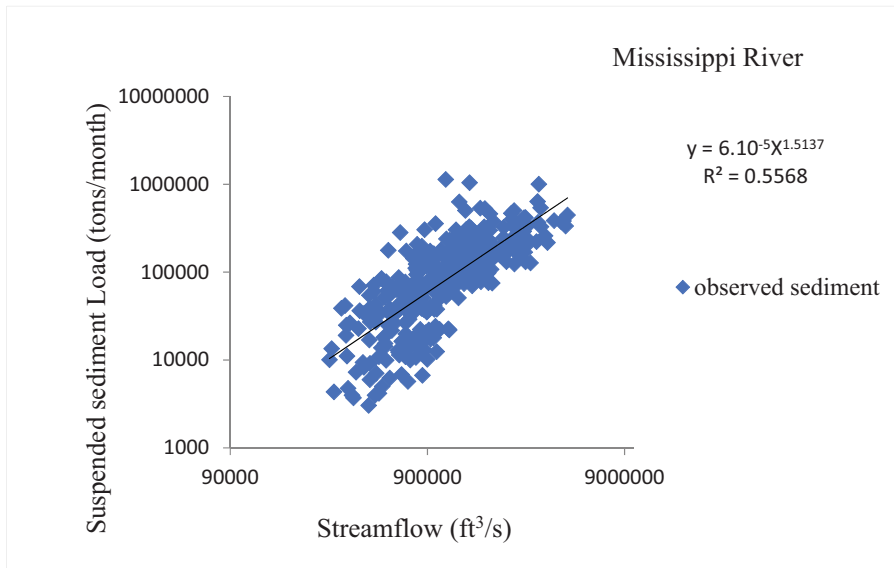


Figure 4 Regression relationship between the monthly suspended sediment load and water discharge of the Mississippi River.

The percent deviation is the observed minus the predicted suspended sediment load.
 $\{(|\text{Observed} - \text{Predicted}|)/\text{Observed}\} \times 100\% = \text{Percent Deviation (Equation 3)}$

Or simply: Deviation = $(|\text{Observed} - \text{Predicted}|)/\text{Observed}$ (Equation 4)

Deviations of similar values indicated high precision of prediction. For each investigated river, the precision of the prediction was tested using sediment data of various record spans. The percent deviation was calculated for each span and the results are included in Table 4. The deviations are graphed in Figure 5.

It is also true that the lower the percent deviation, the higher the correlation coefficient and the higher the accuracy of the prediction, as mentioned in section 4.1.

Regression Relationship	Correlation Coefficients			
	Maumee	Sacramento	Cuyahoga	Mississippi
Daily	0.96	0.89	0.90	0.72
Monthly	0.97	0.87	0.85	0.75
Annual	0.86	0.83	0.78	0.59

Table 2. Correlation coefficients for the daily, monthly, and annual relationships for the four rivers

River	Monthly Regression Equation
Maumee	$Y = 5.10^{-5} X^{1.741}$
Sacramento	$Y = 5.10^{-7} X^{1.9281}$
Cuyahoga	$Y = 0.0002 X^{1.7471}$
Mississippi	$Y = 6.10^{-5} X^{1.5137}$

Table 3. Monthly regression equations for the four rivers

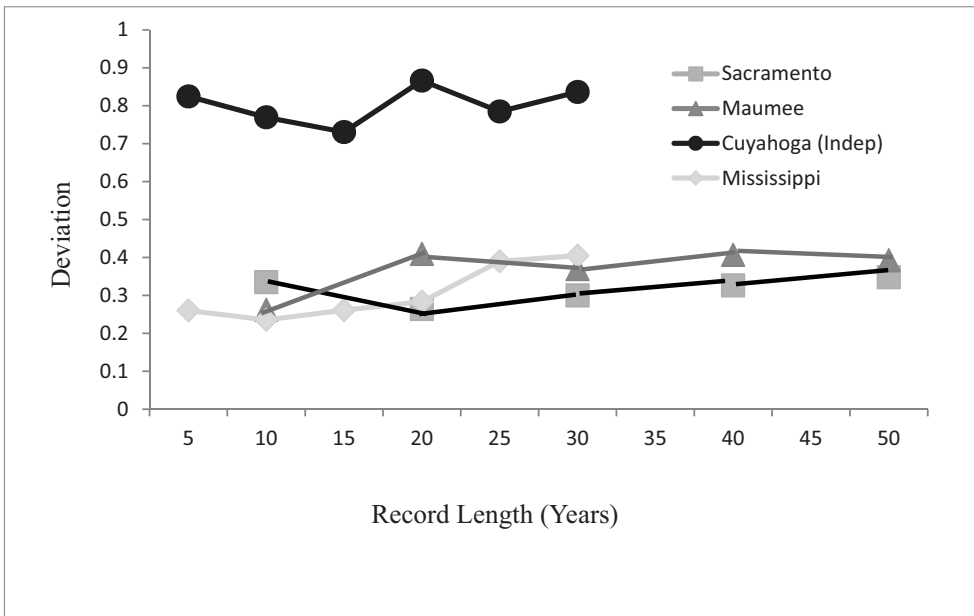


Figure 5. Graph of deviation vs. record length for the four rivers.

5. DISCUSSION

Both cause-and-effect and correlation are statistical relationships among variables. With respect to predicting sediment load that is suspended in a stream, previous studies have established a cause-and-effect relationship among various causes, stream-water discharge being one of them. Going one step further, previous studies have also established statistically a correlation between sediment load and water discharge, e.g., the higher the discharge the greater the sediment load. After establishing causality with these two variables, we could use correlation as a prediction tool.

Accuracy is the closeness of a calculated value (predicted) to the actual value (observed as measured). If these differences are small, the predictions are accurate. Accurate results would have

high correlation coefficients in linear regression analyses. Precision, on the other hand, is the closeness of all values over time to each other, regardless of the closeness to the true values. If the differences among the predicted and measured sediment loads for various lengths of the data record (e.g., 5, 10, 15 years) are similar (if not equal), then the prediction is precise. The ideal analysis is both accurate and precise.

Sediment load carried by a stream is influenced by four factors, as summarized by Bhowmik et al. (1980):

- 1) watershed soils, vegetative cover, and farming practices.
- 2) rainfall and snow and ice melt.
- 3) land use (rural and urban), the channel bed and bank sediments, soil thickness, bank stabilization, and tributaries of the main river.
- 4) engineering alteration of the channel (including dams).

All four factors will affect the results of linear regression, namely the correlation coefficients and percent deviations.

In this study, using linear regression with smearing corrections, the four investigated rivers showed high correlation between the independent (water discharge) and dependent (sediment load) variables. The correlation coefficients of the four rivers varied from 0.72 to 0.97, with a median of 0.85 (Table 2). The high correlation coefficients indicated high prediction accuracy.

It is interesting to note that the Mississippi River showed lower correlation coefficients than the other three rivers, probably because of the huge area of this drainage basin encompassing many climatic zones and various geologic units. As indicated in section 2.4, the Mississippi River has a drainage area which occupies about 2/5 of the United States. The three other rivers had smaller drainage basins, so the factors affecting sediment load within their drainage basins were not as variable.

Raw data were those values collected from USGS daily measurements. We also worked with values of monthly compilations of daily data and annual compilations of monthly data. Linear regression worked best with daily data and monthly compiled values, rather than with compiled annual values for the following reason. With the many data points for daily data, deviations on one side of the regression line were balanced with other data points on the other side of the regression line, producing the highest correlation coefficients. This was also true when performing linear regression using monthly compiled values of daily data. When calculating annual values based on monthly values, fewer data points were less likely to be balanced on either side of the regression line, producing lower correlation coefficients, probably resulting from differences in weather events from year to year. Therefore, the regression relationships using compiled annual values were significantly different as they were produced by processes that were subjected to different intensities of hydrologic changes. The processes were affected by the four factors mentioned above. The processes of the annually compiled values were subjected to large hydrologic changes that occurred during the entire year, which may change from year to year, whereas those of the monthly compiled values were subjected to lesser hydrologic

changes that occurred during each month. Thus, the monthly compiled values seemed to be the logical choice for predicting the suspended sediment load using linear regression. Therefore, this study proposed application of linear regression to the monthly compiled values of daily data for the prediction of the suspended sediment load in rivers and streams as a function of water discharge.

Each of the four rivers showed similar values of deviations between predicted and observed suspended sediment loads (Table 4 and Figure 5). This means that all four rivers had high precision of prediction. For all increments of sediment data record lengths, three rivers (Maumee, Sacramento, and Mississippi) resulted in lower deviation values (Table 3 and Figure 5) than the Cuyahoga River, indicating that the Cuyahoga had lower accuracy than the other three rivers. Therefore, the results of this study clearly revealed that a daily sediment record length of 5 to 10 years was sufficient for using linear regression to predict the suspended sediment load with high accuracy and precision.

The low accuracy of the Cuyahoga showed that other local factors were influencing the outcome of the Cuyahoga prediction. Discussion of these factors is given in the following paragraph.

This linear regression analysis assumes that the suspended sediment load is derived from erosion within the stream channel by the flow of the stream. Some of the sediments, however, can be added to or removed from the channel on the interfluvies and floodplains. We suspect that this was the case with the Cuyahoga at the sampling station used in this study (Independence, Ohio). Intensive urban development in the area of this sampling station produced additional sources of sediment. This local factor was supported by preliminary results on deviation calculations on the Cuyahoga River at another gauging station (Old Portage, Ohio) south of the station used in this study (Independence, Ohio) and farther south from urban (Cleveland, Ohio) development. The deviation at Old Portage station for the 10-year total record length was lower than for the rivers in our study and its correlation coefficient was higher. This gauging station was not chosen for this study because of its mere 10-year record length.

We present one final note on the sediment record length. Leopold and Maddock (1953) used linear regression to predict the suspended sediment load in some ephemeral streams in the western United States. Their data record spanned shorter periods, 30 days or less. Their regression equations resulted in slope values from 1.09 to 1.58 (median of 1.29). Fifty years later, Amin and Jacobs (2007) used the same approach to predict the suspended sediment load in the Rio Puerco, an ephemeral stream also located in the western U.S. in the state of New Mexico. The Rio Puerco data covered a period of 14 years. The resulting slope value was 1.37, which compared very well with the range (1.09 – 1.58) reported by Leopold and Maddock. Therefore, the two studies yielded similar slope values regardless of their sediment record lengths. The agreement of the results clearly showed that the regression relationship between the suspended sediment and stream discharge was non-random.

River		Sediment record length, (Years)	Years	Span (Years)	Deviation (Observed-Predicted)/Observed)
1	Maumee River at Waterville, Ohio	54	1950-1959	10	0.26
			1950-1969	20	0.41
			1950-1979	30	0.37
			1950-1989	40	0.41
			1950-1999	50	0.39
2	Sacramento River near Red Bluff, California	54	1960-1969	10	0.33
			1960-1979	20	0.26
			1960-1989	30	0.30
			1960-1999	40	0.33
			1960-2009	50	0.35
3	Cuyahoga River at Independence, Ohio	37	1972-1976	5	0.82
			1972-1981	10	0.77
			1972-1986	15	0.73
			1972-1991	20	0.87
			1972-1996	25	0.78
			1972-2001	30	0.84
4	Mississippi River at McGregor, Iowa	30	1975-1979	5	0.20
			1975-1984	10	0.25
			1975-1989	15	0.26
			1975-1994	20	0.26
			1975-1999	25	0.38
			1975-2004	30	0.40

Table 4. Assessment of deviation based on sediment record length for the four rivers.

6. CONCLUSIONS

Correlation of observed, monthly suspended sediment load (dependent variable) with stream discharge (independent variable) was high using linear regression analysis with smearing corrections. Data were from four gauging stations on the Maumee, Sacramento, Cuyahoga, and Mississippi rivers. The correlation coefficients ranged from 0.72 to 0.97 (median of 0.85). Comparing deviations between predicted and observed suspended sediment load data calculated for record lengths of 30 to 54 years in increments of 5- or 10-years showed that each of the four rivers had similar values of deviations for all increments: Maumee 0.26-0.41; Sacramento 0.26-0.35; Cuyahoga 0.73-0.87; and Mississippi 0.20-0.40. The high correlation coefficients and the similar deviation values for each river indicated that prediction of sediment load from discharge data using this method is accurate and precise, respectively. Consequently, using this method, we conclude that increasing the sediment record length more than 5 to 10 years is not necessary to improve the precision or accuracy of prediction.

7. ACKNOWLEDGEMENTS

We thank Ms. Chantal W. Sakwe, who performed some preliminary data compilations and calculations for the four rivers investigated in this study in her thesis research (Sakwe 2015). Gauging station data were from the U.S. Geological Survey and provided to the public on line. We also appreciate discussions on this study from participants of the International Association of Sediments and Water (IASW) Symposium in June 2017 in Taormina, Italy.

8. REFERENCES

- American Rivers (2014) America runs on the Mississippi River. Available at: <http://www.americanrivers.org/rivers/fun/america-runs-on-the-mississippi-river/>
- Amin IE, Jacobs AM (2007) Accounting for sediment sources and sinks in the regression analysis of the suspended sediment load of streams: the Rio Puerco, New Mexico as an example. *Environmental Geosciences* 14:1-14.
- Araujo HA, Cooper AB et al (2012) Estimating suspended sediment concentrations in areas with limited hydrological data using a mixed-effects model. *Hydrological Processes* 26: 3678–3688.
- Asselman, NEM (2000) Fitting and interpretation of sediment rating curves. *Journal of Hydrology*, v. 234 pp. 228-248.
- Bhowmik NG, Bonini AP et al (1980) Hydraulics of flow and sediment transport in the Kankakee River in Illinois, Illinois State Water Survey, Report of Investigation 98p.70-94
- Botkin DB, Keller EA (2005). *Environmental Science: Earth as a living planet*, 5th Ed. John Wiley and Sons, New York.
- Cohn TA, Gilroy EJ (1991) Estimating loads from periodic records, U. S. Geological Survey Branch of Systems Analysis Technical Memo. 91.01.
- Cuyahoga Valley National Park Ohio (2013), National Park Service. Available at: <http://www.nps.gov/cuva/index.htm>
- Horowitz, AJ (2003) An evaluation of sediment rating curves for estimating suspended sediment concentrations for subsequent flux calculations. *Hydrological Processes*. v. 17, issue 17, pp 3387-3409.
- Jain SK (2001) Development of integrated sediment rating curves using ANNs. *Hydrologic Engineering* 127:30-37.
- Jones BL, Hawley NL, Crippen JR (1972) Sediment transport in the western tributaries of the Sacramento River, California. Geological Survey Water-Supply Paper 1798-J.
- Leopold LB, Maddock T (1953) The hydraulic geometry of stream channels and some physiographic implications. Geological Survey Professional Paper 252-A.
- Linsely RK, Kohler MA, Paulhus JL (1981) *Hydrology for Engineers*, 3rd Ed. McGraw-Hill, New York.
- McGregor Weather (2015). Available at: <http://www.idcide.com/weather/ia/mcgregor.htm>.
- National Park Service (2017) Mississippi. Available at: <https://www.nps.gov/miss/getinvolved/2016.htm>

Ohio Environmental Protection Agency. n.d. The Maumee River Watershed. Available at:
<http://www.epa.state.oh.us/dsw/tmdl/MaumeeRiver.aspx>

Ohio History Central. n.d. Maumee River. Available at:
http://www.ohiohistorycentral.org/w/Maumee_River?rec=2164

Russell MA, Walling DE, Hodgkinson RA (2001) Suspended sediment sources in two small lowland agricultural catchments in the UK. *Journal of Hydrology* 252:1-24.

Sakwe, CW (2015) Analysis of suspended sediment loads in streams and rivers using linear regression and Pearson correlation. Master's thesis, Youngstown State University

Svendsen LM, Kronvang B et al (1995) Dynamics of phosphorus compounds in a lowland river system: importance of retention and non-point sources. *Hydrological Processes* 9:119–142.

US Climate Data (2015) Climate Red Bluff, California. Available at:
<http://www.usclimatedata.com/climate/red-bluff/california/united-states/usca0919>.

US EPA (2006) Wadable streams assessment: a collaborative survey of the nation's streams, Washington, D.C.

United States Geological Survey Sediment Portal. Available at:
<http://cida.usgs.gov/sediment/>

Walling DE (1977) Limitations of the rating curve technique for estimating suspended sediment loads, with particular reference to British rivers. *Erosion and solid matter transport in inland waters*, IAHS Proceedings of the Paris symposium, July 1977, 122: 34–48.

Waterville, OH Weather (2015) Available at:
<http://www.idcide.com/weather/oh/waterville.htm>