Journal of Engineering and Applied Sciences 14 (Special Issue 9): 10716-10722, 2019
ISSN: 1816-949X

Evaluation of Filling Missing Data Technique in Hydraulic Series

Adnan K. Shathir
College of Engineering, University of Kerbala, Karbala, Iraq

Abstract: The missing data is one of the main problems in the hydraulic research, specially when need large amounts of data, the missing data is made the series of data discrete and not sufficient to research. This problem is very serious in Iraq because of large amount of the discrete time series due wars and other reasons. So, in order to solve this problem in Iraq the simple technique of filling missed data is chosen with the assumption that no information about the relations between the nearest stations and the target station. So, the target stations with missing data is the only series used to estimate missing data. The reliability of the values of estimated data were evaluated by verification of 30 gaging station inflow discharges for United State rivers with different locations and different hydraulic properties. In this study, the year 1975 is chosen randomly to be considered as a missing data. So, the data of this year was removed from the monthly discharge time series for each river and use three approaches of filling missed data, the first method use the mean of the month of time series as estimated value for missing month, the second method use the interpolation between two months to estimate the missing data while the third method use the nearest 2 months to modify the value of mean of missing month. Then we compare the computed values with the actual data to know how these values have percentage of error to decide how it's proper or not to use in analysis research. From the results, the mean method and modified mean method for monthly time series can be used with confidence but it depends on what reason these estimated data will be used in research analysis.

Key words: Hydraulic research, verification, estimated, sufficient, interpolation, approaches

INTRODUCTION

The missing data is one of the main problems in the hydraulic research, especially, when large amounts of data is needed. The missing data is discrete the length of series and make it not sufficient to research analysis. So, in order to solve this problem, many approaches is tried to fill the missing data. These approaches are different due the purpose of using these time series in hydraulic researches and also due to kind of time series itself (short lag time series or long lag time series). For short lag time series like daily series, the problem of missing data is enlarged because of the effect of unpredictable precipitations on discharge daily value. So, filling these missing data is have to take the probability of the rain day or not in considerations to give reality to these values (Stout, 2017). But for long lag time series like monthly time series, the problem is worth to study because the series is smoothed and effect of precipitations is vanished. Several methods are currently used to fill the missing data. Some of these techniques are complex in both analysis and collecting historical record data for stations near the station under the study which is called the target station (McLeod *et al*., 1977).

The forecasting models are one of the complex methods in analysis used to predicate the missing data (Lusajo *et al*., 2018). While the multiple regression methods is difficult because its need the availability of the data for neighboring stations to the target stations (Sattari *et al*., 2016). Also some techniques is used an expensive equipment such as Satellite Radar (Ekeu-wei *et al*., 2018). In this research, the main assumption is that no information about the neighboring stations data is known and the data of target station is available only. So, the attention is focused to the methods deal with target station only (Silva1 *et al*., 2007). The monthly discharge time series is used because it is much needed for operation of reservoirs and plan of management of water. The data is evaluated the reliability of the values of data estimated by filling of missing data by verification of 30 gaging station inflows monthly discharges for United State rivers with different locations and different hydraulic properties.

Randomly year 1975 is chosen as a missing year. Monthly data of year 1975 was removed from the data discharge series for each river and considered as a missing data from the time series and use three approaches offilling missed data, the first method use the mean of the month of time series as estimated value for missing month, the second method use the interpolation between two points to estimate the missing data while the third method use the nearest 2 months to modify the value of mean of missing month. Then, the computed values were compared with the actual data to know how these values have percentage of error to decide how it's proper or not to use in analysis research.

Data: Time series of monthly discharges for thirty gaging stations of United State rivers with different locations and different hydraulic properties was used to estimate the missed data of year 1975.

MATERIALS AND METHODS

Methodology: Three methods were used to fill missed data as mention bellow:

The mean method: In this method the estimated value for missing month (i) equal to the average mean of the same month (i) for the whole time series as indicated:

$$\text{Missed value for month}\left(\text{i}\right)=\frac{\sum_{k=a}^{b}\text{month}\left(\text{i}\right)\text{of year}\left(\text{k}\right)}{\left(\text{b-a+1}\right)},\ \text{for i}=1,2,3,...,12 \quad (1)$$

Where:

a : The first year of time series
b : The last year of time series

With important note that the time series must ignored year with missed data.

The interpolation method: In this method, the estimated value for missing month (i) is value estimated from interpolation of two values, the first is the month (i) of previous year and the second is the same month (i) for the next year to missing year. As indicated in Eq. 2:

$$\text{Missed month}\left(\text{i}\right)=\frac{\text{month}\left(\text{i}\right)\text{of year}\left(\text{k-1}\right)+\text{month}\left(\text{i}\right)\text{of year}\left(\text{k+1}\right)}{2} \quad (2)$$

The modify mean: In this method, the estimated value for missing month (i) equal to the average mean of the same month (i) for the whole time series multiply by a factor (f) which is equal to average deviation of values of previous 2 months from their means as indicated in Eq. 3:

$$\text{Missed month}\left(\text{i}\right)=\text{f. Mean value of month}\left(\text{i}\right) \quad (3)$$

Where:

$$\text{f}=0.5\left(\frac{\text{month}\left(\text{i-1}\right)}{\text{mean of month}\left(\text{i-1}\right)}+\frac{\text{month}\left(\text{i-2}\right)}{\text{mean of month}\left(\text{i-2}\right)}\right) \quad (4)$$

Two methods of checking are used to select the proper method of filling. The first one is the Nash-Sutcliffe model Efficiency coefficient (NSE) which is commonly used in hydrological discharge models (Sattari *et al*., 2016). It is represented as:

$$\text{NSE}=1-\frac{\sum_{i=1}^{n}\left(\left(\text{Xobs}\left(\text{i}\right)\text{-Xest}\left(\text{i}\right)\right)\text{\^{}2}\right.}{\sum_{i=1}^{n}\left(\left(\text{Xobs}\left(\text{i}\right)\text{-mean of Xobs}\left(\text{i}\right)\right)\text{\^{}2}\right.} \quad (5)$$

Where:

xobs (i) : The actual observation value at time i
Xest (i) : The estimated value at time i

While the second method is the Root Mean Square Error (RMSE) with the formula:

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}\left(\left(\text{Xobs}\left(\text{i}\right)\text{-Xest}\left(\text{i}\right)\right)\text{\^{}2}\right.}{\text{n}}} \quad (6)$$

RESULTS AND DISCUSSION

Time series of inflow discharges for thirty gaging stations of United State rivers with different locations and different hydraulic properties was used to estimate the missed data. By three methods, the mean the interpolation and modified mean for the year 1975 was estimated. Figure 1-30 shows these results with comparison with the actual value of year 1975 as seen bellow:

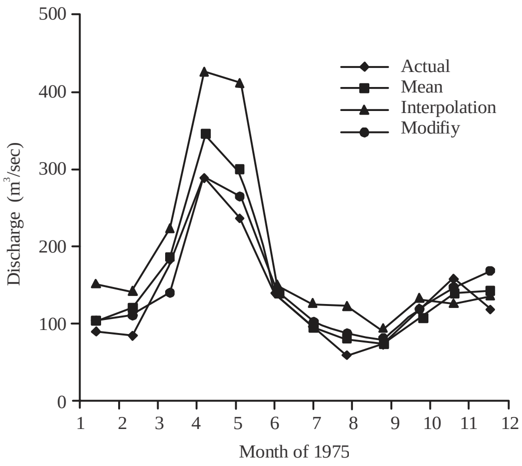

Fig. 1: Comparison of estimated value for Androscoggin river

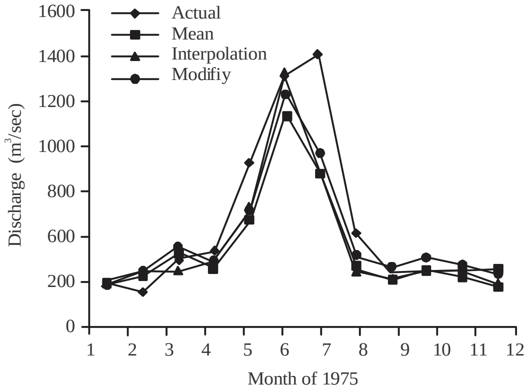

Fig. 2: Comparison of estimated value for Yellowstone river

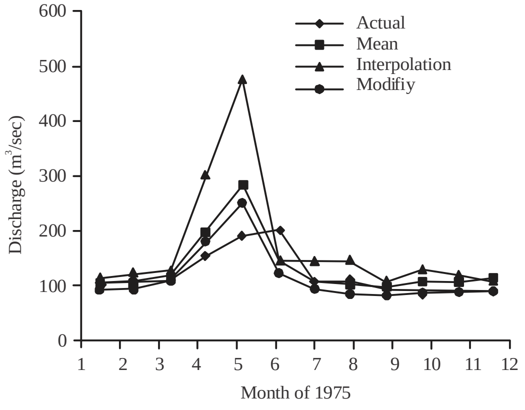

Fig. 3: Comparison of estimated value for Kennebec river

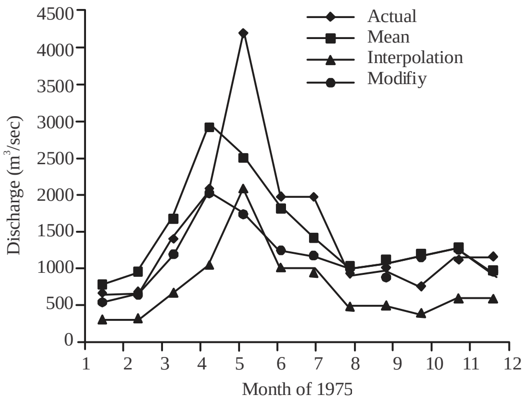

Fig. 6: Comparison of estimated value for Mississippi river

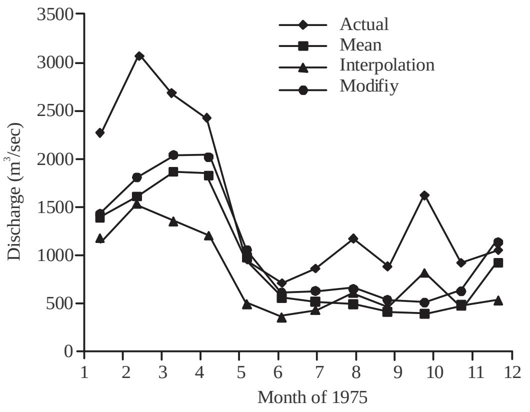

Fig. 4: Comparison of estimated value for Alabama river

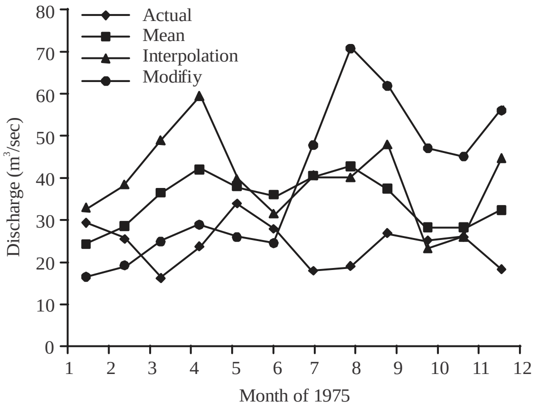

Fig. 7: Comparison of estimated value for Arizona river

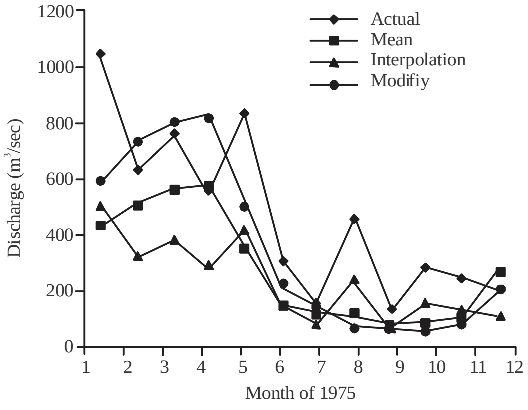

Fig. 5: Comparison of estimated value for Pearl river

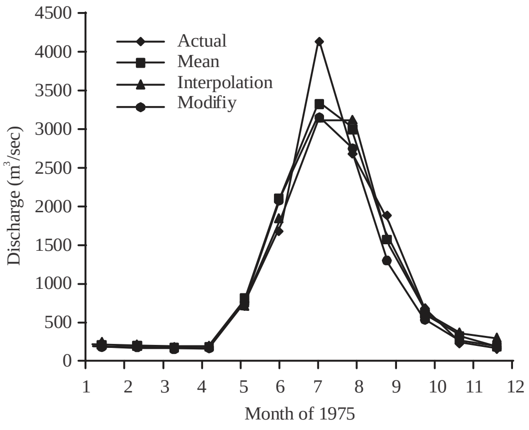

Fig. 8: Comparison of estimated value for Copper river

From the above figures it seems that 21 from 30 rivers gauging stations shows good results for estimated values of missed values with respect to actual values for the mean method and modified mean method as shown in Fig. 1, 2, 3, 6, 8, 9, 10, 11, 14, 15, 16, 17, 18, 19, 22, 24, 26, 27, 28, 29 and 30. This represent 70% of the total rivers used in the research. While 9 rivers gauging stations shows large differences between estimated and actual values. As shown in Fig. 4, 5, 7, 12, 13, 20, 21, 23 and 25. Three main reasons may be considered to understand these results, the first one is

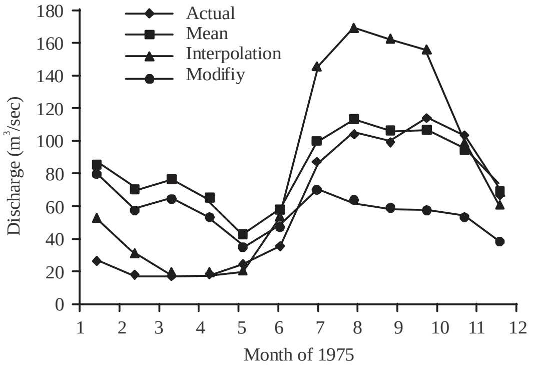

Fig. 9: Comparison of estimated value for St. Johns river

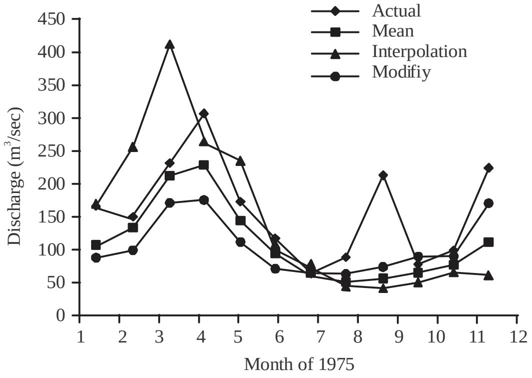

Fig. 10: Comparison of estimated value for Grand river

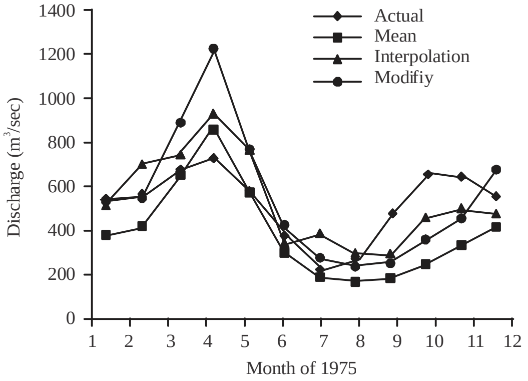

Fig. 11: Comparison of estimated value for Hudson river

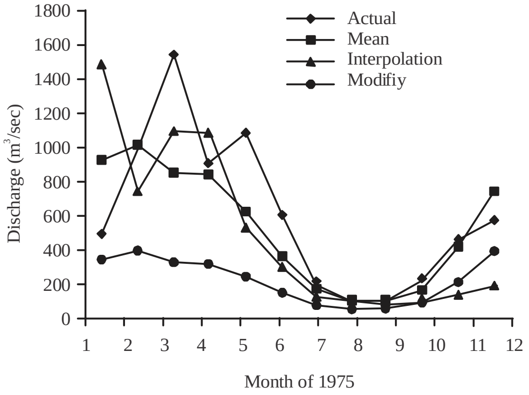

Fig. 12: Comparison of estimated value for Klamath river

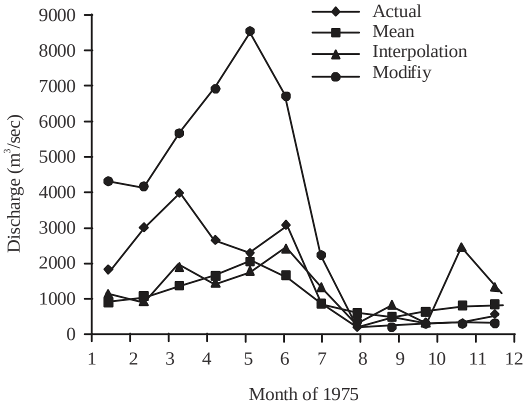

Fig. 13: Comparison of estimated value for Arkansas river

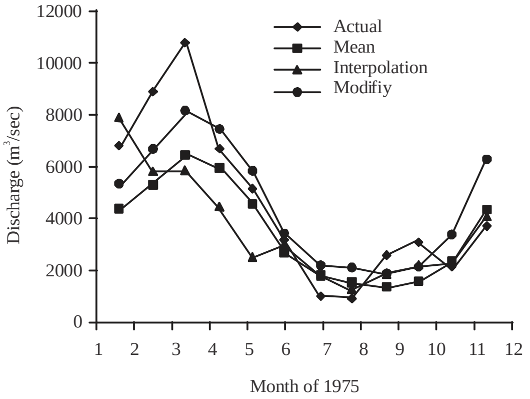

Fig. 14: Comparison of estimated value for Ohio river

sensitivity of the modified method to the values of the nearest 2 months to the missed value, so, if the two month are much higher from the average mean, the estimated values for the whole missed months will become higher than actual and if the 2 months is much lower than their means, the estimated values will become lower than actual for the whole missed months. The second reason is may be due to the decision maker effect of operation of dams or reservoirs upstream the station with missed data. For this research we know nothing about historical record of events upstream these stations which give us the wright answer about the human effect on raw data. The last reason may be due to the year itself which is considered dry year or wet year.

Table 1: The values of RMSE and NSE for 30 rivers for suggested methods

	RMSE			NSE		
River name	Mean	Interp.	Modify	Mean	Interp.	Modify
Androscoggin	33.90	87.80	27.30	0.00	-5.70	0.35
Yellowstone	222.39	198.07	175.17	0.00	0.21	0.38
Kennebec	34.64	97.07	29.27	0.00	-6.85	0.29
Alabama	738.57	865.82	630.15	0.00	-0.37	0.27
Pearl	275.56	272.48	236.27	0.00	0.02	0.26
Mississippi	622.88	878.24	795.09	0.00	-0.99	-0.63
Colorado	13.77	19.80	24.88	0.00	-1.07	-2.27
Copper	324.53	363.95	369.02	0.00	-0.26	-0.29
St. Johns	33.23	35.31	39.25	0.00	-0.13	-0.39
Grand	65.22	98.37	69.49	0.00	-1.28	-0.14
Hudson	196.61	146.43	207.80	0.00	0.45	-0.12
Klamath	289.46	410.40	520.18	0.00	-1.01	-2.23
Arkansas	1138.24	1216.46	2642.41	0.00	-0.14	-4.39
Ohio	1950.58	2113.30	1527.62	0.00	-0.17	0.39
Missouri	376.04	521.93	238.10	0.00	-0.93	0.60
Tennessee	1256.76	1280.65	1549.97	0.00	-0.04	-0.52
Pee Dee	222.30	267.19	201.16	0.00	-0.44	0.18
Brazos	177.87	304.73	451.20	0.00	-1.94	-5.43
James	141.41	189.01	177.51	0.00	-0.79	-0.58
Roanoke	195.49	223.47	218.00	0.00	-0.31	-0.24
Trinity	217.06	318.79	434.96	0.00	-1.16	-3.02
Yukon	1137.52	1681.65	1148.95	0.00	-1.19	-0.02
Sabine	212.16	196.43	164.70	0.00	0.14	0.40
Sacramento	219.84	300.30	347.21	0.00	-0.87	-1.49
Eel	338.58	315.56	459.70	0.00	0.13	-0.84
Tanana	118.95	224.16	102.38	0.00	-2.55	0.26
Delaware	169.86	126.26	173.04	0.00	0.45	-0.04
Wabash	348.15	243.57	265.60	0.00	0.51	0.42
Potomac	225.25	289.36	199.59	0.00	-0.65	0.21
Fox	36.74	50.74	39.16	0.00	-0.91	-0.14

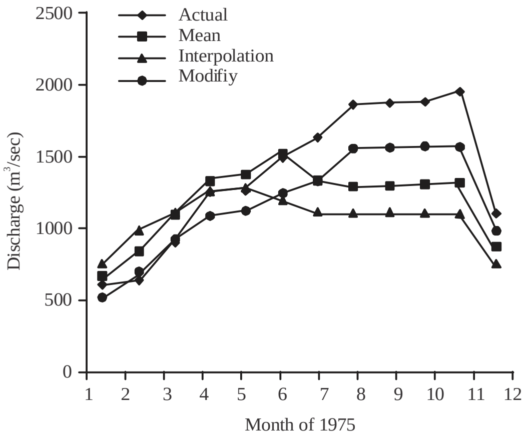

Fig. 15: Comparison of estimated value for Missouri river

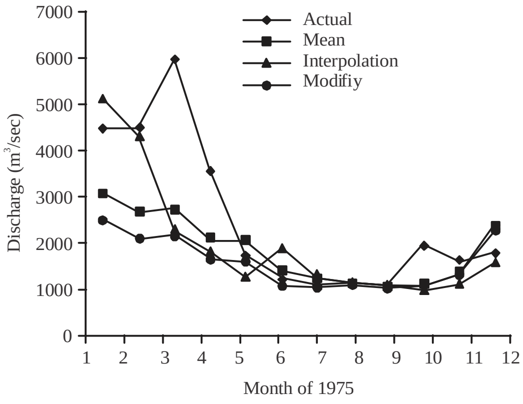

Fig. 16: Comparison of estimated value for Tennessee river

From the results shown in Table 1 the mean method and modified mean method can be used to fill missed data for monthly time series with confidence but it depends on what reason we will use these data estimated in research analysis. These two methods are easy and simple to use to give an acceptable values of missed data but for more accurate values we can use regression method between two nearest stations to fill the missed data.

But for short period time series like weekly or daily discharge time series, the mean method will give more deviation between actual and estimated values because of precipitation effect which is being smoothed in monthly time series. The difficulty of using mean method for precipitation is answering the question if the value of zero precipitation is missed data or not in time series? So, advanced methods must be used for weekly and daily time series like time series generation method or ARIMA forecasting method to fill missed data (Adnan *et al*., 1988).

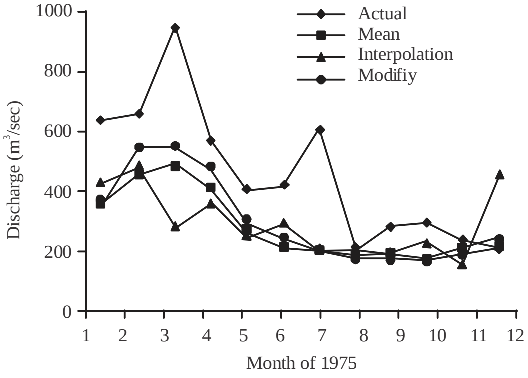

Fig. 17: Comparison of estimated value for Pee Dee river

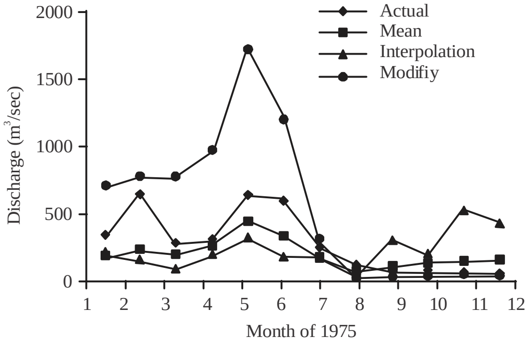

Fig. 18: Comparison of estimated value for Brazos river

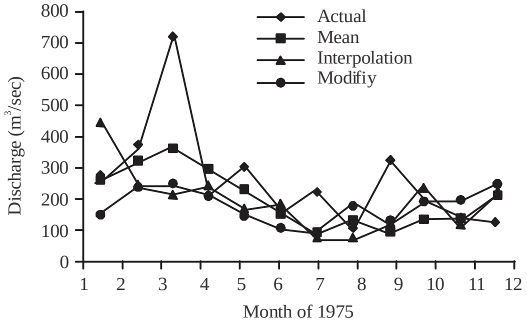

Fig. 19: Comparison of estimated value for James river

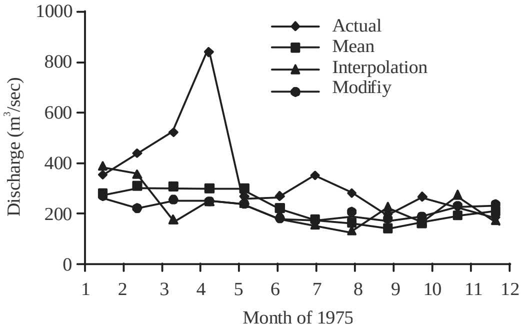

Fig. 20: Comparison of estimated value for Roanoke river

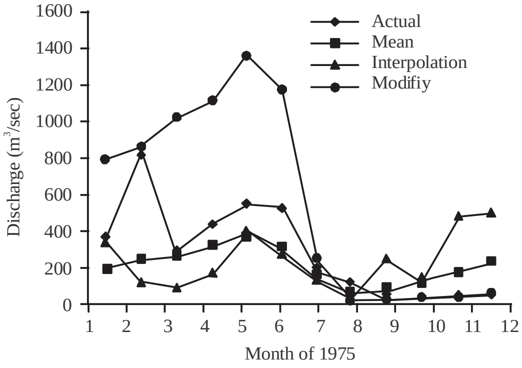

Fig. 21: Comparison of estimated value for Trinity river

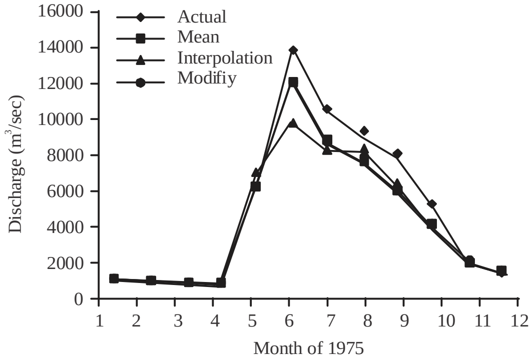

Fig. 22: Comparison of estimated value for Yukon river

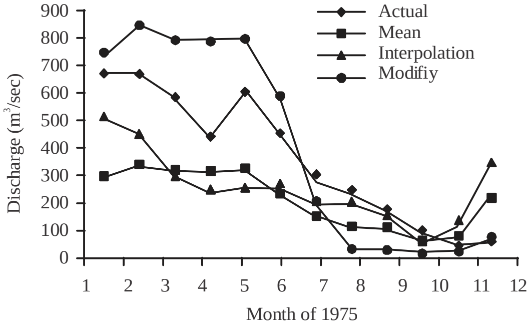

Fig. 23: Comparison of estimated value for Sabine river

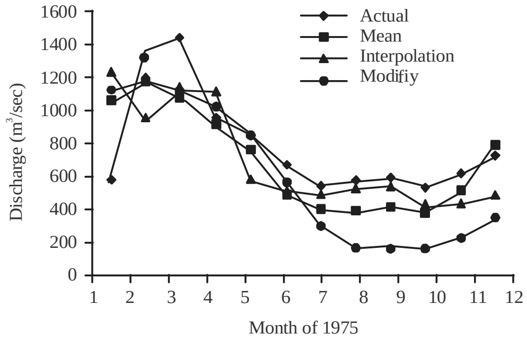

Fig. 24: Comparison of estimated value for Sacramento river

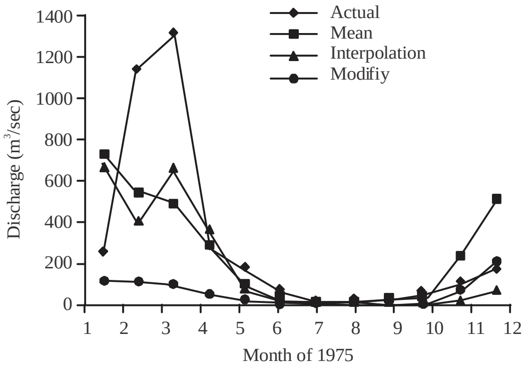

Fig. 25: Comparison of estimated value for Eel river

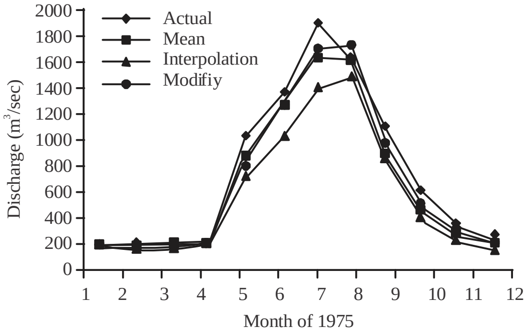

Fig. 26: Comparison of estimated value for Tanana river

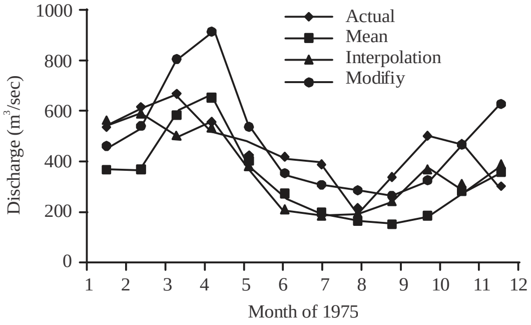

Fig. 27: Comparison of estimated value for Delaware river

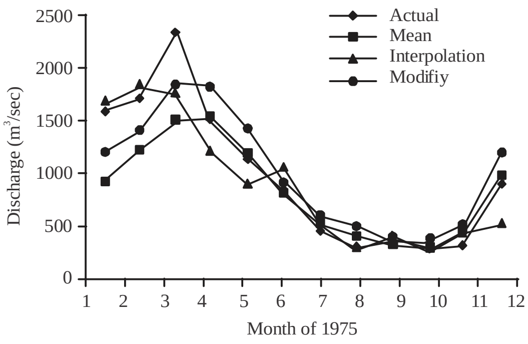

Fig. 28: Comparison of estimated value for Wabash river

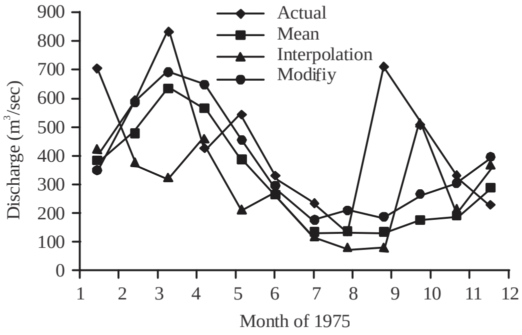

Fig. 29: Comparison of estimated value for Potomac river

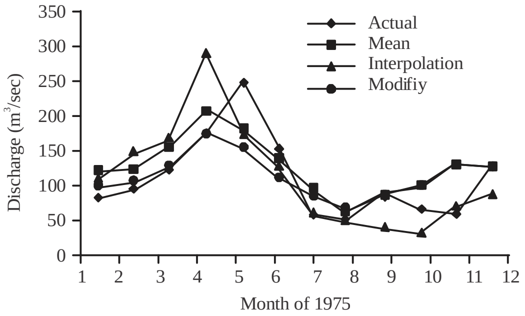

Fig. 30: Comparison of estimated value for Fox river

REFERENCES

Ekeu-wei, I., G. Blackburn and P. Pedruco, 2018. Infilling missing data in hydrology: Solutions using satellite radar altimetry and multiple imputation for data-sparse regions. Water, Vol. 10, No.10. 10.3390/w10101483

Lusajo, M., C.J. Salim and S. Kazumba, 2018. Estimation of missing river flow data for hydrological analysis: The case of great Ruaha river catchment. Hydrol. Current Res., Vol. 9, No 2. 10.4172/2157-7587.1000299

McLeod, A.I., K.W. Hipel and W.C. Lennox, 1977. Advances in box-jenkins modeling: 2. Applications. Water Resour. Res., 13: 577-586.

Sattari, M.T., A. Rezazadeh-Joudi and A. Kusiak, 2016. Assessment of different methods for estimation of missing data in precipitation studies. Hydrol. Res., 48: 1032-1044.

Silva, R.P.D., N.D.K. Dayawansa and M.D. Ratnasiri, 2007. A comparison of methods used in estimating missing rainfall data. J. Agric. Sci., 3: 101-108.

Stout, T.L., 2017. Development and application of hydraulic and hydrogeologic models to better inform management decisions. M.Sc. Thesis, Utah State University, Logan, Utah.