International Journal of Innovation in Science and Mathematics
Volume 3, Issue 4, ISSN (Online): 2347–9051

Application of Different Statistical Methods to Recover Missing Rainfall Data in the Klang River Catchment

Mehrdad Habibi Khalifeloo
B.Eng Student, Civil Engineering,
SEGi University

Munira Mohammad
Lecturer, Civil Engineering Department,
Faculty of Engineering & Built
Environment, SEGi University

Mohammad Heydari
Ph.D. Candidate, Faculty of Engineering,
University of Malaya

Abstract – **Most of hydrological studies are based on statistical science. The first step in water project engineering studies, agricultural development plans and such like is to use of correct data and information. However, because of various reasons, there are several gaps in available data. In this study, for finding missing value (545 records) in our data, we used the daily data from five pluviometry stations in the Klang river basin between 2005 to 2015. Statistical methods used in this study were the linear interpolation, linear regression, trendline command in Excel software and EM method in SPSS. The results obtained from this research showed that the best method for singular missing data is a linear interpolation method and the fastest way to fill the missing gaps was the use of the SPSS software. Also, if the aim is to find the relationship between variables and determine the priorities, the regression method is recommended.**

Keywords – **Data Mining, Knowledge Discovery From Data, Klang River, Interpolation, SPSS.**

I. Introduction

The basis of hydrologic studies is statistical data. Several scientists studied the time series data e.g., temperature, rainfall, sunshine hours per a day, the water level of rivers, etc. to predict weather patterns[1, 2]. These analyses trace the significant changes in hydrological parametersand warn the floods, drought and other environmental damages in advance[3]. But ina most hydrologic data such as river discharge because of not statistics data registering, deletion of false statistic and failure or wasting of measurement tools, it needs to estimate and calculate these data,besides accessibility to sufficient and accurate data from one point of view leads to reduction of study time and from the other perspective it leads to more accurate calculation of goal parameters and reduction of executive costs and future damages resulting of civil plans performances. To removing data gaps in a measurement station, statistic methods are used and have been done by the help of adjacent station's data with hydrology, climatology, and physiography similarity.

Some of the first approaches facing with lost data are to remove records and replacing it with mean or mode. These are common because of simple implementation and understanding. While these approaches may lead to several problems, as an example, removing records with lost value as the rest of the records could not be a good factor in society leads to skew in data. Furthermore, the records are valuable informational, thus its deletion means lost data will be replaced with the average of lost data or the average of available data. Since, this amount will be replaced as all of the lost amounts, average method, decreases the variance of available data in this variable. Furthermore,this method permitted variables, relationships with each other. Therefore, considering the importance of data quality, a more effective method has been provided.

There are widespread researchers about the recovery of hydrologic data. In each of these researches, a special method for recovery has been presented. According to increasing ways in recent years, some scholars have been confused to choose an appropriate method facing with lost amount's issue. One of the calculationsways can be referred to distance-based attributed methods and model-based attributed methods. As an example, there is a comparison among different recovery methods of rain statistic gaps in different time segments in central Alborz with linear regression methods,normal ratio,axis,and statistic ground. For this purpose, 18 stations with the 27 years cycle and without any gap have been chosen. Assessment's results showed the normal ratio with RMSE criterion in 69.2% of all cases as the most appropriate method [4]. Hasan and Croke [5] for estimating missing data in a series of daily precipitation in the basin, Brahmani, Rachi, India investigated the combined approach (probabilistic method for collecting data and interpolation method for matching data). They used Gamma Poisson distribution to generate the date for their problem. The results showed a relatively good approximation of the distribution in most cases except for large rainfall and rainfall away from the target stations. [6, 7] evaluated the estimation of missing hydrological data in the group form. Nguyen, Prentiss [8] used the analysis of rainfall interpolation in the Santa Barbara area, in that study, they concluded that multiple regression analysis provided better results than the inverse distance method for interpolate rainfall data.Researchers have applied various methods to recover the missing data. Some of these methods include artificial neural networks ANN [9], the regression method [10], Kriging method , the inverse distance methods [11] and the linear interpolation [12]. The purpose of this study is to estimate the rainfall missing data for the five stations in Klang River Basin between 2005 and 2015 by using proper statistical methods.

II. Case Study

The Klang River Basin flowing through Selangor and Kuala Lumpur has experienced flooding for more than a decade. As a capital city of a developing country such as Malaysia, it suffers from urbanization and a high rapid population. The catchment area of the Klang River Basin is 1288km^2 with a total stream length of approximately 120km. Located at 3°17’N, 101°E to 2°40’N,101°17’E, it covers areas in Sepang, Kula Langat, Petaling Jaya, Klang, Gombak and Kuala Lumpur. In the case study, only the upper river basin is targeted within an area of 468km^2 (Petaling Jaya, Klang, Gombak and Kuala Lumpur)[13]. Most of the flooding in the Klang River occurred from soil erosion problems and high rainfall intensity adds to the serious degrading. Since 1998, more than RM20 millionhas been spent on flood mitigation on this river. It is essential to study the rainfall intensity effect for the Klang River Basin, and the combination of radar and rain gauge will improve the rainfall estimation. Hence, it can be deployed for further hydrological work.

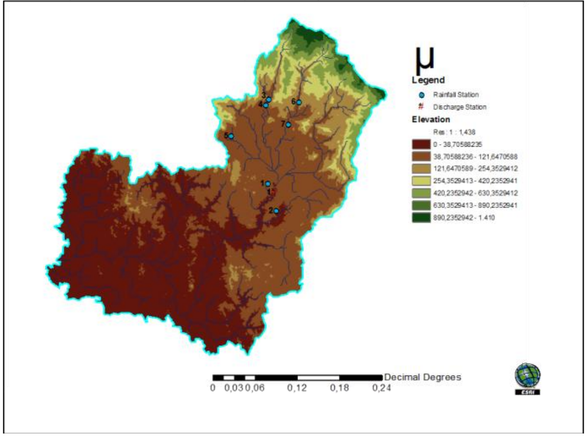


Fig.1. Schematic picture of the river basin study

Figures 2, showing time series data studied in five rain-gauge stations. It can be seen clearly the lack of time-series data in the figures.

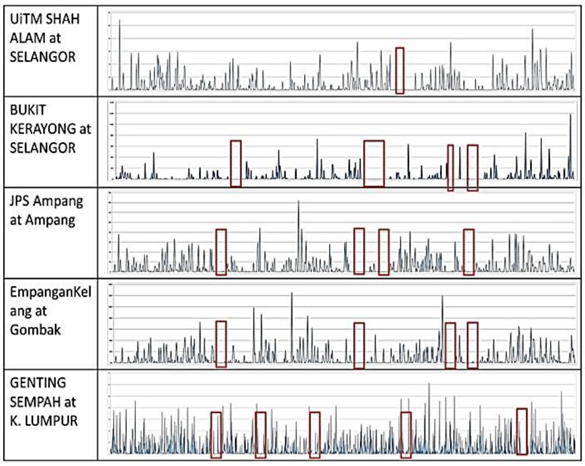


Fig.2. The daily time series rainfall data from 2005 – 2015

III. Statistical methods

A. Linear Interpolation

Data interpolation is one of routine need in most of the sciences and either technical engineering field. The interpolations which are common and useful is linear interpolation. That means that if we have the information of 0 and 1 (on following figure) and want to find the value of Y corresponding X, we draw a connector line from point 0 to 1 and calculate the value of Y.

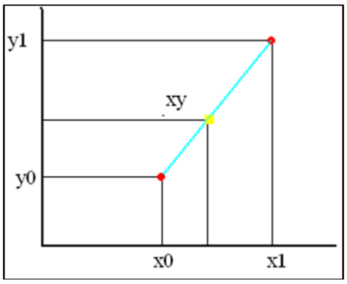


Fig.3. Schematic illustration of a linear interpolation between the two points

$$\frac{f_1(x)-f(x_0)}{x-x_0}=\frac{f(x_1)-f(x_0)}{x_1-x_0} \qquad (1)$$

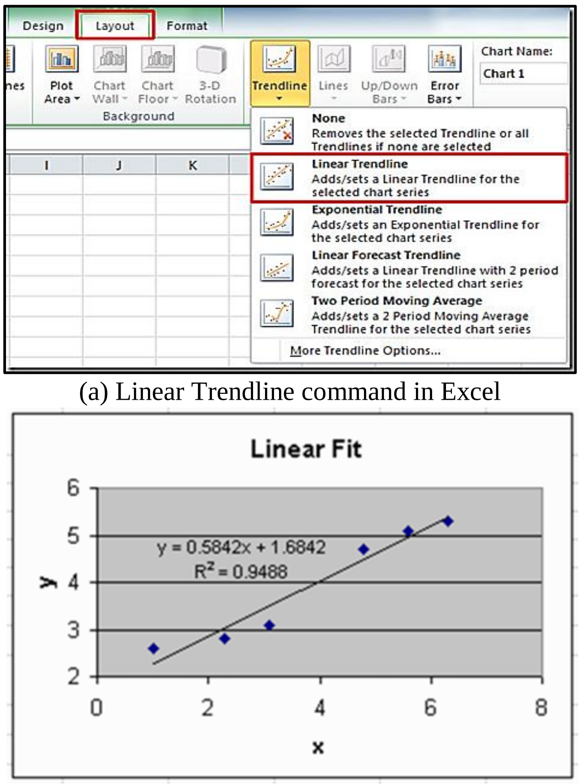
$$f_1(x)=f(x_0)+\frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0) \qquad (2)$$

B. Simple Linear Regression:

If the numbers of points were not more than 2 and there was not any necessity to cross from all points, regression can be a good solution for the problem. Generally, regression follows a mathematical relationship calculation as the quantity of an unknown variable can be determined by known variable or variables. Supposing that there is a cause and effect relationship between two quantitative variables and this relationship is in the linear form; regression equation will be like following equation.

$$Y = A + BX \qquad (3)$$

This figure shows how the linear regression equation in excel can be achieved as best linear equation of available data.



None
Removes the selected Trendline or all Trendlines if none are selected

Linear Trendline
Adds/sets a Linear Trendline for the selected chart series

Exponential Trendline
Adds/sets an Exponential Trendline for the selected chart series

Linear Forecast Trendline
Adds/sets a Linear Trendline with 2 period forecast for the selected chart series

Two Period Moving Average
Adds/sets a 2 Period Moving Average Trendline for the selected chart series

More Trendline Options...

(a) Linear Trendline command in Excel

(b) Linear Fit plot with equation and R^2

Fig.4. How to find the linear regression equation by Linear Trendline in Excel

Underneath relation shows the calculation of R^2. As the amount of R^2 closes to one, it shows that estimated equation has more accuracy.

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = 1 - \frac{SSE}{\sum(yi - \bar{y})^2} \qquad (4)$$

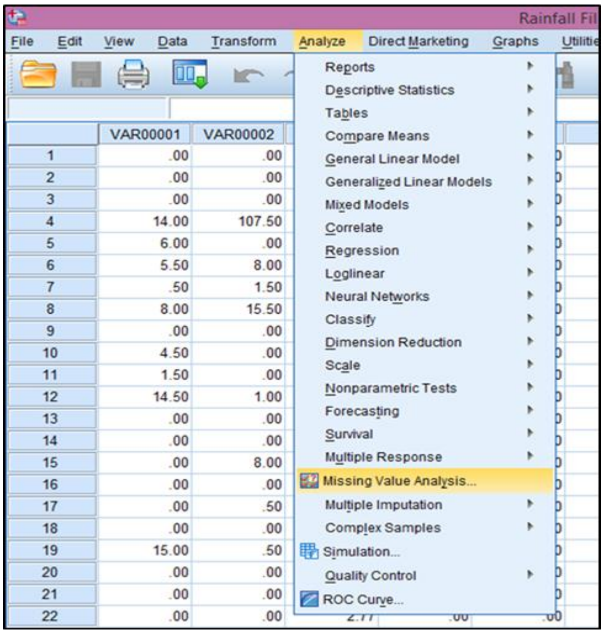
C. Algorithm of EM in SPSS:

In most researches we expose to a huge amount of data which needs huge calculation for execution of operation, too. Therefore, using appropriate statistic software is necessary. The reason of researchers' welcoming to SPSS software is providing outputs with the greater graphical environment and executing of most statistical methods without any need to programming and also having a syntax SPSS editor for professional users.

EM (expectation Maximization): In this algorithm, in order to attribute the amount of one variable, other variables are used. Then algorithm examines whether this amount is most probable one. If not, a more probable amount will be attributed. This action will be continued until achieving the most probable amount. EM is a reasonable technique which frequently applied for analyzing data in handling lost data. In contrast with two proposed methods, EM considers standard error in the equation.

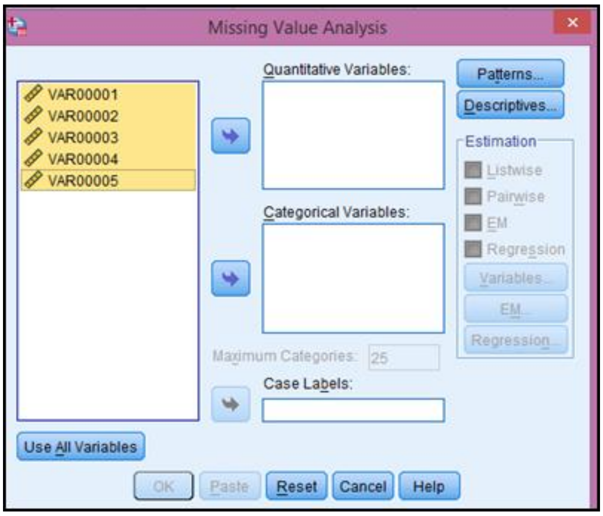
In EM, to calculate the average from variance and covariance, complete data are used. Then, ML process is employed to achieve regression lines which connect each variable to other variables. In this step, there are equations as much as numbers of variables. ML assures us that these formulas give more accurate mean, variance and covariance rather than any other formula.

The missing data recovery process in SPSS by using EM algorithm is shown in the following figure.

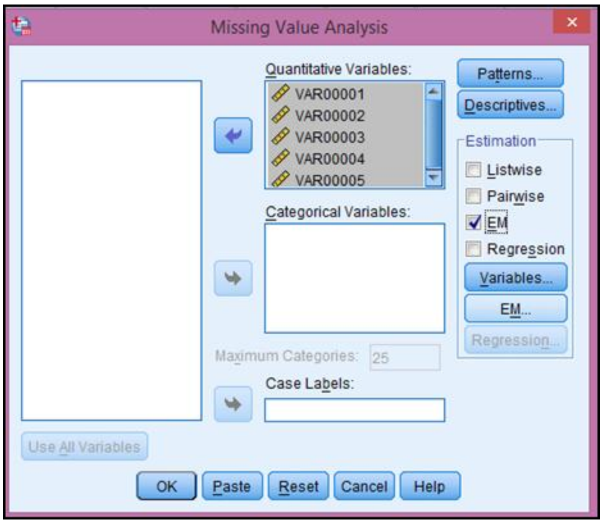


	VAR00001	VAR00002
1	.00	.00
2	.00	.00
3	.00	.00
4	14.00	107.50
5	6.00	.00
6	5.50	8.00
7	.50	1.50
8	8.00	15.50
9	.00	.00
10	4.50	.00
11	1.50	.00
12	14.50	1.00
13	.00	.00
14	.00	.00
15	.00	8.00
16	.00	.00
17	.00	.50
18	.00	.00
19	15.00	.50
20	.00	.00
21	.00	.00
22	.00	.00

The first step



The second step

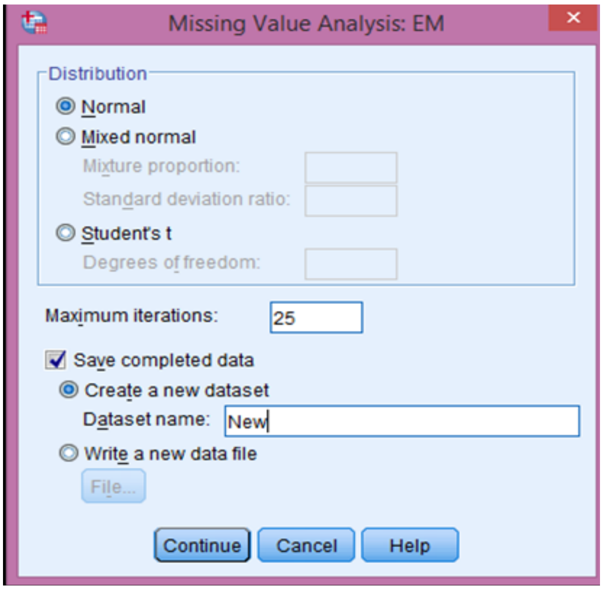


Quantitative Variables: VAR00001, VAR00002, VAR00003, VAR00004, VAR00005

Categorical Variables:

Estimation
□ Listwise
□ Pairwise
☑ EM
□ Regression

The third step



Distribution
● Normal
○ Mixed normal
Mixture proportion:
Standard deviation ratio:
○ Student's t
Degrees of freedom:

Maximum iterations: 25

☑ Save completed data
● Create a new dataset
Dataset name: New
○ Write a new data file

The fourth step

Fig.5. Recovery of lost data steps in SPSS by EM method

D. Time interval based method (Developed method):

The nearest time interval based method is the shortest time frame to neighbor with lost value. This method shows the amount of closer data to the lost data comparing with data with more farfetched time interval, is further actually. The idea of this method is like that for close years, further weights and for far years lesser weights are used. If we use time series like this research, the changes will be an irregular combination of past years instead linear one (Figure 6).

IV. Result and Discussion

Figure 6 shows the comparison interpolation methods with the nearest time interval based method (developed method). As can be seen, the irregular changes during the missing data in developed method are more similar to time series irregular changes.

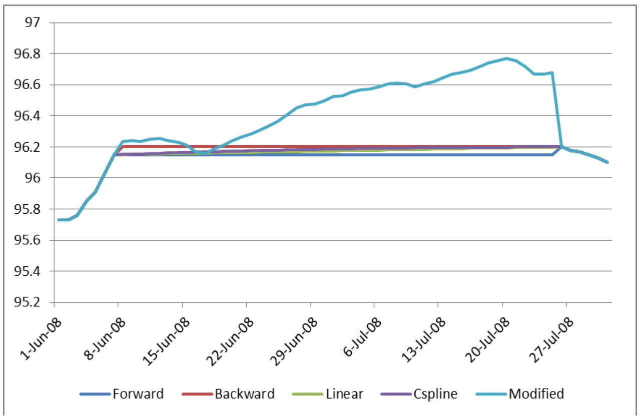


Fig.6. Comparison interpolation method (Forward, Backward, Linear and Cspline methods) with developed attribution time interval-based method (Modified method)

The results of the EM method obtained from SPSS software is given in tables below.

Table 1: Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
Rainfall 1	3651	6.56	15.64	1	0	0	521
Rainfall 2	3651	4.68	16.37	1	0	0	603
Rainfall 3	3246	8.17	15.86	406	11.1	0	421
Rainfall 4	3586	7.25	14.94	66	1.8	0	517
Rainfall 5	3581	7.11	13.76	71	1.9	0	434

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Table 2: Summary of Estimated Means

	Rainfall 1	Rainfall 2	Rainfall 3	Rainfall 4	Rainfall 5
All Values	6.56	4.68	8.17	7.25	7.11
EM	6.56	4.68	8.15	7.24	7.10

Table 3:Summary of Estimated Standard Deviations

	Rainfall 1	Rainfall 2	Rainfall 3	Rainfall 4	Rainfall 5
All Values	15.64	16.37	15.86	14.94	13.76
EM	15.64	16.37	15.83	14.93	13.76

Table 4: EM Means [a]

Rainfall 1	Rainfall 2	Rainfall 3	Rainfall 4	Rainfall 5
6.56	4.68	8.15	7.24	7.10

a. Little's MCAR test: Chi-Square = 13.682, DF = 24, Sig. = .954

Table 5: EM Covariances[a]

	Rainfall 1	Rainfall 2	Rainfall 3	Rainfall 4	Rainfall 5
Rainfall 1	244.67				
Rainfall 2	55.92	268.06			
Rainfall 3	52.56	51.47	250.52		
Rainfall 4	39.37	31.27	104.27	223.05	
Rainfall 5	36.43	28.64	77.50	74.83	189.24

a. Little's MCAR test: Chi-Square = 13.682, DF = 24, Sig. = .954

Table 6: EM Correlations [a]

	Rainfall 1	Rainfall 2	Rainfall 3	Rainfall 4	Rainfall 5
Rainfall 1	1.00				
Rainfall 2	0.22	1.00			
Rainfall 3	0.21	0.20	1.00		
Rainfall 4	0.17	0.13	0.44	1.00	
Rainfall 5	0.17	0.13	0.36	0.36	1.00

a. Little's MCAR test: Chi-Square = 13.682, DF = 24, Sig. = .954

V. Conclusion and Recommendation

1. If the time series data are available, time series analysis should be used for best accuracy.
2. If spatial data was available, GIS-based method such as the Kriging method is recommended.
3. If missing data was a single, linear interpolation method and the Trendline test are recommended.
4. If the aim is to find the relationship between variables and determine the priorities, the regression method is recommended.
5. If the aim of reaching a solution in the shortest time, the use of SPSS software is recommended.

Acknowledgment

The authors would like to thank the reviewers for their valuable suggestions that have led to substantial improvements to the article.

References

[1] Othman, F., et al., Prediction of Water Level And Salinity of Lakes By using Artificial Neural Networks, Case Study: Lake Uremia, in 35th International Association for Hydro-Environmental Engineering and Research (IAHR). 2013: China.
[2] Noori, M., et al., Utilization of LARS-WG Model for Modelling of Meteorological Parameters in Golestan Province of Iran. Journal of River Engineering, 2013. 1.
[3] Heydari, M., et al., The Necessity of Systematic and Integrated Approach in Water Resources Problems and Evaluation Methods, a Review. Advances in Environmental Biology, 2014. 8(19): p. 307-315.
[4] Sobhani, B., et al., West and south-west of the Caspian Sea rainfall modelling using spatial interpolation methods by GIS. Geography and Development 2013(Issue 30).
[5] Hasan, M. and B. Croke. Filling gaps in daily rainfall data: a statistical approach. in 20th International Congress on Modelling and Simulation. 2013.
[6] Elshorbagy, A.A., U. Panu, and S. Simonovic, Group-based estimation of missing hydrological data: I. Approach and general methodology. Hydrological sciences journal, 2000. 45(6): p. 849-866.
[7] Elshorbagy, A.A., U. Panu, and S. Simonovic, Group-based estimation of missing hydrological data: II. Application to streamflows. Hydrological sciences journal, 2000. 45(6): p. 867-880.
[8] Nguyen, R., D. Prentiss, and J. Shively, Rainfall Interpolation for Santa Barbara County. 2004.
[9] Kuligowski, R.J. and A.P. Barros, Using Artificial Neural Networks To Estimate Missing Rainfall Data1. JAWRA Journal of the American Water Resources Association, 1998. 34(6): p. 1437-1447.
[10] Haitovsky, Y., Missing data in regression analysis. Journal of the Royal Statistical Society. Series B (Methodological), 1968: p. 67-82.
[11] Teegavarapu, R.S. and V. Chandramouli, Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology, 2005. 312(1): p. 191-206.
[12] Paulhus, J.L. and M.A. Kohler, Interpolation of missing precipitation records. Mon. Wea. Rev, 1952. 80(5): p. 129-133.
[13] Habibi Khalifeloo, M., M. Mohammad, and M. Heydari, Multiple Imputations for Hydrological Missing Data by using a Regression Method (Klang River Basin). International Journal of Research in Engineering and Technology, 2015. 4(6): p. 519-524.

Author's Profile

Mehrdad Habibi Khalifeloo

is a civil engineering student in the last semester at SEGi University, Kuala Lumpur, Malaysia. He has a keen interest in research, especially in the field of hydrology. So far, three papers of him focusing on “Recovery of missing data” have been accepted in hydrologic journals and conferences.

Munira Mohammad

received her B.Sc.(Hons) in Civil Engineering from University Teknologi Mara, Malaysia in 2005, and her M.Sc. in Water Resources Engineering and Management, from University of Stuttgart, Germany, in 2008 . She is a PhD candidate in Environment and Water Engineering at University of Stuttgart. Her thesis title is “Development of Hydrological Modelling of Klang River Basin for Flood Forecasting”.

Mohammad Heydari

received his B.Sc. and M.Sc. in civil engineering at Islamic Azad University-Tehran branch in 2006 and 2009 respectively. He is a PhD candidate in water resources, completing the last semester at University Malaya (UM), Kuala Lumpur, Malaysia. His research interests are Reservoir Operation, River Engineering, Optimization, Forecasting and Time series analysis. His papers have been published in over 20 journals and conference proceedings.