Infilling Missing Rainfall and Runoff Data for Sarawak, Malaysia Using Gaussian Mixture Model Based K-Nearest Neighbor Imputation

Po Chan Chiu[1,2,3(✉)], Ali Selamat[1,2,4,5], and Ondrej Krejcar[5]

[1] School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
pcchiu@unimas.my
[2] MagicX (Media and Games Center of Excellence), Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
[3] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[4] Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
[5] Faculty of Informatics and Management, University of Hradec Kralove, Rokitanského 62, 500 03 Hradec Kralove, Czech Republic

Abstract. Hydrologists are often encountered problem of missing values in a rainfall and runoff database. They tend to use the normal ratio or distance power method to deal with the problem of missing data in the rainfall and runoff database. However, this method is time consuming and most of the time, it is less accurate. In this paper, two neighbor-based imputation methods namely K-nearest neighbor (KNN) and Gaussian mixture model based KNN imputation (GMM-KNN) were explored for gap filling the missing rainfall and runoff database. Different percentage of missing data entries were inserted randomly into the database such as 2%, 5%, 10%, 15% and 20% of missing data. Pros and cons of these two methods were compared and discussed. The selected study area is Bedup Basin, located at Samarahan Division, Sarawak, East Malaysia. It is observed that the GMM-KNN imputation method results in the best estimation accuracy for the missing rainfall and runoff database.

Keywords: GMM-KNN · KNN · Imputation · Missing rainfall · Runoff data

1 Introduction

Hydrological missing data poses a challenge in hydrological and environmental modelling. Hydrologists are often encountered problem of missing values in a rainfall and runoff database. Rainfall is the quantity of rain that falls in a location over a period of time [1]. Meanwhile runoff refers as amount of water that discharged in surface streams. Missing data occurs when data values are not available or incomplete in the database. The incompleteness of rainfall and runoff data may due to equipment

F. Wotawa et al. (Eds.): IEA/AIE 2019, LNAI 11606, pp. 27–38, 2019.
https://doi.org/10.1007/978-3-030-22999-3_3

malfunctioned, measurement errors and changes to instrumentation over time [2]. If the missing data is left untreated, it reduces the power and the precision of hydrological research. These missing data would result in uncertainty particularly water flow information and it affects the plan ahead of time to deal with extremes such as flood and climate change.

As floods become increasingly more frequent in Malaysia, the analysis of rainfall and runoff plays a significant role in the field of climatology and hydrological studies [3–6]. However, rainfall and runoff data analysis is always challenged by the shortage of consecutive data at Sarawak rivers. In many instances, while analyzing the hydrological data for Sarawak rivers, there is a shortage of rainfall records of several gauging stations at Bedup Basin, and most of these records are incomplete. In addition to that, a study by Ismail *et al.* [7] concluded that the best data treatment method of a target station is different from the other target stations in Peninsular Malaysia. Therefore, this paper explores a suitable technique for handling missing rainfall and runoff data at Bedup Basin, Sarawak.

The remaining of this work is organized as follows. Section 2 reviews the existing handling missing data techniques; Sect. 3 presents a case study; Sect. 4 describes the imputation methods and Sect. 5 reports the experimental results both on KNN and GMM-KNN imputation methods. Finally, Sect. 6 summarizes the main conclusions.

2 Related Work

Numerous techniques have been proposed to estimate missing values [7–9]. Imputation is a procedure that is used to fill in missing values with substitutes [9]. The normal ratio method (NR) and the inverse distance weighting method (IDW) are the two types of the traditional missing data handling methods. The methods are the most popular approach for estimation of missing rainfall records. Suhaila *et al.* [8] adapted the inverse distance weighting (IDW) method for estimation of missing rainfall data. The study reported that the target station could be affected most by the nearest stations. Kamaruzaman *et al.* [10] have compared different methods such as inverse distance weighted (IDW), modified correlation weighted (MCW), combination correlation with inverse distance (CCID) and averaging correlation and inverse distance (ACID) to examine the best imputation methods for treating daily rainfall at 104 stations in Peninsular Malaysia. Meanwhile interpolation techniques such as arithmetic average (AA) method, inverse distance (IDW) method, normal ratio (NR) method and coefficient of correlation (CC) method were compared in a study by Ismail *et al.* [7]. There are several shortcomings, such as the overestimation or underestimation of association among variables and lack of information available from the neighbor stations. If there is no information could be used from the neighbor stations, the mean on the same day and month but at different year will be taken as the estimation of the missing values at the missing entries. Hence, this method is less accurate and time consuming as compared to other missing data imputation techniques.

The nearest neighbor stations are progressively being used to estimate the missing values in the database. Ferrari and Ozaki [11] used the nearest neighbor station to estimate the missing data based on the statistical imputation and quality control procedures to model the drought period. Furthermore, Teegavarapu and Chandramouli [12] used the inverse distance weighting method (IDW) to estimate missing rainfall values which is based on the values recorded at all other available nearby stations.

Other than that, artificial neural networks (ANNs) have become one of the most promising tools for treating missing data problem. In a study by Dastorani *et al.* [13], artificial neural networks (ANNs) and adaptive neuro-fuzzy inference system (ANFIS) methods were proposed to predict the missing flow data using the data from neighboring sites. The study revealed that the ANFIS technique demonstrated a superior prediction of missing flow data in arid land stations. Besides that, Mispan *et al.* [14] employed Levenberg-Marquadt back propagation algorithm in predicting missing stream flow data in Langat River Basin, Malaysia. The training and validation results are satisfactory; which r values range from 0.91 to 0.97 for flow parameters.

Another approach to treat the missing data problem is using Gaussian mixture model based KNN (GMM-KNN) method. Ding and Ross [15] have proposed Gaussian mixture model based KNN (GMM-KNN) imputation method for treating missing scores in biometric fusion. In the study, Ding and Ross [15] reported that GMM-KNN method performs better than the other imputation methods such as K-nearest neighbor (KNN) method, likelihood-based method, Bayesian-based method and multiple imputation (MI) method at multiple training set sizes and missing rates because it retains the natural structure of the original dataset. On the other hands, the other imputation methods such as KNN method did not capture the shape of the original scores very well. Therefore, the study indicates that the GMM-KNN imputation method results in the best recognition accuracy in the context of multibiometric fusion. However, the GMM-KNN imputation method has not been explored in the context of rainfall and runoff in Malaysia. Therefore, this study intends to explore the estimation of missing data using hydrological data from neighboring gauging sites in Sarawak and GMM-KNN imputation method.

3 Material and Method

3.1 Study Area

The study area is located in Sungai Bedup Basin, an upstream of Sadong Basin in Sarawak as shown in Fig. 1. This basin has a maximum stream length of 10 km and is situated approximately 80 km from Kuching city.

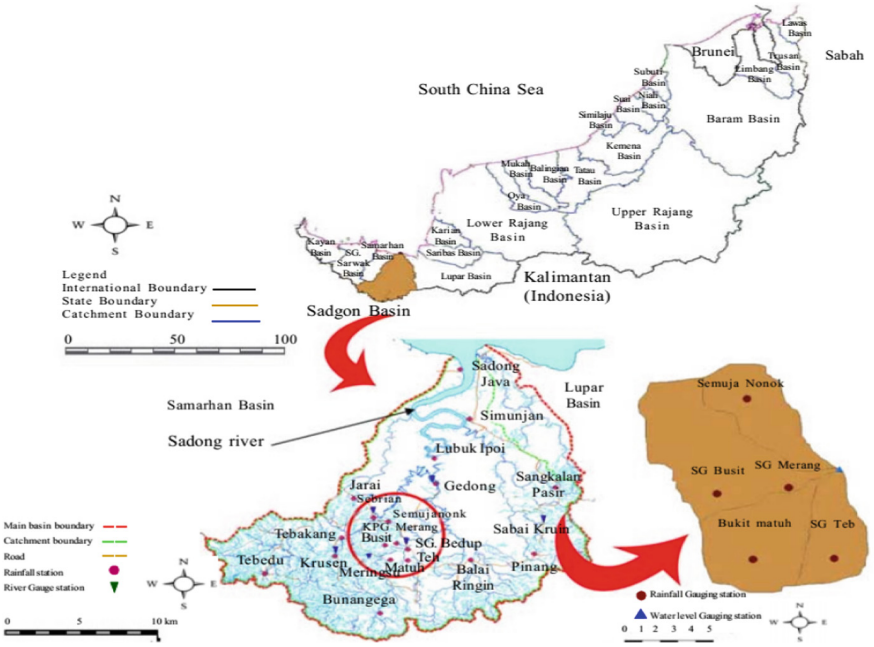


Fig. 1. Locality Map of Bedup Basin and the gauging stations [16]

Sungai Bedup Basin consists of five rainfall gauging stations and one river stage gauging station. The details of the gauging stations are presented in the following Table 1.

Table 1. Gauging stations of Bedup Basin, Sarawak.

Station name	Station number	Latitude	Longitude	Data collected
Bukit Matuh	1005079	001 03 50	110 35 35	Rainfall
Semuja Nonok	1105035	001 06 25	110 35 50	Rainfall
Sungai Busit	1005080	001 05 25	110 34 40	Rainfall
Sungai Merang	1006033	001 05 40	110 36 25	Rainfall
Sungai Teb	1006037	001 03 15	110 37 00	Rainfall
Sungai Bedup	1006028	001 05 10	110 37 50	Runoff

3.2 Data

The dataset used in this study consists of five rainfalls and one runoff data that were collected from Department of Irrigation and Drainage, Sarawak. A daily rainfall and runoff dataset consisting of 24-month records has been selected to evaluate the performance of imputation methods, as shown in Table 2. The selected dataset is prepared with some data's missing. In this work, rate-based schema has been used to randomly select a specific proportion of the entries and then removed from the complete dataset

[17]. Different percentages of missing data are inserted randomly into the dataset. The missing percentage varies as 2%, 5%, 10%, 15% and 20% missing of the total data entries. For each missing entry, a proportion of the rainfall and runoff entries will be randomly selected and removed from the dataset. According to Little and Rubin's [18] missing data mechanism, the missing value in this study has been classified as missing completely at random (MCAR). The reason is because of the occurrence of missingness in the rainfall and runoff data of the area at Bedup basin is not affected by the data in that area or any area.

Table 2. Fragment of the data from the gauging stations of Bedup Basin.

Bukit Matuh (mm)	Semuja Nonok (mm)	Sungai Busit (mm)	Sungai Merang (mm)	Sungai Teb (mm)	Sungai Bedup (m^3)
1	53.5	35.5	53.5	36	1.28
40	4.5	11	2	1	1.546
41	40	?	39.5	55	1.433
0	23	26.5	22.5	46.5	1.556
34.5	0.5	0	0	?	2.23
4	?	34	44.5	30.5	?
116.5	?	7.5	5.5	2	1.764
0	148.5	?	119.5	112.5	1.783
0	0.5	0.5	0.5	0.5	2.789
0	0.5	0	0	0	2.796

Missing data

3.3 Data Correlation Between the Investigated Stations

The correlation of daily rainfall and runoff at different stations is important to calculate the strength of a relationship between the data values. Hence, the correlation between the daily rainfall and runoff at different nearby stations was investigated (Table 3).

Table 3. Correlation matrix of investigated stations

	Bukit Matuh	Semuja Nonok	Sungai Busit	Sungai Merang	Sungai Teb	Sungai Bedup
Bukit Matuh	1.0000	0.0086	0.0043	0.0187	0.0088	0.0545
Semuja Nonok	0.0086	1.0000	0.8139	0.8538	0.7348	0.0956
Sungai Busit	0.0043	0.8130	1.0000	0.8455	0.7897	0.0874
Sungai Merang	0.0187	0.8544	0.8450	1.0000	0.8191	0.0759
Sungai Teb	0.0088	0.7366	0.7876	0.8166	1.0000	0.1042
Sg Bedup	0.0545	0.0970	0.0865	0.0931	0.1170	1.0000

As seen in Table 3, the rainfall at the Semuja Nonok is most correlated with the Sungai Busit, Sungai Merang and Sungai Teb stations respectively. The correlation of the rainfall at the Bukit Matuh station are relatively lower as compared to the rest of the rainfall stations. In addition to that, the runoff at the Sg Bedup has low correlation with all the other rainfall stations, within the range of 0.0545 to 0.1042. Generally, the rainfall and runoff correlation values in different stations are positively correlated with their respective stations.

3.4 Performance Measures

In this study, the root mean square error (RMSE) and the mean absolute error (MAE) are used to evaluate the performance of GMM-KNN and KNN imputation methods. The root mean square error (RMSE) calculates the average square errors of the treated datasets and the error is measured by Eq. (1).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - T_i)^2}{N}} \tag{1}$$

The mean absolute error (MAE) provides the average error in the treated datasets. The error is calculated based on Eq. (2).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|O_i - T_i| \tag{2}$$

N = total number of observations
O = actual values of observation
T = imputed values.

4 Imputation Methods

4.1 K-Nearest Neighbor Imputation

In context of limited data availability such as time series of rainfall and runoff at Sungai Bedup Basin, this study uses the nearest neighbor stations to estimate the missing values of the basin's datasets. Among the nearest neighbor imputation algorithms, K-Nearest Neighbor (KNN) imputation is one of the easiest and efficient methods used to fill in the missing values in the datasets [19].

In this work, the built-in KNN imputation from Matlab is adopted in this study. KNN imputes missing data using nearest-neighbor method. KNN replaces the missing values with a weighted mean of the k nearest-neighbor columns. The weights are inversely proportional to the distances from the neighboring columns in terms of

Euclidean distance. In this study, k = 5 is found to provide the best imputation accuracy in the dataset (not shown here).

4.2 Gaussian Mixture Model Based KNN Imputation (GMM-KNN)

Another efficient nearest neighbor imputation is Gaussian mixture model based KNN imputation (GMM-KNN). The GMM-KNN is proposed by Ding and Ross [15] in their study on handling missing scores in biometric fusion. Two main steps are essential in GMM imputation, that are the density estimation using the GMM assumption and the imputation itself based on this estimated density. For the density estimation, a simulated dataset (s), D_{sim} is generated from Gaussian mixture distribution. The dataset, D_{sim} is simulated from a multivariate normal distribution and then fit a GMM to the data using Matlab. If the D_{sim} = 10, the density estimation will return a GMM with a number of estimated parameters such as ten distinct means, covariances matrices and component proportions to the data. Then KNN imputation process can be used based on the generated Gaussian mixture distribution. The key idea of GMM-KNN is to find the most similar vectors as "donors" in the training set. The Euclidean distance measurement d is employed to find the "nearest" donors for the incomplete score vectors. The GMM-KNN scheme can be summarized in the following steps, as shown in Algorithm 1 [15]. According to our analysis (not shown here), generally, k = 5 and D_{sim} = 1 provide the best imputation accuracy on the dataset in terms of RMSE, MAE and computational time.

Algorithm 1 Gaussian Mixture Model KNN imputation (GMM-KNN)

1. Use the estimated parameters of GMM, to simulate a dataset D_{sim}, having a similar or larger size than D_{ori}
2. For each observation x, apply the distance function d to find k = 5 nearest neighbours in the simulated set D_{sim}
3. The missing variables x, are imputed by the average of corresponding variables from the nearest neighbours taken from D_{sim}

where D_{sim} = simulated dataset (s) that generated from Gaussian mixture distribution
D_{ori} = original dataset

5 Results and Discussion

This study uses the GMM-KNN and KNN methods to impute the missing values in a rainfall and runoff database from East Malaysia. The root mean square error (RMSE) and mean absolute error (MAE) are used to evaluate the performance of GMM-KNN and KNN imputation methods.

Figures 2 and 3 illustrates the imputation performances of GMM-KNN and KNN models at different percentages of missing data.

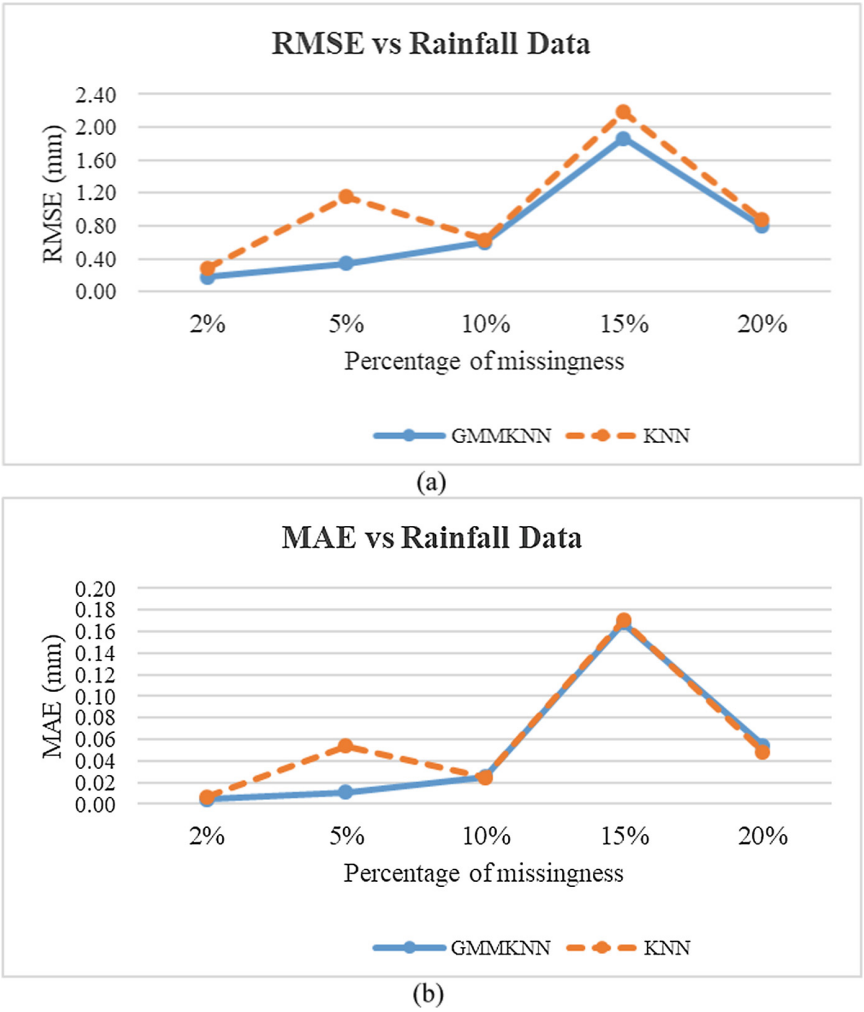
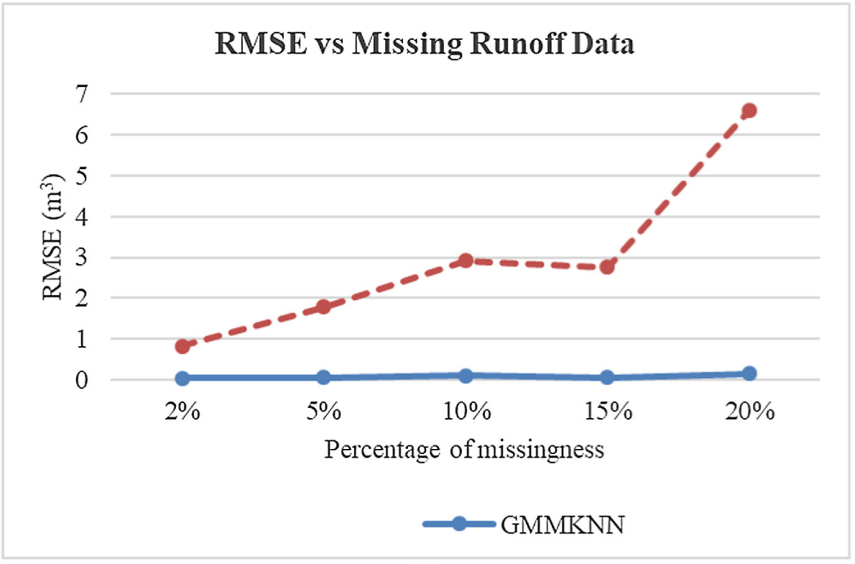


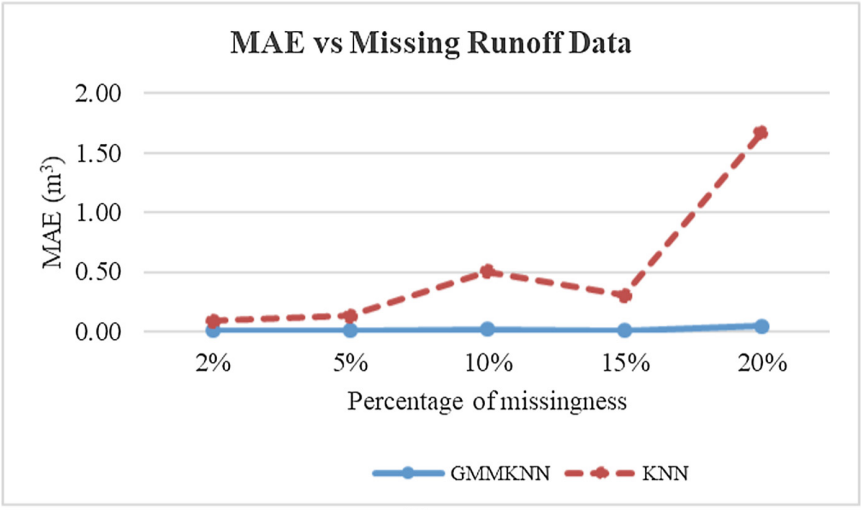
Fig. 2. Comparison of RMSE and MAE plots using GMM-KNN and KNN imputation at various percentage of missing rainfall data

Generally, the GMM-KNN models perform better than KNN model at different missing rainfall and runoff entries, which support the findings by Ding and Ross [15]. As seen in Fig. 2(a), GMM-KNN imputation provides quite accurate predictions at the missing entry of 2%, 5%, 10%, 15% and 20% with the low root mean square error (RMSE), ranges between 0 to 1.8 mm. In addition to this, the observed data points are close to the GMM-KNN model's predicted values at the missing entries between 2% to 10% of missing values in the datasets. However, there are slightly increase in the rainfall value of RMSE for KNN imputation method compared to GMM-KNN.

In Fig. 2(b), the comparison of mean absolute error (MAE) at various percentage of missing data demonstrated that KNN imputation generates higher MAE in rainfall



(a)



(b)

Fig. 3. Comparison of RMSE and MAE plots using GMM-KNN and KNN imputation at various percentage of missing runoff data

values at the missing entries of 2%, 5%, 10%, 15% and 20% as compared to GMM-KNN imputation. In contrast, GMM-KNN offers the best performances at multiple missing entries, with the MAE values range between 0.004 mm to 0.167 mm. Since the GMM-KNN imputation uses a simulated dataset D_{sim} that is synthetically generated for the donor imputation pool, GMM-KNN has a better chance of finding the closest from

the nearest neighbor. However, it is revealed that the values of RMSE and MAE of both methods are not linearly increased with the increasing of the amount of missingness. This could be due to the reason that the methods are not sensitive to the proportion of the missing values in the datasets, as reported by Suhaila et al. in the estimation of missing rainfall data [8].

In addition to that, a study has been conducted by Kamaruzaman *et al.* [10] using weighting methods for treating missing daily rainfall data at the missing percentage of 10% within the context of Peninsular Malaysia. Kamaruzaman *et al.* [10] reported that ACID method obtained good results in the test statistic for treating missing daily rainfall data in terms of RMSE and MAE with an average of 12.78 and 6.28 respectively. As compared to the findings by Kamaruzaman *et al.* [10], the performance of GMM-KNN imputation has improved in terms of RMSE and MAE, where the error values decreased by an average of 11 mm and 6 mm respectively. Furthermore, Ismail *et al.* [7] revealed that IDW method is to be the best missing data estimation method among the interpolation techniques for most of the rainfall stations located at Terengganu, Malaysia. For example, Ismail *et al.* [7] reported the values of RMSE and MAE for station TR_b range between 12.765 mm to 14.76 mm and between 6.042 mm to 6.704 mm respectively. Comparison of this study to the study by Ismail *et al.* [7] shows that GMM-KNN method has lower values of RMSE and MAE, where the error values decreased by an average of 6.97 mm and 3.78 mm respectively at multiple missing rainfall entries. Hence, it is clearly shown that GMM-KNN imputation is the best method for finding the missing rainfall entries at Bedup Basin station.

Meanwhile the example of plots of RMSE and MAE for the runoff missing data are shown in Fig. 3. In Fig. 3(a) and (b), it is observed that the GMM-KNN imputation results in a low value of RMSE, with the range of 0.04 to 0.15 m^3 and MAE, with the range of 0.004 to 0.04 m^3 at all the missing entries. One possible reason is that a good density GMM model positively increases the accuracy of the neighbor-based imputation method.

However, KNN imputation has an increased runoff value of RMSE and MAE when the percentages of missing entries increase. Besides that, KNN imputation has high values of RMSE and MAE, between the range of 0.8 to 6.6 m^3 and the range of 0.08 to 1.7 m^3 respectively. This may be due to the low correlation between the runoff data and the rainfall data of the target stations. Since KNN imputation is based on nearest neighbor method, the accuracy of KNN imputation could highly affected by the correlation of the nearby target stations. As a result, the accuracy of the KNN decreased gradually with the decreased value of correlation between the runoff and the rainfall data. A closer inspection revealed that the decreased value of the correlation between the data could be due to the proportions of missing data. The proportions of missing values that contain more relevance to the target station could lead to overestimate and underestimate the missingness. Therefore, GMM-KNN imputation provides a much better performance for filling rainfall and runoff missing entries than the KNN imputation at Bedup Basin station.

6 Conclusion

In this study, the daily rainfall and runoff data at six gauging stations located in Bedup Basin was considered. The GMM-KNN and KNN methods were applied to fill the missing entries at different percentage of rainfall and runoff missing entries. The results demonstrated that GMM-KNN method performs better than KNN method and it is suitable to be applied for finding the missing rainfall and runoff database.

However, the drawback of GMM-KNN imputation is GMM may fail to work if the dimensionality of the problem is too high. For future work, it is recommended to consider hybrid GMM with other missing data estimation techniques on real world missing datasets.

Acknowledgments. The authors sincerely acknowledge the Department of Irrigation and Drainage (DID), Sarawak, Malaysia for providing the rainfall and runoff data in this study. The authors wish to thank Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, Malaysia Research University Network (MRUN) Vot 4L876 and the Fundamental Research Grant Scheme (FRGS) Vot 5F073 supported under Ministry of Education Malaysia for the completion of the research. The works were also supported by the SPEV project, University of Hradec Kralove, FIM, Czech Republic (ID: 2102–2019). We are also grateful for the support of Ph.D. student Sebastien Mambou in consultations regarding application aspects.

References

1. Selase, A.E., Agyimpomaa, D.E., Selasi, D.D., Hakii, D.M.: Precipitation and rainfall types with their characteristic features. J. Nat. Sci. Res. **5**(20), 1–3 (2015). www.iiste.org
2. Sattari, M.T., Rezazadeh-Joudi, A., Kusiak, A.: Assessment of different methods for estimation of missing data in precipitation studies. Hydrol. Res. **48**(4), 1032–1044 (2017)
3. Kuok, K.K., Harun, S., Shamsudin, S.M.: Global optimization methods for calibration and optimization of the hydrologic Tank model's parameters. Can. J. Civ. Eng. **1**(1), 2–14 (2010)
4. Kuok, K.K., Kueh, S.M., Chiu, P.C.: Bat optimisation neural networks for rainfall forecasting: case study for Kuching city. J. Water Clim. Change (2018)
5. Valizadeh, N., El-Shafie, A., Mirzaei, M., Galavi, H., Mukhlisin, M., Jaafar, O.: Accuracy enhancement for forecasting water levels of reservoirs and river streams using a multiple-input-pattern fuzzification approach. Sci. World J. **2014** (2014)
6. Yaseen, Z.M., El-Shafie, A., Afan, H.A., Hameed, M., Mohtar, W.H., Hussain, A.: RBFNN versus FFNN for daily river flow forecasting at Johor River, Malaysia. Neural Comput. Appl. **27**(6), 1533–1542 (2016)
7. Ismail, W.N., Zin, W.Z., Ibrahim, W.: Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods. Malay. J. Fundam. Appl. Sci. **13**(3), 213–217 (2017)
8. Suhaila, J., Sayang, M.D., Jemain, A.A.: Revised spatial weighting methods for estimation of missing rainfall data. Asia-Pac. J. Atmos. Sci. **44**(2), 93–104 (2008)
9. Eskelson, B.N., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L., Hudak, A.T.: The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scand. J. For. Res. **24**(3), 235–246 (2009)

10. Kamaruzaman, I.F., Zin, W.Z., Ariff, N.M.: A comparison of method for treating missing daily rainfall data in Peninsular Malaysia. Malay. J. Fundam. Appl. Sci. **13**(4–1), 375–380 (2017)
11. Ferrari, G.T., Ozaki, V.: Missing data imputation of climate datasets: implications to modeling extreme drought events. Revista Brasileira de Meteorologia **29**(1), 21–28 (2014)
12. Teegavarapu, R.S., Chandramouli, V.: Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J. Hydrol. **312** (1–4), 191–206 (2005)
13. Dastorani, M.T., Moghadamnia, A., Piri, J., Rico-Ramirez, M.A.: Application of ANN and ANFIS models for reconstructing missing flow data. Environ. Monit. Assess. **166**, 421–434 (2010)
14. Mispan, M.R., Rahman, N.F., Ali, M.F., Khalid, K., Bakar, M.H., Haron, S.: Missing river discharge data imputation approach using artificial neural network. J. Eng. Appl. Sci. **10**(22) (2015)
15. Ding, Y., Ross, A.: A comparison of imputation methods for handling missing scores in biometric fusion. Pattern Recogn. **45**(3), 919–933 (2012)
16. Kuok, K.K., Harun, S., Shamsuddin, S.M., Chiu, P.C.: Evaluation of daily rainfall-runoff model using multilayer perceptron and particle swarm optimization feed forward neural networks. J. Environ. Hydrol. **18**(10), 1–6 (2010)
17. Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I., Ishii, S.: A Bayesian missing value estimation method for gene expression profile data. Bioinformatics **19**(16), 2088–2096 (2003)
18. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, Hoboken (2014)
19. Zainuri, N.A., Jemain, A.A., Muda, N.: A comparison of various imputation methods for missing values in air quality data. Sains Malaysiana **44**(3), 449–456 (2015)