Revised Normal Ratio Methods for Imputation of Missing Rainfall Data

Siti Nur Zahrah Amin Burhanuddin[1], Sayang Mohd Deni[2] and Norazan Mohamed Ramli[3]

[1 2 3] *Center for Statistics and Decision Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA*
40450 Shah Alam, Selangor, Malaysia
**E-mail: [1] misssnzbab@gmail.com, [2]sayang@tmsk.uitm.edu.my, [3]norazan@tmsk.uitm.edu.my*

ABSTRACT

A good quality of rainfall data is highly necessary in hydrological and meteorological analyses. Lack of quality in rainfall data will influence the process of analyses and subsequently, produce misleading results. Thus, this study is aimed to propose modified missing rainfall data treatment methods that produced more accurate estimation results. In this study, the old normal ratio method and the modified normal ratio based on trimmed mean are combined with geographical coordinate method. The performances of these modified methods were tested on various levels of the missing data of 36 years complete daily rainfall records from eighteen meteorology stations in Peninsular Malaysia. The results indicated that both modified methods improved the estimation of missing rainfall values at the target station based on the least error measurements. Modified normal ratio based on trimmed mean with geographical coordinate method is found to be the most appropriate method for station Batu Kurau and Sg. Bernam while modified old normal ratio with geographical coordinate is the most accurate in estimating the missing data at station Genting Klang.

Keywords: *missing data, estimation, normal ratio, geographical coordinate*

INTRODUCTION

Rainfall data is one of the most important climate variables that play a significant role in supplying vital information for environmental planning. A good quality of rainfall data is very important in representing the climatological characteristics truly for an efficient environmental analysis. Moreover, the results accuracy of climatological analyses is extremely dependent on the quality of rainfall data used. Lack of good quality rainfall data will have bad implication on the process of analyses and subsequently, leads to biased results of the analysis.

However, rainfall time series often carries an uncertainty (such as outliers) due to temporal and spatial variability of rainfall measurement, which will affect the quality of data. Furthermore, factors such as faulty instruments, relocation of stations, and human negligence during the measurement of rainfall may affect the continuity of the rainfall records [1]. These problems will affect the estimation output and subsequently, produce inaccurate results. Concerning this situation, this study introduced a practical and reliable approach for developing estimation methods to impute the incomplete (in other words, missing) rainfall data in effort of providing a good quality dataset for public domain.

The effort of filling up the incomplete rainfall records has recently received much attention from many researchers. Various different estimation methods have been applied by the previous studies to fulfill this effort. Nevertheless, the normal ratio (NR) has become the most commonly used method in estimating missing rainfall data as stated in the previous studies [1–7]. Although the NR method is a traditional method, it is still applied by many researchers, such as [3] and [7], due to its simplicity and efficiency in handling the missing data.

The normal ratio method has recorded a long history of treating missing values in rainfall time series. The application of NR method in estimating missing rainfall records was pioneered by [8]. They compared three estimation methods—including the NR method—in imputing the missing data for the United States Weather Bureau database. They found that the performance of the NR method is the most effective for their study.

The NR method was then modified by [5] and [6] through the adaption of correlation coefficients and the effect of distance in the original version of NR method to estimate the missing values in the Malaysian monthly rainfall data. This method is subsequently improved by [1] with some revisions to the methods proposed by [5] and [6] through the combination of these methods with the inverse distance weighting (IDW) method in proposing a more accurate method in estimating the missing daily values in Peninsular Malaysia's rainfall time series. This modified method was revealed to perform very well from its application in the study. The development of the NR method is continually being explored and is recently being studied by [1], [7], and [9]. Similar to [1], [7] also used both versions of the NR method in their study to estimate the missing values in Turkish monthly meteorological data. They compared this method to more sophisticated methods and found that more robust results were produced by the latter.

Geographical coordinate (GC), another simple method that was recently applied in estimating missing rainfall data is also considered in this study. [9] and [10] applied GC method in imputing the missing annual rainfall values in the Iranian time series. They have compared the method and NR with the other advance methods. However, from the priority ranking that they made, it is discovered that the advance methods outperformed these methods.

From all previous applications of the NR method, it can be concluded that it has high capability in estimating missing rainfall records. Therefore, due to its simplicity, NR is considered to be an evergreen method in imputing missing data. However, currently, the limitations of the NR method are disclosed from its applications as found in literature. Thus, it is deeply necessary to revise the NR method in order to enhance its performances.

Consequently, this study is aimed to propose modifications onto the NR method, which will result in more accurate estimations of missing rainfall values. The modification involved the combination of NR with the recently used missing data estimation method, GC method. The methods were compared to existing methods to evaluate its performance in imputing the missing rainfall data based on root mean square error, coefficient of variation root mean square error, and mean absolute error.

MATERIALS AND METHODS

Data Description

The main research area of this study is located in the region of Peninsular Malaysia. Eighteen rainfall measuring stations were selected throughout this region for the purpose of analysis. Table 1 lists the stations with their geographical and statistics information. Three stations were considered as the target station; Batu Kurau, Sg. Bernam, and Genting Klang. This study assigned the neighbouring stations to the target station based on the stations within the radius of 300 km (see Table 2).

Table 1: The geographical coordinates and descriptive information of the selected rainfall stations

Name of Station	Latitude	Longitude	Mean (mm)	Standard Deviation
Batu Kurau	**4.97**	**100.8**	7.42	10.45
Sg. Bernam	**3.68**	**101.33**	6.67	11.23
Genting Klang	**3.23**	**101.75**	6.36	9.97
Sikamat	2.73	101.95	5.3	8.98
Pekan	3.55	103.35	6.38	9.81
Paya Kangsar	3.9	102.43	4.38	7.57
Dispensari Kroh	5.7	101	4.49	7.82
Lawin	5.28	101.05	4.13	7.04
Kemaman	4.22	103.42	6.9	10.06
Kg. Menerong	4.93	103.05	9.38	12.53
Dungun	4.75	103.42	6.54	12.21
Kuala Terengganu	5.32	103.13	6.53	9.68
Ampang Pedu	6.23	100.77	5.01	8.58
Genting Sempah	3.37	101.77	6.44	12.19
Gombak	3.27	101.72	6.52	9.99
Kg. Sg. Tua	3.27	101.68	6.56	10.03
Kota Tinggi	1.75	103.72	5.23	11.2
Johor Bahru	1.47	103.75	6.4	10.2

Table 2: Distances from the target stations to the neighboring stations

Station	Euclidean Distance (km)		
	Batu Kurau	**Sg. Bernam**	**Genting Klang**
Batu Kurau	0.00 (0)	1.39 (154)	1.98 (220)
Sg. Bernam	1.39 (154)	0.00 (0)	0.61 (68)
Genting Klang	1.98 (220)	0.61 (68)	0.00 (0)
Sikamat	2.51 (279)	1.13 (126)	0.54 (60)
Pekan	-	2.02 (224)	1.63 (181)
Paya Kangsar	1.95 (217)	1.12 (124)	0.96 (106)
Dispensari Kroh	0.76 (85)	2.04 (227)	2.58 (287)
Lawin	0.40 (45)	-	2.17 (241)
Kemaman	-	2.15 (239)	1.94 (215)
Kg. Menerong	2.25 (249)	2.12 (236)	2.14 (238)
Dungun	2.63 (291)	2.34(260)	2.25 (250)
Kuala Terengganu	2.35 (261)	2.43 (270)	2.50 (278)
Ampang Pedu	1.27 (141)	2.61(290)	-
Genting Sempah	1.87 (208)	0.54 (60)	0.13 (15)
Gombak	1.93(215)	0.57 (63)	0.05 (5)
Kg. Sg. Tua	1.92 (213)	0.55(61)	0.07 (8)
Kota Tinggi	-	-	2.46 (274)
Johor Bahru	-	-	2.67 (297)

Figure 1 shows the locations of these stations on a topographic map of Peninsular Malaysia. The target and the neighbouring stations are marked as circles (●) and squares (■), respectively. The information of these stations was obtained from the Malaysian Drainage and Irrigation Department (DID). The complete daily rainfall records of a period of 36 years; from January 1, 1975 to December 31, 2010 were used.

Methodology

Existing methods

Normal ratio method: Normal ratio method is based on the mean

ratio of data between the target station and the neighboring stations. [8] proposed old normal ratio (ONR) in estimating missing rainfall records. ONR considered the application of arithmetic mean in the weighting factor. The method is given as follows:

$$Y_t = \frac{1}{N}\sum_{\substack{i=1 \\ i\neq t}}^{N}\left(\frac{\mu_t}{\mu_i}\right)Y_i \qquad (1)$$

where μ_t and μ_i are the sample mean of the available data at target station t and i^{th} neighboring station respectively; Y_t is the missing data at the target station t; Y_i is the concurrently observed data at the i^{th} neighboring station; and N is the number of surrounding stations

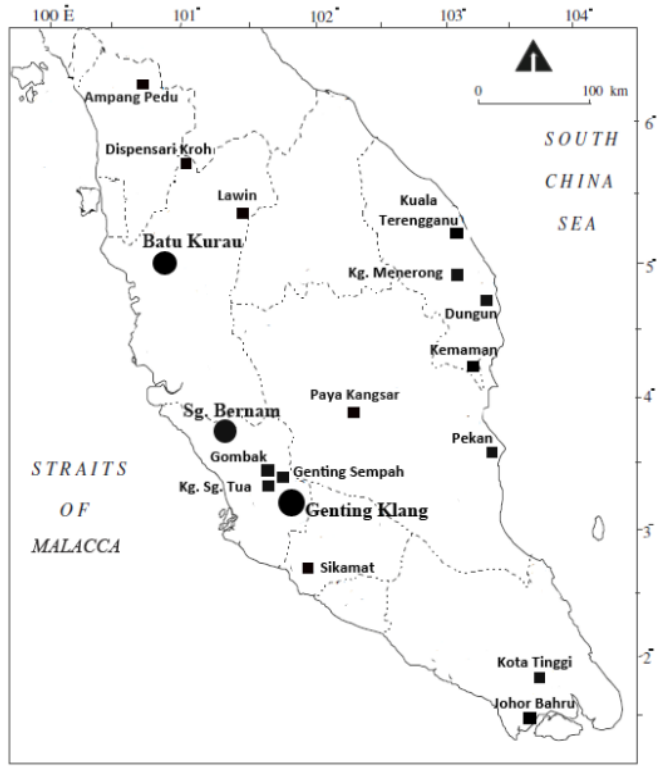


Figure 1: The locations of the target stations and their corresponding neighboring

Geographical coordinate method: Geographical coordinate (GC) method is one of the weighting methods used in imputing the missing rainfall values. This method applied inverse of geographical coordination (the longitude and the latitude) to determine weight coefficient. In GC method, target station is considered as the center point of coordinate system. The distance of each surrounding station is computed according to the center point. The method is as follows:

$$Y_t = \sum_{i=1}^{N} \left(\frac{\frac{1}{x_i^2 + y_i^2}}{\sum_{i=1}^{N} \frac{1}{x_i^2 + y_i^2}} \right) Y_i \qquad (2)$$

where x_i and y_i are the longitude and latitude i^{th} neighboring station; Y_t is the missing data at the target station t; Y_i is the concurrently observed data at the i^{th} neighboring station; and N is the number of surrounding stations

Modified methods: Some modifications have been made to the existing method. The modifications involved the adaptation of robust element into the weighting factor and the combination of both existing methods in order to improve the accuracy of estimation results.

Modified old normal ratio with geographical coordinate method: Combination of both existing methods is proposed in this study as ONRGC method. ONR is often found to be one of the best estimation methods among the researchers. This study attempts to adapt the location element in the method to upgrade its performance. The modification can be expressed as follows:

$$Y_t = \sum_{i=1}^{N} \left(\frac{\left(\frac{1}{x_i^2 + y_i^2} \right) \left(\frac{\mu_t}{\mu_i} \right)}{\sum_{i=1}^{N} \left(\frac{1}{x_i^2 + y_i^2} \right) \left(\frac{\mu_t}{\mu_i} \right)} \right) Y_i \qquad (3)$$

Modified normal ratio based on trimmed mean with geographical coordinate method: By considering the robust element, the arithmetic mean in ONR method is replaced by trimmed mean to reduce the effect of extreme

values exist in rainfall data. The complete rainfall dataset are trimmed 20% before the mean is calculated (20% was selected because it provided the most accurate results among 1, 5, and 10 trimming percentage). The modified normal ratio based on trimmed mean (NRTR) is then combined with the GC method (named as NRTRGC method) and this effort is expected could possibly increase the accuracy of the method. The method is described as follows:

$$\sum \left(\frac{\left(\frac{}{x_i^2 \quad y_i^2} \right) \left(\frac{trim}{trim} \right)}{\sum \left(\frac{}{x^2 \quad y^2} \right) \left(\frac{trim}{trim} \right)} \right) \tag{4}$$

where μ_{trim_t} and μ_{trim_i} are the sample trimmed mean of the available data at target station t and i^{th} neighbouring station respectively.

APPLICATION OF ESTIMATION METHODS

For this purpose of identifying the most appropriate estimation methods to imputing the missing rainfall values, the complete rainfall time series of the target stations were artificially created by randomly removing parts of the data. Six different level of missingness, i.e. 5%, 10%, 15%, 20%, 25%, and 30%, were considered in this study in order to assess the consistency of the estimation results. The estimation methods were then applied to estimate the missing values based on the information obtained from their neighboring stations. The error of the estimated and the observed values were calculated based on root mean square error (RMSE), coefficient of variation root mean square error (CVRMSE), and mean absolute error (MAE) to identify the most appropriate estimation method in imputing the missing data.

RESULTS

The proposed methods, i.e. ONRGC and NRTRGC were compared to the existing methods, i.e. ONR and GC in imputing the missing daily rainfall data at three target stations (Batu Kurau, Sg. Bernam, and Genting Klang) based on three performance criteria. The results are ranked due to their

priority and presented in Table 3 for RMSE, Table 4 for CVRMSE and Table 5 for MAE. Both modified methods showed in the two highest ranking (with the least value of error measurements; RMSE, CVRMSE, and MAE) for all three target stations and most of the level of missingness. In particular, NRTRGC provided the most accurate estimation results for Batu Kurau and Sg. Bernam stations with the RMSE, CVRMSE, and MAE ranged from 8.0857 to 11.7122, 1.1030 to 1.3406, and 5.3834 to 7.1556, respectively. It is followed by ONRGC for the same stations with the RMSE, CVRMSE, and MAE ranged from 8.0815 to 11.7122, 1.0641 to 1.3343, and 5.3847 to 7.1615, respectively. Meanwhile ONR appeared as the worst method for station Batu Kurau and Sg. Bernam and GC is the worst for station Genting Klang.

The comparison of modified and existing methods with different levels of missing data are depicted in Figure 2 based on the RMSE, CVRMSE, and MAE. Station Batu Kurau is selected as the example. The plots show that the methods were not too sensitive to the varying levels of the missing data. However, it could be seen that the results accuracy of the estimation methods are slightly decreased with the increase of the missingness level. It also shows that the modified methods, ONRGC and NRTRGC, were found to be the most accurate for all missing percentages compared to the existing ones, based on the least error measurements. This can be concluded that the proposed methods produced the most accurate estimation of missing rainfall values as expected.

Table 3: Sorting of methods according to their priority rank at various level of missingness based on RMSE

Station	Level of Missingness	RMSE				Priority Ranking			
		ONR	*GC*	*ONRGC*	*NRTRGC*	*ONR*	*GC*	*ONRGC*	*NRTRGC*
Batu Kurau	5%	9.9271	9.8128	9.7588	9.7565	4	3	2	1
	10%	10.3762	10.2229	10.1625	10.1606	4	3	2	1
	15%	10.4308	10.3078	10.2716	10.2705	4	3	2	1
	20%	10.5800	10.4468	10.4137	10.4126	4	3	2	1
	25%	10.6709	10.5901	10.5606	10.5598	4	3	2	1
	30%	10.5719	10.4871	10.4560	10.4553	4	3	2	1
Sg. Bernam	5%	10.5257	10.5020	10.5126	10.5080	4	3	2	1
	10%	10.6419	10.5986	10.5422	10.5416	4	3	2	1
	15%	11.3179	11.2812	11.2313	11.2302	4	3	2	1
	20%	11.6746	11.6707	11.6267	11.6258	4	3	2	1
	25%	11.7063	11.7179	11.6615	11.6599	3	4	2	1
	30%	11.7514	11.7689	11.7137	11.7122	3	4	2	1
Genting Klang	5%	8.0976	8.1232	8.0815	8.0857	3	4	1	2
	10%	8.1873	8.1970	8.1694	8.1751	3	4	1	2
	15%	8.7438	8.7531	8.7272	8.7337	3	4	1	2
	20%	8.6544	8.6503	8.6409	8.6477	4	3	1	2
	25%	8.8595	8.8453	8.8454	8.8536	4	2	1	3
	30%	8.8182	8.8043	8.8036	8.8113	4	2	1	3

Table 4: Sorting of methods according to their priority rank at various level of missingness based on CVRMSE

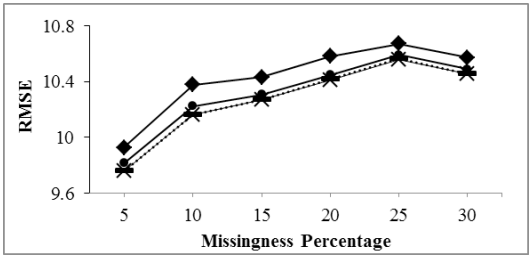
Station	Level of Missingness	CVRMSE				Priority Ranking			
		ONR	*GC*	*ONRGC*	*NRTRGC*	*ONR*	*GC*	*ONRGC*	*NRTRGC*
Batu Kurau	5%	1.1595	1.1462	1.1399	1.1396	4	3	2	1
	10%	1.1836	1.1662	1.1593	1.1591	4	3	2	1
	15%	1.1603	1.1466	1.1426	1.1425	4	3	2	1
	20%	1.1839	1.1690	1.1653	1.1652	4	3	2	1
	25%	1.1699	1.1610	1.1578	1.1577	4	3	2	1
	30%	1.1733	1.1639	1.1604	1.1603	4	3	2	1
Sg. Bernam	5%	1.2551	1.2522	1.2535	1.2529	4	3	2	1
	10%	1.3097	1.3044	1.2975	1.2974	4	3	2	1
	15%	1.3511	1.3467	1.3407	1.3406	4	3	2	1
	20%	1.3341	1.3336	1.3286	1.3285	4	3	2	1
	25%	1.3394	1.3407	1.3343	1.3341	3	4	2	1
	30%	1.3383	1.3403	1.3340	1.3338	3	4	2	1
Genting Klang	5%	1.1047	1.1082	1.1025	1.1030	3	4	1	2
	10%	1.0657	1.0670	1.0634	1.0641	3	4	1	2
	15%	1.1044	1.1055	1.1023	1.1031	3	4	1	2
	20%	1.0853	1.0848	1.0836	1.0844	4	3	1	2
	25%	1.1060	1.1042	1.1042	1.1053	4	2	1	3
	30%	1.1074	1.1057	1.1056	1.1065	4	2	1	3

Table 5: Sorting of methods according to their priority rank at various level of missingness based on MAE

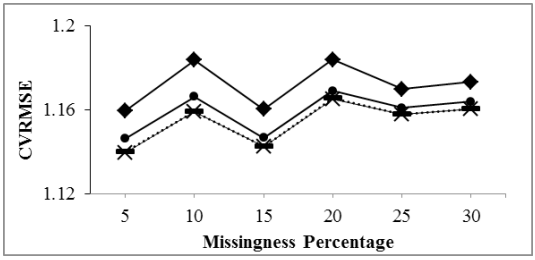
Station	Level of Missingness	MAE				Priority Ranking			
		ONR	*GC*	*ONRGC*	*NRTRGC*	*ONR*	*GC*	*ONRGC*	*NRTRGC*
Batu Kurau	5%	6.9657	6.5708	6.4263	6.4209	4	3	2	1
	10%	7.2072	6.7389	6.5906	6.5851	4	3	2	1
	15%	7.2633	6.8158	6.6881	6.6834	4	3	2	1
	20%	7.2450	6.7676	6.6525	6.6484	4	3	2	1
	25%	7.2588	6.8348	6.7215	6.7176	4	3	2	1
	30%	7.2402	6.8024	6.6875	6.6835	4	3	2	1
Sg. Bernam	5%	6.7577	6.6517	6.5655	6.5578	4	3	2	1
	10%	7.0250	6.8242	6.6564	6.6509	4	3	2	1
	15%	7.3621	7.1627	6.9994	6.9935	4	3	2	1
	20%	7.4775	7.3123	7.1615	7.1556	4	3	2	1
	25%	7.4084	7.2519	7.0942	7.0877	4	3	2	1
	30%	7.4123	7.2716	7.1154	7.1091	4	3	2	1
Genting Klang	5%	5.4585	5.4824	5.3847	5.3834	3	4	2	1
	10%	5.4680	5.4824	5.3929	5.3923	3	4	2	1
	15%	5.6928	5.7043	5.6182	5.6184	3	4	1	2
	20%	5.7053	5.7048	5.6301	5.6307	4	3	1	2
	25%	5.7370	5.7398	5.6714	5.6731	3	4	1	2
	30%	5.7471	5.7405	5.6749	5.6766	4	3	1	2

CONCLUSION

ONRGC and NRTRGC were proposed in this study for the application of estimating missing daily rainfall data. These methods were evaluated by comparing them to existing methods. Eighteen rainfall stations in Peninsular Malaysia were used for the analysis. The performances of the methods were evaluated based on the comparison of error of the estimated and the observed values. Based on the least error measurements, the proposed ONRGC and NRTRGC methods were found to be the most appropriate methods in estimating missing daily rainfall data for the stations considered in this study compared to the existing ones. This shows that the adaption of the location element improved the performance of NR method. Therefore, the modified methods are highly recommended to be applied as an alternative method in any studies related to missing values, in particular. The degree of suitability of the proposed methods to other climatic variables (e.g. temperature) and time scales (e.g. monthly) needs to be determined and validated in future studies.



(a)



(b)

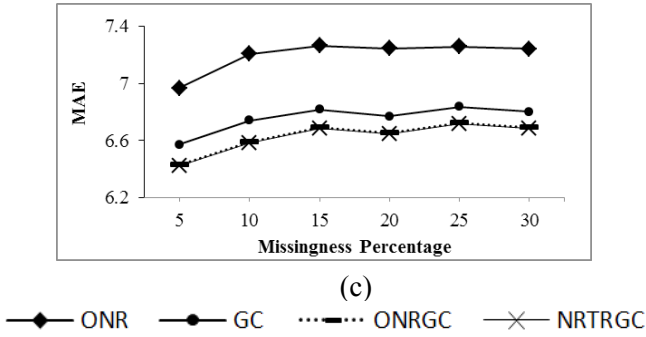


Figure 2: Performance of estimation methods at various level of missingness based on RMSE (a), CVRMSE (b), and MAE (c)

ACKNOWLEDGEMENT

The authors are indebted and thankful to the staff of Malaysian Meteorological Department, and Drainage and Irrigation Department for providing the hourly rainfall data for the usage of this study. This research would not have been possible without the sponsorships of Ministry of Higher Education and also Universiti Teknologi MARA, Malaysia. This research was funded by the Malaysian Fundamental Research Grant Scheme, FRGS/1/2014/ST06/UITM/02/6.

REFERENCES

[1] J. Suhaila, S. M. Deni, and A. A. Jemain, 2008. Revised spatial weighting methods for estimation of missing rainfall data, *Asia-Pacific J. Atmos. Sci*, *Vol. 44*, pp. 93–104.

[2] R. P. De Silva, N. D. K. Dayawansa, and M. D. Ratnasiri, 2007. A Comparison of methods used in estimating missing rainfall data, *J. Agric. Sci.*, *Vol. 3*, pp. 101–108.

[3] M. H. Kashani and Y. Dinpashoh, 2012. Evaluation of efficiency of different estimation methods for missing climatological data, *Stoch. Environ. Res. Risk Assess.*, *Vol. 26*, pp. 59–71.

[4] A. Mair and A. Fares, 2010. Comparison of rainfall interpolation methods in a mountainous region of a tropical island, *J. Hydrol. Eng.*, *Vol. 15*, pp. 61-66.

[5] W. Y. Tang, A. H. M. Kassim, and S. H. Abubakar, 1996. Comparative studies of various missing data treatment methods-Malaysian experience, *Atmos. Res.*, *Vol. 42*, pp. 247–262.

[6] K. C. Young, 1992. A three-way model for interpolating for monthly precipitation values, *Mon. Weather Rev.*, *Vol. 120*, pp. 2561–2569.

[7] C. Yozgatligil, S. Aslan, C. Iyigun, and I. Batmaz, 2013. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data, *Theor. Appl. Climatol.*, *Vol. 112*, pp. 143–167.

[8] J. L. H. Paulhus and M. A. Kohler, 1952. Interpolation of missing precipitation records, *Mon. Wea. Rev*. *Vol. 80*, pp. 129–133.

[9] G. Khosravi, A. R. Nafarzadegan, A. Nohegar, H. Fathizadeh, and A. Malekian, 2014. A modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran, *Theor. Appl. Climatol.*, *119*, pp. 1–10.

[10] Z. Khorsandi, M. Mahdavi, A. Salajeghe, and S. Eslamian, 2011. Neural network application for monthly precipitation data reconstruction, *J. Environ. Hydrol.*, *Vol. 19*, pp. 1–12.