ENGINEER - Vol. XLIX, No. 04, pp. [1-8], 2016

A Comparison of Methods of Estimating Missing Daily Rainfall Data

H.P.G.M. Caldera, V.R.P.C. Piyathisse and K.D.W. Nandalal

Abstract: The availability of a long and complete rainfall record is very important for carrying out a hydrological study successfully. However in general, the data series in these records may contain gaps for various reasons. The objective of this study is to analyse the different methods available for filling gaps in rainfall data records and propose a method suitable for a river basin situated in a mountainous area in Sri Lanka. Towards this end, daily rainfall data from ten gauging stations in the upper catchment area of BaduluOya were collected. Seven different techniques were studied to ascertain their suitability. The methods studied were the Arithmetic Mean method, Normal Ratio method, Inverse Distance Weighting method, Linear Regression method, Weighted Linear Regression method, Multiple Linear Regression method and the Probabilistic method. The data generated for the target stations were compared with actual observations made, based on error statistics, Error Standard Deviation (STD),Root Mean Square Error (RMSE) and Correlation Coefficient (CC). The results of the study showed that for target stations that have only one neighbouring station with a high correlation coefficient, the Probabilistic method and the Linear Regression method give good predictions. For stations that have relatively low correlation coefficients with the neighbouring stations, the Inverse Distance Squared method and the Normal Ratio method outperformed the others. To obtain accurate results from the Multiple Linear Regression method and the Weighted Linear Regression method, it is necessary to have a set of neighbouring stations that have fairly high correlation coefficients with the target station.

Keywords: Infilling methods, Rainfall time series, Error statistics, Probabilistic method

1. Introduction

A lengthy rainfall data series plays a major role in all water related studies. Consistency and continuity of rainfall data series are very important for obtaining reliable results from such studies. However, these rainfall data series very often contain gaps or missing values due to various reasons such as the absence of observers, problems with measuring devices, loss of records etc. The use of a rainfall data series with missing values may critically influence the statistical power and accuracy of a study. By estimating and filling the missing rainfall data, a series could be made longer to make the water related study more reliable. Diverse techniques have been proposed and adopted in filling missing data with a view to obtaining a continuous and lengthy rainfall data series.

Basically, the procedures can be grouped into three major classes as deterministic, stochastic and artificial intelligence based methods [1]. Deterministic approaches are more suitable because of their robustness, ease of implementation and computational efficiency [1, 2]. They are mathematical models that always produce the same output from a given initial condition and they neither contemplate on the existence of randomness nor do they attribute the results to a probability of occurrence. Arithmetic mean method, normal ratio method and inverse distance weighting method are the examples of deterministic methods. Stochastic and artificial intelligence approaches are sophisticated but they are more costly and complex [3, 4]. Stochastic methods provide probabilistic estimates of the outcome. Artificial intelligence methods such as artificial neural networks (ANNs)have a complex mathematical formulation and are thus difficult to implement.

The best method for estimating missing rainfall data can vary for different areas depending on their rainfall patterns and spatial distributions.[2] This study filled data at monthly time steps.

Eng.H.P.G.M. Caldera, *AMIE****(Sri Lanka),*** *B.Sc.(Eng)Hons., Lanka Hydraulics Institute, Katubedda, Moratuwa. Email:gangacaldera@gmail.com*

V.R.P.C. Piyathisse*, B.Sc.(Eng)Hons., Department of Civil Engineering, University of Peradeniya. Email:pavithranipiyatisse@gmail.com*

Eng.(Prof) K.D.W. Nandalal*, IntPE(SL), C.Eng, FIE(Sri Lanka), B.Sc.(Eng)Hons., MEng(AIT), Ph.D.(The Netherlands), Senior Professor, Department of Civil Engineering, University of Peradeniya. Email:kdwn@pdn.ac.lk*

The objective of this paper is to present the analysis carried out to evaluate the few methods that are available for filling gaps in rainfall data records and propose a novel method in their place. A river basin situated in a mountainous area in Sri Lanka was used for the study. The study area is shown in Figure 1.

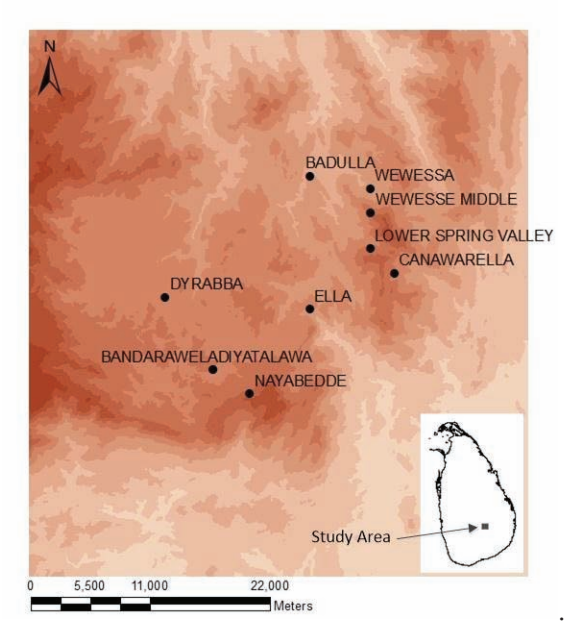

Figure 1 - Distribution of ten rain gauging stations in BaduluOya upper catchment

2. Methodology

2.1 Data Collection and Study Area

Daily rainfall data at ten gauging stations in BaduluOya upper catchment collected over a period of 10 years were considered for the study based on their availability and spatial variability. The elevation of the catchment selected, varied between 740m and 1440 m above mean sea level and it covered an area of about 600 km^2. Figure 1 shows the gauging stations and their data availability is shown in Figure 2. The data were obtained from the Department of Meteorology.

Gauging Station	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Badulla										
Bandarawela										
Canawarella										
Diyathalava										
Dyrabba										
Ella										
Lower spring valley										
Nayabedda										
Wewessa										
Wewessa middle										

Figure 2 - Rainfall data availability at the gauging stations

2.2 Method

Seven different techniques for estimating the missing data were used to evaluate the suitability of each method for mountainous areas.The seven methods used included six deterministic methods and one probabilistic method. Gauging stations located at Bandarawela, Badulla and Lower Spring Valley, each of which had 100% data availability during a period of 10 years were selected for the evaluation.

As shown in Table 1, during the analysis, rainfall data of the above mentioned three stations (target stations)relating to one or more months of a every year were randomly deleted.These months were considered as months for which rainfall data were missing. Thereafter in respect of each station, the missing (deleted) data of each month were estimated using rainfall data available at other neighboring gauging stations. This was repeated for all the seven methods. Subsequently, the estimated data were compared with the actual observations making use of three error statistics.

The stations which were nearest and spread out well around the target station were the ones that were selected for each target station. Thereafter, the correlation coefficients of each target station with the above selected neighbouring stations were calculated. As some neighbouring stations had gaps, a time period where all the stations had 100% data were considered in calculating the correlation coefficients. The neighbouring stations were ranked according to their correlation coefficients for rainfall data estimation. The distribution of neighbouring stations for each target station is shown in Figure 3.

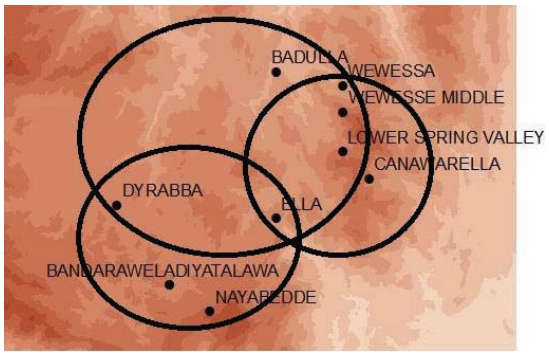

Figure 3 - Neighbouring stations selected for each target station

The names of target stations with the names of their corresponding neighbouring gauging stations along with their respective correlation coefficients are given in Table 2.

Table 1 - Randomly deleted months of the three gauging stations

Year	Deleted months		
	Bandarawela	Lower Spring Valley	Badulla
2004	February	February	March
2005	January	January	June
2006	August	July	September
2007	March	April	December
2008	May	May	January
2009	September	September	February
2010	October	October	April
2011	June	March	November
2012	April	June	July
2013	December	November	August

Table 2 - Correlation coefficients for the neighbouring gauging stations

Target Station	Neighbouring Stations	Distance between gauging stations/(km)	Correlation Coefficient
Bandarawela	Nayabedda	3.317	0.803
	Diyathalawa	1.563	0.891
	Ella	8.912	0.614
	Dyrabba	7.817	0.515
Lower Spring Valley	Canawarella	3.127	0.784
	Wewessa	5.638	0.763
	Wewessa middle	3.317	0.747
	Ella	7.816	0.639
Badulla	Wewessa	6.721	0.669
	Wewessa middle	6.444	0.509
	Lower Spring Valley	8.634	0.646
	Ella	12.163	0.518
	Dyrabba	17.266	0.562

The seven methods used in the estimation of missing data are given below.

1. Arithmetic Mean (AM) Method

If the normal annual rainfalls at surrounding gauges are within 10% of the normal annual precipitation at the stations concerned, then the arithmetic procedure could be adopted to estimate the missing data [5]. This assumes equal weights from all nearby rain gauge stations and uses the arithmetic mean of the precipitation data.

$$p_x = \frac{1}{m}\sum_{i=1}^{m} p_i \qquad \ldots(1)$$

2. Normal Ratio (NR) Method

This method is used if the normal annual precipitation of any surrounding gauges exceeds 10% of the gauge that is under consideration. This weighs the effect of each surrounding station [6].

$$p_x = \frac{1}{m}\sum_{i=1}^{m} \left(\frac{N_x}{N_i}\right) p_i \qquad \ldots(2)$$

3. Inverse Distance Weighting(IDW) Method

In this method, the weight for each station is assumed to be inversely proportional to its squared distance of the target station from the neighbouring station with data [7].

$$p_x = \frac{\sum_{1=i}^{m} \frac{1}{d_i^2} p_i}{\sum_{1=i}^{m} \frac{1}{d_i^2}} \qquad \ldots(3)$$

4. Linear Regression (LR) Method

The correlation coefficients between the target station and each of the neighbouring stations are initially calculated and then ranked. Then the missing data are estimated using a linear regression equation with the station that has the highest correlation. The correlation and the equation of the regression line are obtained using Microsoft Excel software.

In order to be able to generate zero values together with non-zero values, the regression line is forced through the origin.

$$p_x = c_1 p_i \qquad \ldots(4)$$

5. Weighted Linear Regression (WLR) Method

This method takes into account the impact of the distance between the target station and data station in addition to their correlation with respect to data. It uses Equation 5, in which a weighting factor is introduced[1].

$$p_x = \frac{\sum_{1=i}^{n} \frac{1}{d_i^2}}{\sum_{1=i}^{n} \frac{1}{d_i^2}} c_1 p_i \qquad \ldots(5)$$

6. Multiple Linear Regression (MLR) Method

Rainfall data are estimated considering a linear correlation between the target station and some of the other (multiple) neighboring stations [1].

$$p_x = \sum_{i=1}^{n} c_i p_i \quad \ldots(6)$$

7. Probabilistic (PRB) Method

The PRB method assumes that the missing rainfall values and available data for a particular gauging station have similar statistical properties[3]. The data estimation procedure starts with the calculation of the monthly total rainfall at the missing station using the IDW method and based on the observations made at the neighbouring rain gauging stations located within a 10 km radius. Thereafter using non-zero values in the data series, a probability distribution is fitted into monthly rainfall data over a ten year period and the parameters of the probability distribution are determined. The daily rainfall data are usually very much right skewed and therefore probability distributions that match well with the mare used in the study. Only non-zero values for the missing month are randomly generated using estimated probability distribution parameters until the average of the generated values become equal to the monthly rainfall estimated using the IDW method. Finally, the distribution of generated datachronologically is done for each month by matching them with the data available at the gauging station that has the highest correlation.

The data were generated using three probability distributions, viz., Generalized Gamma, Weibull and Pearson 6 in order to determine the probability distribution that was most suitable for the daily rainfall data.

The notations used for the seven methods are given below.

p_x = Estimate for the target station (X)
p_i= Rainfall values of rain gauges used for estimation
m= Number of surrounding stations
N_x= Normal annual precipitation of the X station
N_i =Normal annual precipitation of the surrounding stations
d_i = Distance from each location to the point being estimated
c_1, c_i= Regression coefficients

2.3 Comparison of Estimates

The data estimated for the target stations were compared with the actual observations made, based on the following error statistics:

a) Error Standard Deviation (STD)
b) Root Mean Square Error (RMSE)
c) Correlation Coefficient (CC)

3. Results and Discussion

The AM method could not be applied to any of the target stations since the average annual rainfalls at their surrounding gauges were not within the 10% range of the normal annual precipitation at the target station. This means that the annual rainfall values can be significantly different among the gauging stations even though they are located close to each other, probably due to the considerable variations in their elevations.

Table 3 presents an example of a comparison of the observed rainfalls with the generated rainfalls at the three gauging stations for three different months. Data generated were for the months given in Table 1. As Table 3 reveals, the results indicate that different methods perform differently for the three stations.

When the correlation coefficients with the neighboring stations are each more than 0.7, the MLR method performed acceptably. As Table 2 shows, the Bandarawela station has only two neighbouring gauging stations each having a correlation coefficient greater than 0.7 whereas the Badulla station has no neighbouring gauging station with a correlation coefficient exceeding 0.7. Therefore, the MLR method could not be used for these two stations. However for the Lower Spring Valley station, each of the neighbouring stations except one station had a correlation coefficient greater than 0.7. Therefore, the MLR method generated fairly good results for the Lower Spring Valley station.

The performance of the MLR method was poorer than the LR method. It requires a set of neighbouring stations with good correlation coefficients (greater than 0.7) for it to perform well. Sometimes there are negative coefficient values due to the presence of zero values on the daily scale. This method may perform well on the monthly scale because monthly records usually consist of non-zero values.

Table 3 – Comparison of generated rainfalls with actual rainfalls at the three stations

		Bandarawela							Lower Spring Valley								Badulla					
		Actual	NR	IDW	LR	WLR	PRB		Actual	NR	IDW	LR	WLR	PRB	MLR		Actual	NR	IDW	LR	WLR	PRB
February 2004	1	0	0.0	0.0	0.0	0.0	0	October 2010	73.2	58.1	60.0	80.2	66.4	53.4	45.4	December 2007	0	0.2	0.2	0.0	0.2	0.0
	2	0	0.0	0.0	0.0	0.0	0		22	49.8	34.9	28.3	23.6	21.0	17.9		2.7	2.3	1.8	2.7	1.7	3.1
	3	1.5	1.0	0.6	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		0	0.4	0.5	0.4	0.6	0.2
	4	0	0.4	1.3	1.6	1.3	0.4		0	0.0	0.0	0.0	0.0	0.0	0.0		2.6	4.0	2.5	0.8	2.3	0.4
	5	0	0.0	0.0	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		3.5	8.6	9.6	7.9	8.5	8.0
	6	0	0.0	0.0	0.0	0.0	0		1.2	0.2	0.2	0.4	0.3	3.0	2.6		1.7	6.4	6.1	2.4	3.1	1.4
	7	0	0.4	0.0	0.0	0.0	0		0.7	0.0	0.0	0.0	0.0	0.0	0.0		2.5	12.2	11.5	3.1	5.8	3.6
	8	4.5	6.0	6.3	6.0	4.9	11.2		0	0.0	0.0	0.0	0.0	0.0	0.0		3	5.0	5.1	4.7	5.8	7.9
	9	0	0.0	0.0	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		12.5	9.1	6.9	13.4	8.3	14.7
	10	0	0.0	0.0	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		0	1.4	1.2	0.0	0.0	0.0
	11	0	0.0	0.0	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		3.3	6.3	5.4	0.8	0.9	0.4
	12	0.7	0.3	0.9	1.0	0.8	0.2		0	0.0	0.0	0.0	0.0	0.0	0.0		19.4	13.3	12.5	9.8	10.1	11.1
	13	3.5	6.4	4.5	3.2	2.7	1.7		0	0.0	0.0	0.0	0.0	0.0	0.0		4.9	10.4	9.8	4.3	4.5	5.7
	14	1.6	3.5	2.7	1.4	1.1	0.3		0	0.0	0.0	0.0	0.0	0.0	0.0		64.9	23.7	22.5	26.7	21.8	26.2
	15	1	1.0	1.3	0.7	0.6	0.1		0	0.0	0.0	0.0	0.0	0.0	0.0		9.8	12.6	11.2	13.4	10.3	16.5
	16	0	0.0	0.0	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		20.4	18.8	16.8	17.3	14.6	20.0
	17	0	0.0	0.0	0.0	0.0	0		12.6	0.0	0.0	0.0	0.0	0.0	0.0		21.5	9.7	11.5	14.9	13.8	17.5
	18	0	0.0	0.0	0.0	0.0	0		0	7.9	4.5	3.7	6.0	7.2	6.1		6.5	19.6	15.6	9.4	11.4	8.5
	19	0	0.0	0.0	0.0	0.0	0		0	0.6	0.9	1.5	1.1	3.5	3.0		8.1	19.1	21.1	18.8	22.2	20.4
	20	0	0.0	0.0	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		6.8	16.9	15.4	9.8	10.4	9.2
	21	0	0.0	0.0	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		63.3	56.7	56.8	55.8	56.4	48.1
	22	19.5	12.9	13.6	12.4	10.2	11.4		0	0.0	0.0	0.0	0.0	0.0	0.0		25.1	21.5	21.7	36.1	22.5	47.4
	23	4.8	5.9	4.5	3.2	2.7	1.9		0	0.0	0.0	0.0	0.0	0.0	0.0		17.8	21.6	24.2	23.6	28.3	23.4
	24	0	0.0	0.0	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		0	2.0	2.3	0.0	2.5	0.0
	25	0	0.0	0.0	0.0	0.0	0		0	0.0	0.0	0.0	0.0	0.0	0.0		0	0.0	0.0	0.0	0.0	0.0
	26	4	5.0	6.5	3.7	3.0	9.3		6.3	13.7	13.2	17.8	16.8	18.9	16.0		0	0.0	0.0	0.0	0.0	0.0
	27	0	1.4	3.4	3.7	3.0	5.6		6.5	10.1	14.7	23.8	18.4	20.8	17.7		0	0.0	0.0	0.0	0.0	0.0
	28	0	0.0	0.0	0.0	0.0	0		0.3	1.2	1.8	3.0	2.3	6.7	5.7		0	0.0	0.0	0.0	0.0	0.0
	29	0	0.0	0.0	0.0	0.0	0		54.3	21.8	27.5	40.9	31.3	24.8	21.1		0	0.0	0.0	0.0	0.0	0.0
									2.9	13.4	11.3	11.9	9.1	14.5	12.3		0	0.0	0.0	0.0	0.0	0.0
									5.7	11.6	8.6	7.4	5.7	11.5	9.8		0	0.0	0.0	0.0	0.0	0.0
Total		41.1	44.1	45.6	37.1	30.4	42.0		185.7	188.4	177.7	218.8	181	185.3	157.5		300.3	301.8	292.1	276.0	265.8	293.7

When generating data using the PRB method, the monthly average rainfalls obtained from the IDW method were used as constraints. Initially, the most representative probability distribution for the data set at the station where data were missing was identified. Out of the probability distributions, Generalized Gamma,Weibull and Pearson 6, Generalized Gamma was found to be the most suitable. Non-zero values for the missing month were randomly generated thereafter. The monthly average rainfall computed based on the IDW method is in itself estimation and thus can be erroneous. Moreover, random values were generated only up to 100 seeds or trials. These facts can reduce the accuracy of the final outcome of the PRB method.

The WLR method showed that when the highest correlated stations are closer to the target station, the generated data are acceptable. For instance, for the Bandarawela station, this method gave significantly better results than for the Badulla station.Finally as shown in Table 4, the generated data were compared with the observed data based on the three above mentioned error statistics.

When data values generated using the above mentioned methods are compared with yje actual data values, there were no significant differences among the STD values in most of the cases. However when the LR method was used, Bandarawela and Badulla stations gave minimum STD values for most of the years as far as the overall performance was concerned. For the Lower Spring Valley station, both LR and NR methods gave the minimum STDs.

The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions. This also shows results similar to the STD. For the Bandarawela and Badulla stations, the LR method showed acceptable results. For these stations, the PRB method gave results similar to the LR method. However for the Lower Spring Valley station, both the IDW method and the NR method gave the lowest RMSE values.

In many cases, the CCs at the Bandarawela station are relatively higher than those at the other two stations because of the good correlation that the Bandarawela station has with its neighbouring stations as shown in Table 2. The Bandarawela station showed high CCs for the LR method indicating the latter's superiority for this particular station compared to the other methods. The NR method showed high CC values for the Lower Spring Valley station. For the Badulla station the PRB, IDW and LR methods all performed in a similar manner.

Table 4 – Error statistics at the three stations

		STD					RMSE					CC				
	Year	NR	IDW	LR	WLR	PRB	NR	IDW	LR	WLR	PRB	NR	IDW	LR	WLR	PRB
Bandarawela	2004	1.498	1.466	1.616	1.615	2.581	1.476	1.448	1.594	1.587	2.536	0.930	0.928	0.933	0.929	0.745
	2005	4.420	5.048	5.131	5.116	5.220	4.477	4.970	5.066	5.034	5.136	0.494	0.685	0.415	0.449	0.397
	2006	3.774	0.834	1.315	0.812	2.864	4.176	0.850	1.304	0.799	2.853	0.854	0.609	0.542	0.853	0.340
	2007	12.424	7.095	3.584	3.997	4.788	12.338	7.022	3.605	3.959	4.717	0.957	0.764	0.999	0.990	0.963
	2008	12.907	10.143	11.168	6.562	6.860	13.015	10.280	11.860	10.462	7.163	0.722	0.397	0.727	0.742	0.884
	2009	6.190	5.458	5.397	5.424	5.482	6.218	5.663	5.642	5.661	5.617	0.153	-0.115	0.239	0.210	0.164
	2010	6.004	1.805	1.551	2.502	4.377	5.907	1.832	1.526	2.470	4.306	0.994	0.953	0.997	0.995	0.978
	2011	0.351	0.339	0.239	0.363	0.527	0.350	0.340	0.250	0.360	0.524	0.988	0.977	0.992	0.991	0.911
	2012	5.468	6.797	6.050	5.320	4.369	5.680	6.987	5.949	5.231	4.725	0.874	0.848	0.839	0.866	0.922
	2013	2.571	6.242	2.879	2.826	4.407	2.779	6.153	2.872	2.851	4.465	0.814	0.972	0.947	0.966	0.888
Lower Spring Valley	2004	2.487	2.406	2.979	2.988	4.517	2.444	2.399	2.936	2.943	4.439	0.882	0.909	0.844	0.852	0.656
	2005	7.726	9.258	11.848	10.013	9.665	7.606	9.301	11.716	9.867	9.556	0.629	0.762	0.539	0.620	0.618
	2006	2.944	3.409	4.078	3.450	1.822	2.908	3.387	4.073	3.446	1.792	0.884	0.887	0.813	0.928	0.954
	2007	12.357	12.330	12.774	12.350	13.010	12.164	12.159	12.634	12.169	12.762	0.649	0.655	0.634	0.653	0.615
	2008	7.661	7.523	8.431	7.923	9.562	7.538	7.505	8.457	8.082	9.465	0.628	0.647	0.515	0.574	0.487
	2009	2.276	2.404	2.175	2.421	1.841	2.256	2.505	2.349	2.624	1.966	0.860	0.841	0.884	0.884	0.901
	2010	9.118	6.923	5.575	5.978	8.292	8.970	6.815	5.587	5.883	8.157	0.911	0.829	0.944	0.935	0.876
	2011	5.188	4.634	4.302	4.355	4.689	5.180	4.646	4.573	4.594	4.786	0.629	0.577	0.631	0.625	0.613
	2012	1.276	1.718	1.320	1.191	1.633	1.271	1.702	1.299	1.174	1.625	0.280	0.562	0.120	0.116	0.229
	2013	11.471	14.118	11.370	13.637	13.259	11.278	15.786	11.657	13.713	13.190	0.827	0.826	0.788	0.689	0.725
Bandarawela	2004	5.874	5.432	5.257	5.476	6.001	5.793	5.365	5.359	5.520	5.916	0.579	0.490	0.636	0.574	0.557
	2005	1.476	1.624	1.213	1.422	1.505	1.466	1.605	1.252	1.407	1.491	0.589	0.656	0.813	0.537	0.728
	2006	8.353	8.371	9.203	8.832	9.355	8.215	8.232	9.107	8.762	9.199	0.563	0.566	0.424	0.504	0.560
	2007	9.393	9.286	8.116	9.093	9.326	9.139	9.138	8.022	9.014	9.177	0.830	0.830	0.873	0.837	0.821
	2008	6.753	6.304	7.162	6.855	8.164	6.725	6.388	7.454	7.066	8.148	0.914	0.893	0.892	0.914	0.860
	2009	2.259	1.917	1.933	1.879	1.882	2.227	1.913	1.926	1.883	1.874	0.204	0.085	0.154	0.200	0.263
	2010	15.715	15.263	16.942	15.732	20.838	15.478	15.145	17.330	16.170	20.533	0.506	0.480	0.410	0.451	0.436
	2011	5.374	8.036	5.289	5.524	12.650	5.469	8.191	5.326	5.440	12.882	0.910	0.908	0.916	0.909	0.876
	2012	3.041	2.842	3.808	3.792	2.509	3.041	2.841	3.874	3.850	2.524	0.801	0.797	0.801	0.791	0.932
	2013	6.925	7.016	6.147	6.234	7.027	6.723	6.908	6.265	6.310	6.922	0.124	0.106	0.171	0.121	0.171

Table 5 presents the most suitable methods for filling missing data at the three stations based on error statistics.

Table 5 - Most suitable methods for filling missing data based on error statistics

Station	Method		
	STD	RMSE	CC
Bandarawela	LR	LR, PRB	LR, PRB
Lower Spring Valley	LR	IDW, NR	NR
Badulla	IDW, NR	IDW, NR	IDW

The determination of the best method for each station needs to be based on the overall performance of all three error statistics. However, it is not always possible to select the most suitable method based only on the smallest values of the STD and RMSE since small rainfalls will result in small STD values. This may lead to an improper decision. Thus, the most suitable methods were determined by comparing correlation coefficients and monthly total rainfalls along with STD and RMSE values. Thus, both PRB and LR methods performed well for the Bandarawela station. The NR method is the preferred method for the Lower Spring Valley station while the IDW and NR methods performed well for the Badulla station.

Data generation was done on a daily basis and thereafter their monthly totals were computed. Table 6 presents recorded and generated monthly rainfalls at the three locations.

As shown in Table 7 for the Bandarawela station, the IDW and PRB methods have given the nearest monthly rainfall values. The NR and WLR methods performed best for the Lower Spring Valley station while the NR and IDW methods were the most suitable methods for the Badulla station.

Table 6 –Comparison of monthly rainfalls

	Year	Actual	NR	IDW	LR	WLR	PRB
Bandarawela	2004	41.1	44.1	45.6	37.1	40.2	42.0
	2005	79.9	112.9	86.1	66.7	76.6	77.3
	2006	21.5	80.8	28.4	16.5	22.5	35.3
	2007	94.7	42.4	70.8	71.5	80.6	102.6
	2008	230.9	142.3	154.2	92.4	136.6	156.4
	2009	66.5	28.3	12.3	8.9	9.5	19.1
	2010	141.4	143.0	155.3	142.4	135.0	140.4
	2011	8.0	6.2	6.0	5.4	5.6	5.8
	2012	216.8	271.7	277.9	217.0	214.4	275.9
	2013	137.5	173.2	147.5	122.7	117.9	170.6
Lower Spring Valley	2004	56.0	56.2	43.9	49.6	48.7	53.2
	2005	208.7	199.2	150.2	245.6	226.7	179.1
	2006	38.0	30.0	23.2	16.2	19.4	38.4
	2007	401.7	383.7	375.7	443.1	377.2	397.8
	2008	153.2	86.9	101.9	114.5	158.0	120.5
	2009	59.3	50.8	33.7	30.2	26.1	36.3
	2010	185.7	188.4	177.7	218.8	181.0	185.3
	2011	134.9	107.5	107.0	81.2	83.5	95.4
	2012	9.0	15.1	15.3	7.7	6.2	16.6
	2013	334.2	335.7	101.6	235.1	247.9	273.9
Badulla	2004	82.2	69.5	67.2	38.7	44.9	70.0
	2005	15.8	21.9	20.9	4.4	11.2	21.4
	2006	100.3	94.8	96.0	69.3	65.2	104.5
	2007	300.3	301.8	292.1	276.0	265.9	293.7
	2008	258.3	226.0	210.8	182.8	192.9	215.7
	2009	17.4	12.0	7.9	6.2	6.9	8.6
	2010	328.9	301.4	267.7	185.4	187.5	287.7
	2011	225.4	183.0	290.2	190.9	216.2	326.0
	2012	33.2	14.6	17.5	2.6	3.6	16.8
	2013	70.1	52.5	61.0	19.2	24.1	59.4

Table 7 - Most suitable methods for filling missing data based on comparison with observed data

Station	Methods
Bandarawela	IDW, PRB
Lower Spring Valley	NR, WLR
Badulla	NR, IDW

The zeros of the missing months cannot be exactly predicted using any of the methods. However for the Bandarawela station, the PRB and LR methods gave almost the same number of percentage zeros with respect to the observed data. For the Lower Spring Valley station, the PRB, LR and WLR methods gave almost similar results. Both NR and IDW methods performed well for the Badulla station.

The Bandarawela station has a very high correlation with its neighboring Diyathalawa station. These two stations are close to each other. The differences in elevation and distance between the two stations are very small. In the case of the Bandarawela station, the PRB method outperformed all the other methods, suggesting that the PRB method is highly dependent on the correlation between the target and the neighboring station. The MLR method also depends on the correlation coefficient.

The Lower Spring Valley station has the highest annual rainfall in the catchment and its rainfall pattern varies highly from those of neighboring stations. The data generation based on one single neighboring station was not effective for the Lower Spring Valley station and thus the PRB method or the LR method did not perform well. Four or five surrounding gauging stations were considered in the NR method and the results obtained were acceptable.

Compared to the other two stations, the Badulla station has a relatively lower CC value and higher STD and RMSE values. It has the lowest elevation above the mean sea level and all of its neighboring gauging stations are located at considerably higher elevations. Besides, the distances to the neighboring stations are also relatively high. Therefore, the rainfall pattern at the Badulla station will be different from those of its neighboring stations. These facts lead to lower CCs at the Badulla station compared to those of the others. The IDW method gave good predictions for the Badulla station. Table 8 summarizes the methods observed to be suitable based on the CCs and the number of neighboring stations that were considered.

Table 8 - Most suitable methods based on the correlation coefficient and the number of neighboring stations

Type of target station	Method
With only one neighboring station (high cc)	PRB LR
With more than one neighboring station (relatively low cc)	IDW NR
With more than one neighboring station (relatively high cc)	MLR WLR

4. Conclusions

Based on the analysis, it can be concluded that it is not possible to name one single method from among the seven methods studied as the most suitable one for all of the stations. However, the analysis identified appropriate methods to be employed in filling missing rainfall data at a gauging station based on the number of neighboring gauging stations available for use and their correlations with that particular station for which data are filled.

References

1. Campozano, L., Sanchez, E., Aviles, A. & Samaniego, E., 2014. Evaluation of Infilling Methods for Time Series of Daily Precipitation and Temperature: The Case of the Ecuadorian Andes. *MASKANA*, vol 5, pp.101-15.

2. De Silva, R., Dayawansa, N. & Ratnasiri, M., 2007. A Comparison of Methods Used in Estimating Missing Rainfall Data. *The Journal of Agricultural Sciences*, vol3, pp.101-08.

3. Hasan, M. & Croke, B., 2013. Filling Gaps in Daily Rainfall Data: A Statistical Approach. In *20th International Congress on Modelling and Simulation*. Adelaide, Australia, 2013.

4. Yoo, H., 2013. Linear Programming Method Considering Topographical Factors used for Estimating Missing Precipitation. *Journal of Hydrologic Engineering*, vol18, pp.542-51.

5. Chow, V. T., Maidment, D. R. & Mays, L. W. 1988. Applied Hydrology, McGraw-Hill Book Company, Singpore.

6. Singh, V. P. (1994) Elementary Hydrology, Prentice Hall of India, New Delhi.

7. Chen, F. W. & Liu, C. W. 2012. Estimation of the Spatial Rainfall Distribution using Inverse Distance Weighting (IDW) in the Middleof Taiwan, Paddy Water Environ., DOI 10, 1007/s10333-012-0319-4.