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:pavithranpiyathisse@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.

![Study Area](image)

**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 km². Figure 1 shows the gauging stations and their data availability is shown in Figure 2. The data were obtained from the Department of Meteorology.

![Rainfall data availability at the gauging stations](image)

**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.

![Neighbouring stations selected for each target station](image)

**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.
2. Methodology

The objective of this paper is to present the method in their place. A river basin situated in rainfall data records and propose a novel methods that are available for filling gaps in variability. The elevation of the catchment study based on their availability and spatial distribution of neighboring stations for each method. Gauging stations located at deterministically any surrounding gauges precipitation of any surrounding gauges exceeds 10% of the gauge that is under consideration. This weighs the effect of each surrounding station [6].

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

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].

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

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.

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

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].

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

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.

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

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

### Table 1 - Randomly deleted months of the three gauging stations

| Year | Lower Bandarawela | Spring Valley | Badulla |
|------|-------------------|---------------|---------|
| 2004 | February          | March         |         |
| 2005 | January           | June          |         |
| 2006 | August            | September     |         |
| 2007 | March             | April         | December|
| 2008 | May               | January       |         |
| 2009 | September         | February      |         |
| 2010 | 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  | Canawarella           | 3.127                                  | 0.784                   |
| Valley         | Wewessa               | 5.638                                  | 0.763                   |
|                | Wewessa middle        | 3.317                                  | 0.747                   |
|                | Ella                  | 7.816                                  | 0.639                   |
| Wewessa        | Ella                  | 6.721                                  | 0.669                   |
|                 | middle                | 6.444                                  | 0.509                   |
| Wewessa        | Lower Spring Valley   | 8.634                                  | 0.646                   |
|                 | Ella                  | 12.163                                 | 0.518                   |
|                 | Dyrabba               | 17.266                                 | 0.562                   |

The seven different techniques for estimating the coefficients are given in Table 2.

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.

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

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

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].

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

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].

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

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.

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

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].

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

### Table 1 - Randomly deleted months of the three gauging stations

| Year | Lower Bandarawela | Spring Valley | Badulla |
|------|-------------------|---------------|---------|
| 2004 | February          | March         |         |
| 2005 | January           | June          |         |
| 2006 | August            | September     |         |
| 2007 | March             | April         | December|
| 2008 | May               | January       |         |
| 2009 | September         | February      |         |
| 2010 | 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  | Canawarella           | 3.127                                  | 0.784                   |
| Valley         | Wewessa               | 5.638                                  | 0.763                   |
|                | Wewessa middle        | 3.317                                  | 0.747                   |
|                | Ella                  | 7.816                                  | 0.639                   |
| Wewessa        | Ella                  | 6.721                                  | 0.669                   |
|                 | middle                | 6.444                                  | 0.509                   |
| Wewessa        | Lower Spring Valley   | 8.634                                  | 0.646                   |
|                 | Ella                  | 12.163                                 | 0.518                   |
|                 | Dyrabba               | 17.266                                 | 0.562                   |

The seven different techniques for estimating the coefficients are given in Table 2.
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].

\[ p_X = \sum_{i=1}^{n} c_i p_i \]  \( \ldots(6) \)

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 data chronologically 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.
The data estimated for the target stations were generated using three particular gauging station that have similar rainfall values and available data for a month. The PRB method assumes that the missing rainfall was identified. Out of the probability distributions that match the most suitable for the daily rainfall data. The AM method could not be applied to any of the stations even though they are located close to the PRB method, therefore probability distributions that match the daily rainfall series, a probability distribution is fitted into the monthly rainfall estimated using the IDW method and based on the observed data. The monthly average rainfall variations in their elevations. 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 IDW 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.

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.

| Date       | Bandarawela | Actual | NR | IDW | LR | WLR | PRB |
|------------|-------------|--------|----|-----|----|-----|-----|
| January 2004 | 1 0 0.0 0.0 0.0 0.0 0 0 | 1 22.0 13.8 32.9 558.1 60.0 80.2 66.4 53.4 45.4 185.7 188.4 177.7 218.8 181 185.3 157.5 | 54.3 60.0 6.6 22.0 13.8 32.9 |
| October 2010 | 1 0 0.0 0.0 0.0 0.0 0 0 | 1 22.0 13.8 32.9 558.1 60.0 80.2 66.4 53.4 45.4 185.7 188.4 177.7 218.8 181 185.3 157.5 | 54.3 60.0 6.6 22.0 13.8 32.9 |

| Table 3 - Comparison of generated rainfalls with actual rainfalls at the three stations | Lower Spring Valley | Actual | NR | IDW | LR | WLR | PRB | MLR |
|---------------------------------|-------------------|--------|----|-----|----|-----|-----|-----|
| January 2004 | 1 0 0.0 0.0 0.0 0.0 0 0 | 1 22.0 13.8 32.9 558.1 60.0 80.2 66.4 53.4 45.4 185.7 188.4 177.7 218.8 181 185.3 157.5 | 54.3 60.0 6.6 22.0 13.8 32.9 |
| October 2010 | 1 0 0.0 0.0 0.0 0.0 0 0 | 1 22.0 13.8 32.9 558.1 60.0 80.2 66.4 53.4 45.4 185.7 188.4 177.7 218.8 181 185.3 157.5 | 54.3 60.0 6.6 22.0 13.8 32.9 |

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 target area 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.
Table 4 – Error statistics at the three stations

| Year | STD | RMSE | CC    | STD | RMSE | CC    | STD | RMSE | CC    |
|------|-----|------|-------|-----|------|-------|-----|------|-------|
| 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 | 6.220 | 4.477 | 4.970 | 5.066 | 5.034 | 6.136 | 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 | 11.848 | 10.013 | 9.665 | 7.606 | 9.301 | 8.176 | 9.567 | 0.659 | 0.762 | 0.813 | 0.928 | 0.954 |
| 2008 | 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 |
| 2009 | 6.761 | 7.823 | 8.431 | 7.923 | 9.562 | 7.538 | 3.050 | 8.457 | 8.082 | 9.465 | 0.568 | 0.647 | 0.515 | 0.574 | 0.487 |
| 2010 | 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 |
| 2011 | 9.118 | 6.923 | 5.575 | 9.292 | 8.292 | 8.970 | 6.815 | 5.883 | 8.157 | 0.911 | 0.829 | 0.944 | 0.935 | 0.876 |
| 2012 | 5.388 | 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 |
| 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 |

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, PRB| LR  | PRB  |     |
| Lower Spring Valley | LR, PRB | IDW | NR   |     |
| Badulla           | IDW, NR| IDW | NR   |     |

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 the 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.

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 5 - 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 Diyalathalawa 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 6 - Comparison of monthly rainfalls

| Year | Actual | NR | IDW | LR | WLR | PRB |
|------|--------|----|-----|----|-----|-----|
| 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| 153.3| 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| 117.0| 170.6|

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         |

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) | WLR |

4. Conclusions

Based on the analysis, it can be concluded that it is not possible to name a 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, Singapore.

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.