Estimation of Rainfall from Climatology Data Using Artificial Neural Networks in Palembang City South Sumatera

E Suhartanto1*, S Wahyuni1, K M Mufadhal1

1Water Resources Engineering Department, Faculty of Engineering, Universitas Brawijaya, Malang, 65145, Indonesia.

*Corresponding author: erysuhartanto@ub.ac.id

Abstract. Estimation of climatological parameters, especially rainfall is a data requirement for all regions of Indonesia. The availability of rainfall data is used for early warning of flood or drought disasters. The study location is in Palembang City, South Sumatra Province, where floods and droughts often occur and lack of availability of rainfall data. This study aims to obtain the best model in estimating rainfall from climatological data. The analysis was carried out to estimate the rainfall from the climatological data using the Artificial Neural Networks method. The Artificial Neural Networks were applied and showed some results with the best calibration was at 16 years using TRAINLM with 1500 epochs that is the performances NSE = 0.54, RMSE = 99.37, and R = 0.74. Whereas the best validation was at 1 year that is the performances NSE = 0.41, RMSE = 87.32, and R = 0.65.

keywords: Artificial neural networks, rainfall estimate, calibration, validation.

1. Introduction
Precipitation plays an important role in the hydrologic cycle. Precipitation is also the main focus in climatological studies. Learning about precipitation is very important in (a) precipitation characteristics identifying (b) precipitation statistical modeling and forecasting, and (c) floods and droughts mitigation. In tropical areas, the term precipitation has been replaced by rainfall, where snow is generally absent. The term rainfall is more commonly used than precipitation. The continuity and consistency of rainfall data are very significant in statistical analysis, such as time series analysis. Both continuity and consistency will be disturbed due to changes in the observational and incomplete records, which could vary in length from one or two days until many years [1].

Artificial Neural Networks (ANNs) are a strong method of computational that has been mainly used for pattern recognition, classification, and prediction. The major advantage of the ANNs as an alternative to the conventional and physical methods is that we do not use an explicit description in the mathematical equations for the complex processes of the system under development. Therefore, ANNs could generalize the robust nonlinear patterns of natural phenomena, including aggregation and disaggregation of rainfall with stabilization [2].

The objective study determines the calibration of rainfall and climatological data based on the ANNs models and the validation rainfall from the ANNs models and field stations.
2. Materials and Method

2.1. The study location
The study location is Palembang City, South Sumatra Province. Palembang City has an area of 400.61 km². Geographically, it is located between 2°52’ to 3°5’ South Latitude and 104°37’ to 104°52’ East Longitude, whereas elevation about 8 meters above seawater level.

![Figure 1. Palembang City, South Sumatra Province](image)

2.2. Data Collection
a. Monthly climatological data for 20 years (1999 to 2018).
b. Monthly rainfall data for 20 years (1999 to 2018).

2.3. Stage of rainfall analysis
a. Analysis of the quality of hydrological data based on the rainfall data that has been obtained. The tests used include; (i) consistency test using Rescaled Adjusted Partial Sums (RAPS) Method, and (ii) stationary test with F-test and t-test.
b. Rainfall analysis using the Artificial Neural Networks Models. In this study, data sharing is used by input, target, and modelling output. The division of data composition from the calibration and validation process into several parts as 15-5 years, 16-4 years, 17-3 years, 18-2 years, 19-1 years. This means that in the distribution of data for 15-5 years, the initial 15 years (1999-2013) are used as calibration process and 5 years outside of that year are used as validation process. The network architecture is made using various layers and epochs using the backpropagation algorithm.
2.4. Consistency test

2.4.1. Rescaled Adjusted Partial Sums (RAPS)

The RAPS is a method where data consistency is indicated by the cumulative value of the deviation from the average value [3].

\[
\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}
\]  

(1)

\[
S_k^* = (Y_i - \bar{Y})_t + (Y_i - \bar{Y})_{t+1}
\]

(2)

\[
D_y^2 = \frac{(Y_i - \bar{Y})^2}{n}
\]

(3)

\[
D_y = \sqrt{\sum_{i=1}^{n} D_y^2}
\]

(4)

\[
S_k^{**} = \frac{S_k^*}{D_y}
\]

(5)

\[
R = S_k^{**} \text{ max} - S_k^{**} \text{ min}
\]

(6)

where, \(Y_i\) is the observed data, \(\bar{Y}\) is the average of data observed, \(n\) is the number of the total observations.

| \(n\) | \(Q_{90\%}\) | \(Q_{95\%}\) | \(Q_{99\%}\) | \(R_{90\%}\) | \(R_{95\%}\) | \(R_{99\%}\) |
|-------|-------------|-------------|-------------|-------------|-------------|-------------|
| 10    | 1.050       | 1.140       | 1.290       | 1.210       | 1.280       | 1.380       |
| 20    | 1.100       | 1.220       | 1.420       | 1.340       | 1.430       | 1.600       |
| 30    | 1.120       | 1.240       | 1.460       | 1.400       | 1.500       | 1.700       |
| 40    | 1.130       | 1.260       | 1.500       | 1.420       | 1.530       | 1.740       |
| 50    | 1.140       | 1.270       | 1.520       | 1.440       | 1.550       | 1.780       |
| 100   | 1.170       | 1.290       | 1.550       | 1.500       | 1.620       | 1.860       |
| \(\infty\) | 1.220       | 1.360       | 1.630       | 1.620       | 1.750       | 2.000       |

2.4.2. Stationary test

The stationary test is to test the variance stability and average values of a time series. It is also to determine whether or not the variance (F-test) and average (t-test) values are homogeneous [4].

2.4.2.1. F-Test

\[
F = \frac{N_1 S_1^2 (N_2-1)}{N_2 S_2^2 (N_1-1)}
\]

(7)

where, \(N_1\) is the total number of the 1st group sample, \(N_2\) is the total number of the 2nd group sample, \(S_1\) is deviation standard of the 1st group sample, \(S_2\) is deviation standard of the 2nd group sample.

2.4.2.2. t-test

\[
\sigma = \sqrt{\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2}}^{0.5}
\]

(8)

\[
t = \frac{X_1 - X_2}{\sigma \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}^{0.5}
\]

(9)
where, $\bar{X}_1$ is average data of the 1st group sample, $\bar{X}_2$ is average data of the 2nd group sample, $N_1$ is the total number of the 1st group sample, $N_2$ is the total number of the 2nd group sample, $S_1$ is deviation standard of the 1st group sample, $S_2$ is deviation standard of the 2nd group sample.

2.4.3. Artificial Neural Networks (ANNs)

ANNs are inspired by biological neurons in the human brain and consist of the computational unit interaction. ANNs build a relationship between the input and output, also produce a good response by following the biological processes of human brain activities such as saving information, learning, and training. The structure of an ANN includes the input layer, hidden layer, and output layer [5].

![Figure 2. An example schematic of ANN](image)

2.4.4. Calibration and Validation

Calibration of a model is selecting a combination of parameters or an optimization process based on parameter values to improve the coherence of the observed and simulated watershed hydrological response [6].

The definition of validation or validity measures the extent to which the difference in scores reflects the actual difference between individuals, groups, or situations regarding the characteristics to be measured or the actual error in the same individual or group from one situation to another [6].

2.4.4.1. Nash- Sutcliffe Efficiency

The Nash-Sutcliffe efficiency (NSE) is a normalized statistic that determines the relative magnitude of the residual variance ("noise") compared to the measured data variance ("information"). NSE indicates how well the plot of observed versus simulated data fits the 1:1 line [7].

$$NSE = 1 - \left[ \frac{\sum_{i=1}^{n}(Y_{i}^{obs} - Y_{i}^{sim})^2}{\sum_{i=1}^{n}(Y_{i}^{obs} - Y_{mean})^2} \right]$$

(10)

where, $Y_{i}^{obs}$ is the $i$th observed for the constituent being evaluated, $Y_{i}^{sim}$ is the $i$th simulated value for the constituent being evaluated, $Y_{mean}$ is the mean of observed data for the constituent being evaluated, and $n$ is the total number of observed [7].

| NSE      | Performance Rating |
|----------|--------------------|
| 0.75 < NSE < 1.00 | Very good          |
| 0.65 < NSE < 0.75 | Good               |
| 0.50 < NSE < 0.65 | Satisfactory       |
| NSE < 0.50     | Unsatisfactory     |
2.4.4.2. Root Mean Squared Error
Several error indices are commonly used in model evaluation. These include root mean square error (RMSE). These indices are valuable because they indicate an error in the units (or squared units) of the constituent of interest, which aids in analyzing the results. RMSE values of 0 indicate a perfect fit [7].

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_i - Y_i)^2}{n}}
\]  
(11)

where \(X_i\) is the observed data, \(Y_i\) is the simulated value, and \(n\) is the total number of observations.

2.4.4.3 Correlation coefficient
It is an indicator of the strength of the relationship between observations and estimates. Higher positive coefficients indicate that estimates will be high or low when actual is high or low, respectively giving evidence about the suitability of the estimation method [7].

\[
R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}}
\]  
(12)

where, \(X_i\) is the observed data, \(\bar{X}\) is the average of observed data, \(Y_i\) is the simulated value, \(\bar{Y}\) is the average of simulated value [8].

### Table 3. General Performance Rating for R

| R          | Performance Rating |
|------------|--------------------|
| 0.80 < R < 1.00 | Very Strong       |
| 0.60 < R < 0.799 | Strong            |
| 0.40 < R < 0.599 | Moderate          |
| 0.20 < R < 0.399 | Low               |
| 0.00 < R < 0.199 | Very Low          |

3. Results and Discussion

#### 3.1. Consistency test
The results of the consistency test with RAPS Method at the degree of confidence = 5% and 1%, \(n = 240\), then obtained the value of \(Q_{critical} = 1.36\) for 5% and \(Q_{critical} = 1.75\) for 1%, while \(Q_{calculated} = 0.588\). Because \(Q_{calculated} (0.588) < Q_{critical} (1.75)\), so the rainfall data is considered consistent.

#### 3.2. Stationer test
The results of the stationary test with the F-test showed stable climatological data except for wind speed, while the t-test showed stable climatological data except for air temperature.

### Table 4. Summary Statistic of F-test and t-test

| Climatological Data | F\(_{calculated}\) | \(\alpha\) | F\(_{critical}\) | Performance |
|---------------------|-------------------|-----------|----------------|-------------|
| Wind Speed          | 2.41              | 1%        | 1.54           | Unstable    |
| Long sunshine       | 1.00              | 1%        | 1.54           | Stable      |
| Temperature         | 1.02              | 1%        | 1.54           | Stable      |
| Rainfall            | 0.96              | 1%        | 1.54           | Stable      |

| Climatological Data | t\(_{calculated}\) | \(\alpha\) | t\(_{critical}\) | Performance |
|---------------------|-------------------|-----------|----------------|-------------|
| Wind Speed          | -1.56             | 1%        | 2.58           | Stable      |
| Long sunshine       | 2.33              | 1%        | 2.58           | Stable      |
| Temperature         | -9.12             | 1%        | 2.58           | Unstable    |
| Rainfall            | -1.00             | 1%        | 2.58           | Stable      |
3.3. Artificial Neural Networks (ANNs)

ANNs are an information processing system that has characteristics resembling a human neural network. ANNs are created as an abstraction of mathematical models and human understanding. The characteristics of the ANNs can be determined by the network architecture, training, and activation function.

Feed-forward backup network is one of the algorithms that is often used in solving complex problems. It is possible because the network with this algorithm is trained using the guided learning method. The network is given a pattern consisting of the input pattern and the desired pattern. This exercise is done repeatedly so that all patterns have an output that can meet the desired pattern. In this study, the software used in determining the ANNs models in Matlab R2017a by utilizing the available toolbox. An example of configuration a network architecture on 16 years as a calibration stage and four years as a validation stage using 1500 epochs.

| Table 5. Configuration of Matlab R2017a |
|----------------------------------------|
| **Menu**                  | **Configuration**                     |
| Network Type               | feed-forward backup                   |
| Input data                 | input                                 |
| Target data                | target                                |
| Training Function          | TRAINLM                               |
| Adaption Learning Function | LEARNGDM                               |
| Performance Function       | MSE                                    |
| Number of Layers           | 5                                     |
| Number Of Neurons          | 3-4-5-3-1                             |
| Transfer Function          | LOGSIG-LOGSIG-LOGSIG-LOGSIG-PURELIN   |

The results of the ANNs models analysis based on performance, training state, and regression showed in Fig. 3, Fig. 4, and Fig. 5.

![Figure 3. Best Validation Performance](image1)

![Figure 4. Training State and Validation](image2)

The following results were obtained when ANNs models were developed and applied in the Palembang City (Table 6). The Summary statistics (Table 6) indicated satisfactory calibration and validation using 20 years of data, although the results showed that the calibration is better than the validation.

The best value of NSE used 1500 epochs during calibration is 0.54, and validation is 0.41. According to ANNs models simulated rainfall during calibration is satisfactory. However, on the other hand, validation is unsatisfactory, as shown by the statistical results.
Table 6. Summary statistics from the ANNs models analysis

| Epoch | Calibration (15 Years) | Validation (5 Years) | Calibration (16 Years) | Validation (4 Years) | Calibration (17 Years) | Validation (3 Years) | Calibration (18 Years) | Validation (2 Years) | Calibration (19 Years) | Validation (1 Years) |
|-------|-------------------------|-----------------------|-------------------------|-----------------------|-------------------------|-----------------------|-------------------------|-----------------------|-------------------------|-----------------------|
|       | NSE | RMSE | R | NSE | RMSE | R | NSE | RMSE | R | NSE | RMSE | R |
|-------|-----|------|--|-----|------|--|-----|------|--|-----|------|--|
| 1000  | 0.43| 110.93| 0.66 | -0.81 | 173.38 | 0.17 |       |       |   |       |       |   |
| 1500  | 0.43| 110.91| 0.66 | -0.22 | 152.84 | 0.19 |       |       |   |       |       |   |
| 2000  | 0.40| 113.95| 0.63 | -0.17 | 150.07 | 0.23 |       |       |   |       |       |   |
| 1000  | 0.41| 112.37| 0.64 | 0.26  | 109.10 | 0.54 |       |       |   |       |       |   |
| 1500  | 0.54| 99.37 | 0.74 | -0.86 | 172.60 | 0.11 |       |       |   |       |       |   |
| 2000  | 0.40| 113.88| 0.63 | -0.22 | 139.94 | 0.20 |       |       |   |       |       |   |
| 1000  | 0.49| 104.79| 0.73 | 0.03  | 117.49 | 0.57 |       |       |   |       |       |   |
| 1500  | 0.51| 102.54| 0.72 | -2.91 | 235.24 | 0.11 |       |       |   |       |       |   |
| 2000  | 0.52| 101.26| 0.72 | 0.11  | 112.07 | 0.52 |       |       |   |       |       |   |
| 1000  | 0.27| 124.91| 0.63 | 0.02  | 107.44 | 0.48 |       |       |   |       |       |   |
| 1500  | 0.45| 108.22| 0.68 | 0.27  | 92.67  | 0.56 |       |       |   |       |       |   |
| 2000  | 0.35| 117.85| 0.59 | 0.19  | 97.65  | 0.52 |       |       |   |       |       |   |
| 1000  | 0.44| 107.49| 0.67 | 0.29  | 96.19  | 0.55 |       |       |   |       |       |   |
| 1500  | 0.45| 106.85| 0.67 | **0.41** | **87.32** | **0.65** |       |       |   |       |       |   |
| 2000  | 0.44| 108.33| 0.66 | 0.31  | 94.81  | 0.60 |       |       |   |       |       |   |

The best RMSE values used 1500 epochs during calibration is 99.37, and validation is 87.32. It indicated that the ANNs model's performance for rainfall error in the square unit on validation was a closer perfect fit than calibration.

The coefficient correlation values varied from 0.59 to 0.74 during calibration and from 0.11 to 65 during validation. It indicated that the ANNs model's performance for correlation on calibration is stronger than validation.
The evaluation of applied the ANNs models in the Palembang City, situations possibility that produces contrary performance results for various location and/or output indicators.

In conditions with contrary performance results, those differences should be clearly declared. As an example, if the result for one output indicator in one location generates unbalanced performance results of "unsatisfactory" on NSE, "closer perfect fit" on RMSE, and "strong" on R, then the overall performance should be declared conservatively as "unsatisfactory" for that one location and that one output indicator. However, it should be preferable to declare the performance of trends (NSE) as "satisfactory or unsatisfactory," the error in the square unit (RMSE) as "closer perfect fit, and in the simulation of correlation (R) as "strong."

4. Conclusions
The conclusions are summarized below:
1. Estimation of rainfall from climatological data showed that the best calibration results using 16 years of data (1999-2014) with 1500 epochs in a simulation of trends (NSE) = 0.54 (satisfactory), in a simulation of error in the square unit (RMSE) = 99.37 (closer perfect fit), and in a simulation of correlation (R) = 0.74 (strong).
2. Estimation of rainfall from climatological data showed that the best validation results using 1-year data (2018) with 1500 epochs in a simulation of trends (NSE) = 0.4 (unsatisfactory), in a simulation of error in the square unit (RMSE) = 87.32 (closer perfect fit), and in a simulation of correlation (R) = 0.65 (strong).
3. By applying the ANN Models to estimate the rainfall using climatology data, the calibration performance is better than validation, as shown by the statistical results.

References
[1] 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 no 2 pp 101–108
[2] S. Kim and V. P. Singh 2015 Spatial disaggregation of areal rainfall using two different artificial neural networks models Water (Switzerland) vol 7 no 6 pp 2707–2727 doi: 10.3390/w7062707
[3] S. Harto Br 1993 Analisis Hidrologi Jakarta: Gramedia Pustaka Utama
[4] Soewarno 1995 Hidrologi, Aplikasi Metode Statistik untuk Analisa Data, Jilid 2, 2nd ed. Bandung: Nova
[5] A. Jabbari and D. H. Bae 2018 Application of Artificial Neural Networks for accuracy enhancements of real-time flood forecasting in the Imjin basin Water (Switzerland) vol 10 no 11 doi: 10.3390/w10111626
[6] Indarto 2010 Hidrologi, Dasar Teori dan Contoh Aplikasi Model Hidrologi Jakarta: Bumi Aksara
[7] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith 2007 Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations Trans. ASABE vol 50 no 3 pp 885–900 doi: 10.13031/2013.23153
[8] L. Suryaningtyas, E. Suhartanto, and Rispiningtati 2020 Hydrological Analysis of TRMM (Tropical Rainfall Measuring Mission) Data in Lesti Sub Watershed Civ. Enviromental Sci. J. vol 3 no 1 pp 18–30