Bias correction of daily satellite precipitation data using genetic algorithm

A W Pratama¹, A Buono¹, R Hidayat², H Harsa¹,³
¹Department of Computer Science, Faculty of Mathematics and Natural Science, Bogor Agricultural University, Bogor, Indonesia
²Department of Geophysics and Meteorology, Faculty of Mathematics and Natural Science, Bogor Agricultural University, Bogor, Indonesia
³Meteorological, Climatological, and Geophysics Agency of Indonesia (BMKG), Kemayoran, Jakarta, Indonesia

Email: angga_jisung13@apps.ipb.ac.id

Abstract. Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS) was produced by blending Satellite-only Climate Hazards Group InfraRed Precipitation (CHIRP) with Station observations data. The blending process was aimed to reduce bias of CHIRP. However, Biases of CHIRPS on statistical moment and quantile values were high during wet season over Java Island. This paper presented a bias correction scheme to adjust statistical moment of CHIRP using observation precipitation data. The scheme combined Genetic Algorithm and Nonlinear Power Transformation, the results was evaluated based on different season and different elevation level. The experiment results revealed that the scheme robustly reduced bias on variance around 100% reduction and leaded to reduction of first, and second quantile biases. However, bias on third quantile only reduced during dry months. Based on different level of elevation, the performance of bias correction process is only significantly different on skewness indicators.

1. Introduction
In field of precipitation analysis, the problem which is commonly faced, is availability of observation data [1]. Rain gauge stations in Indonesia are unevenly distributed and the quantity is limited [2]. To address limitations of observation data, Satellite Precipitation Products (SPPs) is frequently used as alternative [1]. Generally, near global coverage SPPs are only available on spatial resolution 0.25° or lower [3]. However, Climate Hazard Group InfraRed Precipitation (CHIRP) dan Climate Hazard Group InfraRed Precipitation with Stations (CHIRPS) provide 0.05° spatial resolution. CHIRPS was developed by U.S Geological Survey (USGS) and University of California (UCSB). CHIRPS was developed to support drought monitoring system Famine Early Warning Systems Network (FEWS NET) and focus over Africa area [4]. However, [5] and [3] revealed that CHIRPS still showed systematic bias. [6] performed preliminary study to find out the performance of CHIRPS over Java Island, and concluded that CHIRPS still has systematic bias on statistical value.

Bias correction is required to do before uses SPPs for further analysis, because the accuracy of SPPs is affected by coastal and topography aspects [2, 7]. Bias correction process on several earlier studies evidently improved performance of SPPs data [1, 2, 7]. In this paper, Nonlinear Power Transformation...
(NL) combined with Single Objective Genetic Algorithm (SOGA) is utilized to address systematic bias of CHIRPS over Java Island. The equation of GA used in this research is adopted from Pratama et al. (2018), however we focus to investigate the impact of the equation and search space based on spatial area (separated by elevation).

2. Methodology

Flow charts of this research is showed by Figure 1. Precipitation data of fourteen observation stations are used (Table 1). January precipitation data of Pondok Betung station is used as training data in training process. Training process is aimed to search optimal genetic algorithm parameter, contain of probability crossover (pc), probability mutation (pm), population size, and maximum generation. Temporal resolution of data is daily-type data, from January 2005 to December 2016.

Table 1. List of station.

| No | Location      | Latitude | Longitude | Elevation (masl) |
|----|---------------|----------|-----------|-----------------|
| 1  | Pondok Betung | -6.2615  | 106.751   | 27              |
| 2  | Kemayoran     | -6.1556  | 106.84    | 4               |
| 3  | Darmaga       | -6.5     | 106.75    | 207             |
| 4  | Bandung       | -6.8836  | 107.597   | 791             |
| 5  | Tegal         | -6.8682  | 109.121   | 1               |
| 6  | Cilacap       | -7.7189  | 109.015   | 8               |
| 7  | Semarang      | -6.9847  | 110.381   | 6               |
| 8  | Sangkapura    | -5.85    | 112.63    | 0               |
| 9  | Juanda        | -7.3846  | 112.783   | 3               |
| 10 | Karang Ploso  | -7.9008  | 112.598   | 590             |
| 11 | Karang Kates  | -8.15    | 112.45    | 285             |
| 12 | Kalianget     | -7.05    | 113.97    | 0               |
| 13 | Sawahan       | -7.74    | 111.79    | 835             |
| 14 | Banyuwangi    | -8.215   | 114.355   | 52              |

Figure 1. Research method.

2.1. Single Objective Genetic Algorithm

Search space and equation used are based on Pratama et al. (2018). The equation and search space is shown in Equation 1. Parameter of genetic algorithm is explored to minimize time required in searching process of solution. Reference value of pc and pm is adopted from Arkeman et al. (2012), pc is between 0.7 and 0.9 then explored value of pm is between 0.01 and 0.2. Population sizes explored are 10, 15, 20, 25 while number of maximal generation is 100, 200, 300, 400, 500.

The objectives on estimating parameter of NL method are to minimize the difference in both mean and CV of corrected data to observation data. In this paper, post process is added to mitigate possibility of more than one solution which have maximum fitness value. When the mentioned condition occurs,
A single solution is selected based on criteria of minimal root mean square error (RMSE) on monthly precipitation. Figure 2 shows the process of SOGA with additional post process.

\[ P^* = \begin{cases} 
   a_1, p^{b_1}, & P > \theta_1 \\
   a_2, p^{b_2}, & 2 \theta_3 < P < \theta_2 \text{ or } \theta_2 < P < 2 \theta_3 \\
   P, & \text{otherwise} 
\end{cases} \]

Subject to,

\[
\begin{align*}
\bar{x}_{\text{CHIRP}} + s_{\text{CHIRP}} - 2 < \theta_1 < \bar{x}_{\text{CHIRP}} + 2 * s_{\text{CHIRP}} + 2 \\
(Q_1^{\text{CHIRP}} * 2) < \theta_2 < (Q_2^{\text{CHIRP}} + 3) \\
(Q_1^{\text{CHIRP}}/2) < \theta_3 < (Q_1^{\text{CHIRP}} + 1) \\
0.1 < a_2 < 1 \\
0.1 < b_2 < 1 \\
0.7 < a_3 < 4 \\
0.7 < b_3 < 4
\end{align*}
\]

Objective function of SOGA in this study is a combination of the objectives. As a replacement of CV, we use variance ($S^2$) and combine it with mean ($\bar{x}$) to construct objective function. Objective function of SOGA is shown in Equation 2. c is constants which is weighted the absolute difference of mean value. In this paper, value of c selected is three. Equation 2 is invers of the objective. Maximum value of Equation 2 is one.

\[
\max f(\text{obs}, \text{cor}) = \frac{1}{1 + \frac{|S^{2}_{\text{obs}} - S^{2}_{\text{cor}}| + c * |\bar{x}_{\text{obs}} - \bar{x}_{\text{cor}}|}{2}}
\]

2.2. Evaluation

Evaluation on this research is evaluating the obtained results based on statistical moment and quantile value. Furthermore, corrected CHIRP constructed is evaluated on different month and elevation. Elevation is grouped by 0 – 20 masl, 20 – 200 masl, 200 – 500 masl, and above 500 masl.
3. Result and Analysis

3.1. Single Objective Genetic Algorithm
Optimal parameters of SOGA obtained from training process were probability crossover (pc) 0.75, probability mutation (pm) 0.01, population size 15, and number of generations 500. Result of SOGA is shown in Table 2.

Table 2. Fitness Values of Stations.

| Month | BMKG Stations |
|-------|---------------|
| 1 | 0.52 0.27 0.61 0.77 0.21 0.84 0.49 0.82 0.94 0.41 0.93 0.29 0.42 0.46 |
| 2 | 1.00 0.68 0.60 0.72 0.34 0.97 0.95 0.73 0.79 0.78 0.94 0.88 0.21 0.81 |
| 3 | 0.83 0.73 0.98 0.99 0.90 0.49 0.31 0.43 0.88 0.29 0.40 0.30 0.24 0.21 |
| 4 | 0.56 0.21 0.31 0.88 0.31 0.95 0.48 0.64 0.92 0.39 0.69 0.64 0.35 0.99 |
| 5 | 0.81 0.96 0.76 0.26 0.54 0.29 0.84 0.25 0.62 0.24 0.28 0.23 0.41 0.57 |
| 6 | 0.98 0.96 0.88 0.82 0.49 0.87 0.97 0.56 0.57 0.33 0.98 0.48 0.66 0.95 |
| 7 | 0.99 0.98 0.59 0.99 0.66 0.93 0.61 0.45 0.71 0.35 0.60 0.75 0.31 0.68 |
| 8 | 0.44 0.50 0.46 0.66 0.52 0.41 0.69 0.99 0.94 0.70 0.88 1.00 0.74 0.99 |
| 9 | 0.75 0.90 0.49 0.82 0.51 0.70 0.94 0.46 0.82 0.22 0.84 0.93 0.97 0.68 |
| 10 | 0.41 0.48 0.56 0.29 0.49 0.57 0.85 0.66 0.78 0.31 0.55 0.59 0.44 0.40 |
| 11 | 0.96 0.96 0.63 0.80 0.44 0.02 0.94 0.34 0.28 0.59 0.83 0.41 0.57 0.67 |
| 12 | 0.45 0.51 0.36 0.69 0.39 0.67 0.29 0.83 0.46 0.45 0.74 0.88 0.99 0.47 |

GA doesn’t always product solution with high fitness value. The low fitness value represents the problem should not solve with single objective. However, high fitness value reveals we can adjust mean and variance of corrected CHIRP equal to observation data with some degree, higher fitness value means higher the similarity of mean and variance of corrected data to observation.

3.2. Evaluation
Both CHIRP, CHIRP, and corrected CHIRP (using SOGA) product bias against observation data. with calculating bias of fourteen station, we have mean absolute error (MAE) of tho those stations. Statistical test analysis of variance (ANOVA) is used to investigate the difference of MAE, Tuckey HSD test is applied when MAE is significantly different. P value of ANOVA is shown by Table 3.

Table 3. P value of ANOVA.

|            | DJF     | MAM     | JJA     | SON     |
|------------|---------|---------|---------|---------|
| Q1         | 8.51E-25** | 1.46E-18** | 3.07E-11** | 2.78E-05** |
| Q2         | 2.05E-33** | 6.28E-17** | 5.74E-14** | 7.17E-06** |
| Q3         | 0.638    | 0.407   | 0.0005** | 0.296    |
| Mean       | 6.67E-16** | 4.57E-09** | 3.52E-08** | 3.12E-11** |
| Variance   | 6.21E-38** | 1.32E-23** | 5.38E-15** | 3.66E-16** |
| Skewness   | 1.89E-10** | 1.09E-05** | 0.003**   | 0.0097**  |
| Kurtosis   | 0.325    | 0.552   | 0.714    | 0.683    |

** significant with level α = 0.01
Kurtosis indicator of three data is not significantly different (Table 3), the bias correction can not reduce bias on this indicator. Bias on first and second quantile was significantly different, therefore bias on third quantile is different only during dry season (JJA).

Performance of CHIRPS during dry months is great. It is represented by quantile values which are nearly identical to observation and the bias on variance reduces relative to CHIRP (Table 5). This result corresponds to the background of CHIRPS establishment which is to support drought monitoring [4]. Comparing to our corrected CHIRP, MAE of $Q_3$ indicator between CHIRPS and corrected CHIRP is not significantly different (Table 4). Most of stations used in this paper have first quantile value equal to zero ($Q_1 = 0$). This condition occurs also during wet months, the number of non-rain days are around 25% from total wet months. Both CHIRP and CHIRPS cannot estimate number of non-rain days well during wet months, it is represented from first quantile value not equal to zero. After bias correction process, bias on first quantile reduces between 74% and 96% (Table 5).

### Table 4. Mean Absolute Error of quantile indicator with tuckey HSD index.

|        | Q1  | DJF | MAM | JJA | SON |
|--------|-----|-----|-----|-----|-----|
| CHIRP  | 4.187a | 2.817a | 0.167a | 0.901a |
| CHIRPS | 2.876b | 0.525b | 0b | 0.126b |
| NL-SOGA| 0.994c | 0.661b | 0.008b | 0.227b |

|        | Q2  | DJF | MAM | JJA | SON |
|--------|-----|-----|-----|-----|-----|
| CHIRP  | 5.607a | 4.83a | 1.055a | 2.51a |
| CHIRPS | 5.94a | 3.432b | 0.005b | 1.422b |
| NL-SOGA| 1.912b | 1.55c | 0.352c | 0.697b |

|        | Q3  | DJF | MAM | JJA | SON |
|--------|-----|-----|-----|-----|-----|
| CHIRP  | 3.319a | 4.122a | 2.353a | 2.738a |
| CHIRPS | 3.889a | 4.784a | 1.104b | 2.48a |
| NL-SOGA| 3.319a | 4.052a | 1.315b | 2.078a |

Bias on third quantile is reduced only during dry months (Table 5). Proposed approach robustly reduces bias on first and second quantile. However Bias on kurtosis increases after bias correction process (shown with negative values).

### Table 5. Mean of percentage of bias reduction relative to CHIRP (%).

| Month | Q1    | Q2    | Q3    | Mean  | Var   | Skew  | Kurt  |
|-------|-------|-------|-------|-------|-------|-------|-------|
| DJF   | 76.25 | 65.90 | 0.00  | 69.38 | 99.96 | 11.66 | -43.22|
| MAM   | 76.55 | 67.91 | 1.71  | 46.83 | 99.82 | 35.13 | 5.43  |
| JJA   | 95.37 | 66.63 | 44.13 | 53.84 | 99.92 | 48.92 | 19.77 |
| SON   | 74.83 | 72.24 | 24.14 | 51.81 | 97.90 | 40.55 | 22.27 |
Figure 3. Percentage of bias reduced based on different station elevation.

For different spatial area, bias on mean indicator continually increases on August (Figure 3). It is impact of adjusting variance of corrected chirp identical to observation data thus bias on mean increases, because basically mean value of chirp approaches observation although the variance value is significantly different. Skewness and kurtosis represent shape and asymmetry of distribution [9]. Bias on skewness can be handled robustly only on 0 - 20 masl elevation station group. Value of skewness relates to extreme value. Second threshold in Equation 1 was added to transform high precipitation of CHIRP becoming extreme precipitation value because Pratama et al. (2018) found that CHIRP cannot predict extreme precipitation value of observation data. However, when high precipitation values of CHIRP is equivalent to observation, the bias correction process will cause the high precipitation value getting further to observation which products higher bias. In other hand, bias reduction on third quantile during dry months does not happen on stations which have elevation between 200 and 500 masl, it means number of not-rain days during dry months on the stations are less than 75% of total data.

4. Conclusion
The brief result of this study is similar to Pratama et al. (2018), the bias correction process cannot handle bias on mean, skewness, kurtosis robustly since the bias on several months increase after correction process. The performance of corrected CHIRP is not separated by different spatial area (elevation). However, performance of corrected CHIRP is separated by different season.

Selecting low value of c (in Equation 2) makes genetic algorithm less sensitive to difference of mean between corrected data and observation data. The impact is, model cannot handle bias on mean because the model only focus on address the difference on variance.

References
[1] Mamenun 2014 Validasi dan koreksi data satelit TRMM pada tiga pola hujan di Indonesia Jurnal Meteorologi dan Geofisika, 15 13-23
[2] Vernimmen R. R. E., Hooijer A, Mamenun and Aldrian E 2012 Evaluation and bias correction of satellite rainfall data for drought monitoring in Indonesia Hydrology and Earth System Sciences, 16 133–146
[3] Katsanos D, Retalis A and Michaelides S 2016 Validation of a high-resolution precipitation database (CHIRPS) over Cyprus for a 30-year period Atmospheric Research, 169 459-464
[4] Funk CC, Peterson PJ, Landsfeld MF, Pedreros DH, Verdin JP, Rowland JD, et al. A quasi-global
precipitation time series for drought monitoring. US Geological Survey, 2014 2327-638X.

[5] Toté C, Patricio D, Boogaard H, Wijngaart R V D., Tarnavsky E and Funk C 2015. Evaluation of satellite rainfall estimates for drought and flood monitoring in Mozambique Remote Sensing, 7 1758-1776

[6] Pratama AW, Buono A, Hidayat R, and Harsa H 2018 Estimating parameter of nonlinear bias correction method using NSGA-II in daily precipitation data TELKOMNIKA Accepted, 17

[7] Valdés-Pineda R, Demaría E M C, Valdés J B, Wi S and Serrat-Capdevil A 2016 Bias correction of daily satellite-based rainfall estimates for hydrologic forecasting in the Upper Zambezi, Africa Hydrology and Earth System Sciences Discuss

[8] Arkeman Y, Gunawan H, and Seminar K B 2012 Algoritma genetika teori dan aplikasinya untuk bisnis dan industri (Bogor: IPB Press)

[9] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1996 Numerical recipes in C (Cambridge: Cambridge University press) pp 610-613