Sriwijaya Journal of Environment
Article
ojs.pps.unsri.ac.id

The Application of Modeling Gamma-Pareto Distributed Data Using GLM Gamma in Estimation of Monthly Rainfall with TRMM Data

Herlina Hanum[1*], Aji Hamim Wigena[2], Anik Djuraidah[2], and I Wayan Mangku[3]

[1]*Department of Mathematics, Sriwijaya University, Kampus Inderalaya Km 32, Ogan Ilir 30662, Indonesia.*
[2]*Department of Statistics, Bogor Agricultural University, Jalan Meranti, Darmaga, Bogor ,Indonesia*
[3]*Department of Mathematics, Bogor Agricultural University, Jalan Meranti, Darmaga, Bogor, Indonesia*
**Corresponding author email: linhanum@yahoo.com.au*

Article history

Received	Received in revised form	Accepted	Available online
24 April 2017	10 May 2017	15 May 2017	30 May 2017

Abstract: As a recently developed distribution, the application of Gamma-Pareto is limited to single variable modeling. A specific transformation of Gamma-Pareto (G-P) yields gamma distribution. Therefore, it is possible to use analysis based on gamma distribution (e.g. GLM) for modeling G-P distributed data. In this paper we study the application of modeling G-P distributed data using GLM gamma for monthly rainfall which observed in Sukadana Station. The modeling aims to analyze whether Tropical Rainfall Measuring Mission (TRMM) satellite data is a good estimator for unobserved station's data. The transformed of station's data were considered as response variable in GLM gamma. The explanatory variable is TRMM data in 9 grids around the station. There are two kinds of modeling i.e. model for whole data and extreme data. The results show that for both data the station's data are G-P distributed and the transformed data are gamma distributed. TRMM rainfall data at each grid around the station can be used to estimate the observed data of monthly rainfall. The best model for both data contains dummy variables which correspond to inter quantile data. The coefficients of dummy variables in the best model may substitute the grouping or the correction in the previous studies.

Keywords: Gamma-Pareto, gamma, GLM, monthly rainfall, TRMM

Abstrak (Indonesian): Sebagai sebaran yang baru dikembangkan, aplikasi sebaran Gamma-Pareto (G-P) masih terbatas pada pemodelan peubah tunggal. Transformasi spesifik terhadap sebaran G-P menghasilkan sebaran gamma. Oleh karena itu, dimungkinkan menggunakan analisis berbasis sebaran gamma untuk pemodelan data bersebaran G-P. Aplikasi untuk beberapa data simulasi menunjukkan bahwa pemodelan data bersebaran G-P dengan menggunakan model linier terampat (GLM) gamma menghasilkan estimasi yang hanya tergantung pada kondisi peubah penjelas. Dalam makalah ini dikaji pemodelan data bersebaran G-P menggunakan GLM gamma untuk curah hujan bulanan yang diamati di Stasiun Sukadana. Peubah penjelas adalah Tropical Rainfall Measuring Mission data satelit (TRMM) di 9 grid di sekitar stasiun. Pemodelan bertujuan untuk menganalisis apakah data TRMM adalah estimator yang baik untuk data yang tidak teramati di stasiun. Hasil transformasi data stasiun digunakan sebagai peubah respon dalam GLM gamma. Ada dua model yang dibentuk yaitu model untuk data keseluruhan dan untuk data ekstrim. Hasil menunjukkan untuk keduanya data stasiun bersebaran G-P dan transformasinya mengikuti sebaran gamma. Data curah hujan TRMM pada setiap jaringan di sekitar stasiun dapat digunakan untuk memperkirakan data curah hujan bulanan yang diamati di stasiun. Model terbaik, baik untuk data keseluruhan maupun data ekstrim, mengandung peubah boneka yang berhubungan dengan data antarkuantil. Koefisien peubah boneka dapat menggantikan pengelompokan atau koreksi dari penelitian sebelumnya.

Kata kunci: Gamma-Pareto, gamma, GLM, curah hujan bulanan, TRMM.

1. Introduction

G-P distribution is a combination of gamma and Pareto distribution with pdf

$$g(y) = \frac{\theta^{-1}}{\varrho^{\alpha}\varrho(\alpha)}\left(\log\left(\frac{y}{\theta}\right)\right)^{\alpha-1}\left(\frac{y}{\theta}\right)^{(-\varrho-1)} \qquad (1)$$

where $\alpha, \varrho, \theta > 0$ and $y > \theta$. The distribution shows a better fit than some distributions for three types of data by [1]. While [2] used G-P distribution in modeling monthly extreme rainfall. The application of G-P distribution is still limited for modeling single variable data.

Furthermore (Alzaatreh et al.)[1] noted that transformation Y which is G-P distributed into $\log(\frac{Y}{\theta})$ results in a variable following gamma distribution. With this transformation, it is possible to analyze G-P distributed data using analysis based on gamma distribution i.e. GLM gamma. The GLM gamma is regression analysis which is developed for gamma distributed response variable [3]. Hanum et al. [4] used GLM gamma to analyze the relationship between simulated G-P distributed response variable with explanatory variable. The result showed that goodness of the model only depend on the goodness of fit the response variable to G-P and the strength of the relationship of response and explanatory variable. This result is just like common modeling problem.

Rainfall data is very important in climate study. Unfortunately, there are some reasons which cause the rainfall data is being unobserved. In order to estimate the unobserved data, we try to use the data which is observed by TRMM satellite. TRMM is satellite which is operated by the collaboration between National Aeronautics and Space Administration (NASA), and Japan Aeronautics Exploration Agency (JAXA) [5]. Some researches and techniques were established to study the used of TRMM data as the completion of station's data. Within these research, in Indonesia [6] yield the correction forms to TRMM data in 3 pattern of rainfall in Indonesia. While [7] used downscaling technique to estimate the rainfall data based on TRMM rainfall data.

In this research, we used TRMM data as explanatory variable (X) in order to estimate the rainfall data in Sukadana station (Y). This research has two goals. The first goal is to apply the modeling G-P distributed data using GLM gamma, while the second is to assess the goodness of TRMM data as the estimate of unobserved rainfall data in Sukadana station.

2. Experimental Sections

2.1. Data Source

This research used two data sets. The first data is monthly rainfall data from Sukadana station Inderamayu West Java, while the second is 9 grids TRMM's rainfall data around Sukadana station. TRMM data is from type 3B43 version 7. Both data are taken from the period of 1998-2012. This 'old' data means to compare with [7], and to adjust to the goals of estimating the unobserved data at station. In this research, both data are divided into analysis (1998-2010) and validation data (2011-2012). The analysis data is used for modeling, while validation data is used for assessing the validation of the model to another data. Figure 1 showed the position of Sukadana station and 9 grids of TRMM. Extremes rainfall data is contained station's rainfall data which exceed quantile 75 %.

Figure 1. Position of Sukadana station and 9 grids TRMM

2.2. Fitting station's rainfall data to Gamma-Pareto distribution

Fitting data begins with parameter estimation of the certain distribution based on the data. Parameter estimation of G-P follows the method in [1] and [2]. Based on the estimator of the G-P parameter, then we determined the quantile values of G-P using quantile function of G-P in [2]. Kolmogorov-Smirnov test [8] is used to assess the goodness of fit between data and quantile values.

2.3. Modeling G-P distributed data using GLM gamma

The station's data (Y) which follows G-P distribution is transformed using $log\,(\frac{Y}{\theta})$. Parameter θ is estimated by $Y_{(1)}$ the minimum value of Y. The result of the transformation (U) which is taken as response variable, with one of 9 grid TRMM as the explanatory variable (X), is analyzed using GLM gamma to obtain the estimator of U, that is $\hat{U}$. The estimator of Y, that is $\hat{Y}$, is obtained by reverse transform $\hat{U}$ using $\hat{Y} = Y_{(1)}e^{\hat{U}}$.

2.4. Model selection

Data analysis yields some models, whether due to different explanatory variable or due to the amount of the explanatory variables in the model. The best model is selected based on some criteria. Those criteria are Akaike Information Criteria (AIC) [9] in GLM gamma, Mean Absolute Percent Error (MAPE) [10], correlation between Y and $\hat{Y}$, and Root Mean Square Error (RMSE) [11]. The best model is the model with smallest AIC, MAPE, and RMSE, and greater correlation coefficient.

3. Results and Discussion

3.1.The result of fitting monthly rainfall data to G-P distribution

In order to certain that this data can be analysis using GLM gamma, first we fit the response variable *Y*, that is monthly rainfall data of Sukadana station in years 1998-2010, to G-P distribution. The histogram of *Y* in Figure 2 shows that the distribution of *Y* is not symmetrical. The distribution has right tail that a bit far from the mode's values. This form of distribution of the data may fit to Gamma-Pareto distribution.

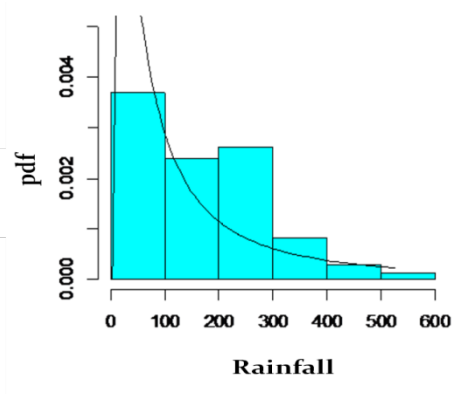

Figure 1. Distribution of monthly rainfall of Sukadana station year 1998-2010 and the pdf of Gamma-Pareto

The pdf of G-P distribution of rainfall data described by the black curve in Figure 2 where only for rainfall between 200-300 mm/month the pdf doesn't fit the data. The estimator of parameters of G-P distribution for the data are α = 11.3499, ϱ = 0.4128, and θ = 1. Kolmogorov- Smirnov test gives the p-value of 0.0988 which is greater than the significant level of 0.05. This means that the rainfall data of Sukadana station has accordance with Gamma-Pareto distribution.

Transformation of Y into $U = log\,(\frac{Y}{\theta})$ yield variable U with parameters α = 11.3499 and ϱ = 0.4128 in gamma distribution. Variables U fit to gamma distribution with P-value 0.0988 in Kolmogorov-Smirnov test. This is consistent with the statement of [1]. The similarity of Y and U is not only in parameter α and ϱ but also at the level of conformity of Y to G-P distribution and U to gamma distribution.

3.2.The modeling by grid 7

Based on the certainty that U is follows gamma distribution, we start the analysis using the GLM gamma with U as the response variable. In the first model we used only rainfall data of TRMM at Grid 7 (we denote it as variables *grid 7*) as explanatory variable. Variable *grid 7* provides estimator which has MAPE value 1.04, the correlation with the data station 0.5917, and p-value 0.0182 in Kolmogorov-Smirnov test. These measures of goodness of indicate that the model with only *grid 7* does not provide a good estimation on rainfall data at Sukadana station. Given the correlation value for other grids with the station data is almost equal to the value of the correlation to the *grid 7*, the goodness of the models using other grids is expected to be nearly equal to the goodness of model with *grid 7*.That is the reason why we do not model U with other grid at this point.

In order to improve the estimation, we try the modeling with the addition of dummy variable. Dummy variable *D1* is used to separate the lower (set to 1) and the upper (set to 0) of quantile 50 (*q50*). The used of *D1* is based on [6] which notes that the TRMM satellite data has good estimate to low rainfall, but not good enough for high rainfall. This means that there is different distribution between low and high rainfall. Dummy variable *D1*, gives different models between low and high rainfall. The addition *D1* into model with explanatory variables *grid 7* was able to significantly improve the model. It decreases the value of MAPE to 0.6869. This means that *D1* can minimize the distance between the data and the estimator. On the other hand, the increasing correlation to 0.7978 showed that *D1* improve the conformity of fluctuations between data and the estimator. With MAPE > 0.5 means this model is not good enough.

Previous study [7] grouping Sukadana station rainfall data into three sections is enough to obtain a good model. Two of them are above *q50*; they are 165-400 mm/month for group 2 and greater than 400 for group 3. This means that there are different models for the data above *q50*. Accordingly we try again another dummy variable that can separate models for the data above *q50* using dummy variables *D2* and *D3*. Dummy variable *D2* is intended to separate the model on data between *q50* to *q75* with other data. Meanwhile *D3* is used to obtain the model for the data over *q75*.The addition of *D2* and *D3* on model with only *grid 7* is not much different in the goodness from the model with *D1*.The addition of *D2* and *D3* clearly improved the goodness than the model with only *grid 7*.Unfortunately the value of MAPE is still large (> 0.5). So we form again several dummy variables that can separate the model on the other inter quantile data. The dummy variables are presented in Table 1.

Table 1 . Dummy variables

Data	*D1*	*B1*	*B2*	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*	*D8*	*D9*
<q25	1	1	0	0	0	0	0	0	0	0	0
q25-q50	1	0	1	0	0	0	0	0	0	0	0
q50-q75	0	0	0	1	0	0	0	0	0	1	1
q75-q90	0	0	0	0	1	1	0	0	0	1	1
q90-q95	0	0	0	0	1	0	1	1	0	0	1
>q95	0	0	0	0	1	0	1	0	1	0	0

Some models are developed based on those dummy variables. The result is presented in Table 2. The separation of the model to the data above *q50* is not much different than the separation models by *D1*. It can be seen that the goodness of *Model 3* and *Model 4*

is almost similar with the *Model 2*. Instead, the model with separation of the data by q25 (*Model 5*) is significantly better than *Model 2*. It can be seen from the comparison *Model 5, 6, 7* and *8* with *Model 2*. The separation of the data above *q50,* given *B2,* provides little improvement of the goodness of model.

Table 2. The goodness of fit for model with grid 7 and dummy variables

Model	Explanatory variables	AIC	cor(y,yh)	MAPE	RMSE
1	*grid 7*	409.77	0.5917	1.05	207.4300
2	*grid 7+D1*	377.42	0.7978	0.6869	95.9893
3	*grid 7+D2+D3*	378.83	0.8199	0.6791	96.0099
4	*grid 7+D2+D4+D5*	380.60	0.8335	0.6785	99.3500
5	*grid 7+B1+B2*	311.37	0.8661	0.3223	60.4200
6	*grid 7+B2 +D7+D9*	311.75	0.9112	0.3451	50.0000
7	*grid 7+B2+D6+D7+D8*	312.19	0.9270	0.3364	46.5527
8	*grid 7+ B2+D2+D4 +D6+D7*	312.00	0.9561	0.3106	36.4043

In Table 2 *Model 8* is the best model based on greater correlation between *Y* and its estimate and smaller AIC, MAPE, and RMSE. On the other hand, *Model 5* could be considered as simplest model with good criteria. These two models yield the estimate which is have correlation to *Y* more than 0.85, MAPE less than 0.33, and RMSE less than 100. The form of GLM gamma $g(\mu)$ of *Model 5* and *Model 8* are

$$\widehat{M5} = 1.5753 + 0.0004\ grid\ 7 - 0.7502\ B1 - 0.2248\ B2$$

$$\widehat{M8} = 0.832 + 0.00025\ grid\ 7 + 0.5409\ B2 + 0.7412\ D2 + 0.803\ D4 + 0.8493\ D6 + 0.8697\ D7$$

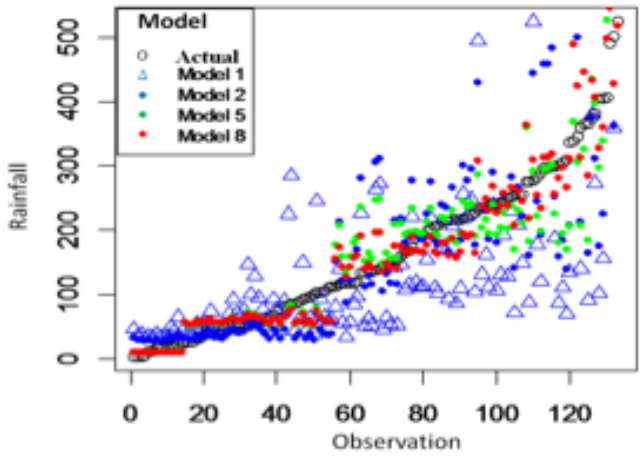

Figure 2. The estimation of rainfall by grid 7 and dummy variables

The estimation of station's rainfall by *Model 1*, *Model 2*, *Model 5*, and *Model 8* is presented in Figure 3. It can be seen that Model 1 yield estimators mostly below the value of 200 mm/month while more than one half of the real data are above that value. The addition of D1 highly improve the goodness of the estimation, although is not good enough. The best estimation is given by *Model 8* where the estimate very close to the real data.

3.3.The estimation by another grid

Based on the best model for grid 7, we used another 8 grid to estimate the real data using *Model 8*. Table 3 showed that the goodness of the model yield by those 8 grids almost similar with the goodness of Model with grid 7. Generally, for *Model 8*, all grids can well explain rainfall data observed at Sukadana station. All grid generates the estimation with the correlation to the real data more than 0.9.

Table 3. The goodness of estimation by each grid TRMM around Sukadana Station

Explanatory variable	cor(X,Ŷ)	AIC	cor(Y,Ŷ)	MAPE	RMSE
grid 7	0.803133	312.00	0.9561	0.3106	36.4043
grid 8	0.783181	311.15	0.9577	0.3060	35.5269
grid 9	0.764774	311.13	**0.9581**	0.3067	**35.3818**
grid 12	0.797844	311.04	0.9522	0.3073	38.1748
grid 13	0.779445	311.59	0.9545	0.3064	36.8621
grid 14	0.759237	311.19	0.9571	**0.3059**	35.5690
grid 17	0.759399	311.52	0.9506	0.3100	38.5425
grid 18	0.768038	311.49	0.9517	0.3071	38.0500
grid 19	0.767223	**310.96**	0.9542	0.3069	36.8470

With approximately 0.3 MAPE value shows that each grid can be used to predict rainfall data station pretty well. There is no significant difference in the goodness estimation between those grids. It means each grid of TRMM could be used for estimating the rainfall at station. Therefore, the prediction of rainfall in 2011 and 2012 below, we only use grid 7 as explanatory variables in *Model 5* and *Model 8*.

3.4. Validation of the best model

Validation of the best models i.e. *Model 5* and *Model 8* used validation data of year 2011 and 2012. Determination of the values of dummy variables is based on the average value of monthly rainfall for last 5 years i.e. 2006-2010. The limits for the dummy variable monthly rainfall data Sukadana station based on data from 1998-2010 is b1 <23 23≤b2 <113, 113≤D2 <229, 229≤D4 <303, 303≤D6 <369, and D7≥369. A particular dummy variable will be set to 1 if the average value falls within its range, unless a dummy variable equal to zero. For example, the average rainfall in June is 68 mm which falls in range of D2, so D2 take value of 1, and 0 for the others.

In general, the prediction of rainfall at Sukadana station in 2011 and 2012 by data grid 7 TRMM using *Model 5* and *Model 8*, is good enough. Both models provide the estimation which approaching the actual. For rainfall data in May of 2011 and 2012 as well as the April 2012 the estimation is not good enough for their considerable difference between the average value of 5 years and the observed data in these months. This difference causes an error in determining the value of dummy variables for those months. As the consequence there is a considerable distance between the data observations and the estimate. Error in determining the

value of dummy variables has become a problem if the model contain dummy variable. To avoid the mistake, *Model 5* is better to use. Rainfall in January and February 2011 was much lower than the average monthly rainfall of 2006-2010. *Model 5* and *Model 8* estimate rainfall of these two months higher because, based on the average value, both fall in the range of D4. Meanwhile, rainfall in December 2012 was not observed, but the data is made zero. Since December is the month of rain, a value of zero is not a reasonable rainfall data. Both models provide appropriate true value for the month, which is close to the average value.

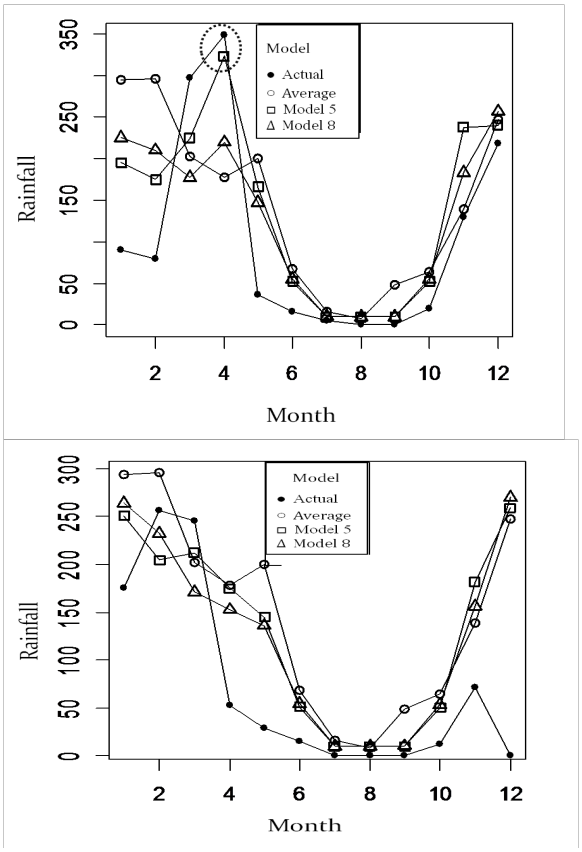

Figure 3. Estimation of rainfall at Sukadana Station year 2011 and 2012

3.5.Modeling the extremes rainfall

The extreme rainfall data in this research is 25% highest rainfall or data above q75 at Sukadana stations. For this data q75 value is 229 mm/month. This extreme rainfall is G-P distributed with $\theta = 233$, $\alpha = 1.1458$, and $\varrho = 0.247$. This set of data is very good fit to G-P distribution with p-value 0.9844 in Kolmogorov-Smirnov test. Based on the results of modeling the whole data, the modeling of extreme rainfall also uses data in grid 7 of TRMM as explanatory variable. *Model Q7* in Table 4 is the model with single variable grid 7. Similar with *Model 1* for whole data, *Model Q7* with high MAPE and low correlation is not a good model for extreme rainfall.

Table 4. Models for extreme rainfall

Name	GLM Model	AIC	cor(y,yh)	MAPE	RMSE
Q7	-2.1335 + 0.0024 grid 7	-18.31	0.4483	0.1663	68.52
Q74	-1.313 + 0.0016 grid 7 -1.2994 D4	-41.39	0.8361	0.0922	41.89
Q746	-1.0546 + 0.0014 grid7 ---1.4818 D4 – 0.3808 D6	-40.60	0.9109	0.0777	30.52

The addition of dummy variable D4 separates the model for data in range q75 to q90. *Model Q74* with D4 in Table 4, has AIC, MAPE and RMSE significantly lower than *Model Q7*. On the other hand the correlation between estimated and observed data is significantly improved. The goodness of *Model Q74* can be slightly improved by adding dummy variable D6 which separates the model for data between q90 and q95.

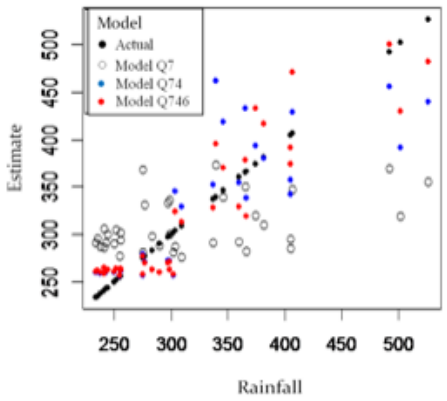

Figure 4 Estimation of extreme rainfall

Figure 5 shows that the model without the dummy variable gives estimation about the value of 250-350 mm/month for all value of rainfall. The value of the actual rainfall spreads from 233 to 526. The addition of D4 lowered the estimated value of data under q90, i.e. 303, and changed the slope of estimation for data between q75 and q90. The addition of D6 into *Model Q74* only improves predictions at the data above q90. Unfortunately, both models variables are not able to estimate two outliers around 500.

4. Conclusion

Modeling Gamma-Pareto distributed data can be done using the GLM gamma. In order to provide gamma distributed response variable in GLM gamma, we firstly transform the variable which follows Gamma-Pareto distribution into variable which follows gamma distribution. The estimation of Gamma-Pareto distributed data will be obtained from re-transformation the result of GLM gamma.

This modeling can be applied to rainfall data in stations Sukadana which is Gamma-Pareto distributed. The result of modeling shows that the rainfall data from 9 grid satellite TRMM located around Sukadana station can be used to estimate the unobserved monthly rainfall in Sukadana station. The best model contains dummy variables. The coefficients of dummy variables in the

best model improve the grouping or the correction in the previous studies.

The monthly extreme rainfall of Sukadana station is very well fitted to Gamma-Pareto distribution. The best model for the data also includes dummy variable for data between quantile 75 and quantile 95. There is difference model for inter quantile data.

References

[1] A. Alzaatreh, F. Famoye, and C. Lee, "Gamma-Pareto Distribution and Its Applications," *J. Mod. Appl. Stat. Methods*, vol. 11, no. 1, p. Article 7, 2012.

[2] H. Hanum, A. H. Wigena, A. Djuraidah, and I. W. Mangku, "Modeling extreme rainfall with Gamma-Pareto distribution," *Appl. Math. Sci.*, vol. 9, no. 121, pp. 6029–6039, 2015.

[3] R. Balaji, "GLM with a Gamma-distributed Dependent Variable," 2013. [Online]. Available: http://civil.colorado.edu/~balajir/CVEN6833/lectures/GammaGLM-01.pdf.

[4] H. Hanum, A. H. Wigena, A. Djuraidah, and I. W. Mangku, "Modeling Gamma-Pareto distributed data using GLM Gamma," *Glob. J. Pure Appl. Math.*, vol. 12, no. 4, pp. 3569–3575, 2016.

[5] TRMM, "Tropical Rainfall Measuring Mission," 2015. [Online]. Available: http://trmm.prnm.nasa.gov.

[6] M. Mamenun, H. Pawitan, and A. Sophaheluwakan, "Validasi dan koreksi data satelit TRMM pada tiga pola curah hujan di Indonesia," *J. Meteorol. dan Geofis.*, vol. 15, no. 1, pp. 13–23, 2014.

[7] A. D. Warawati, "Prakiraan curah hujan Station Sukadana dengan teknik downscaling berdasarkan data Satelit TRMM," Institut Pertanian Bogor, 2013.

[8] H. Hassani and E. Silva, "A Kolmogorov-Smirnov Based Test for Comparing the Predictive Accuracy of Two Sets of Forecasts," *Econometrics*, vol. 3, pp. 590–609, 2015.

[9] K. P. Burnham and D. R. Anderson, "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociol. Methods Res.*, vol. 33, no. 2, pp. 261–304, 2004.

[10] J. J. M. Moreno, A. P. Pol, A. S. Abad, and B. C. Blasco, "Using the R-MAPE index as a resistant measure of forecast accuracy.," *Psicothema*, vol. 25, no. 4, pp. 500–506, 2013.

[11] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, pp. 1247–1250, 2014.