Volume 8 No 2 December (2025)

e-ISSN: 2615-3270
p-ISSN: 2615-3599

Eigen Mathematics Journal

homepage: https://eigen.unram.ac.id/index.php/eigen

Prediction of Rainfall in Lampung Province Using Tweedie Mixture Distribution with PCA Reduction

Sari Utami[1], Ma'rufah Hayati[2*], Reni Permatasari[1]

[1] *Department of Statistics, Faculty of Science and Teknologi, Nahdlatul Ulama University Lampung, Indonesia*
[2] *Department of Actuarial Science, Faculty of Science, Institut Teknologi Sumatera, Indonesia*
[*] *Corresponding author: marufah.mt@at.itera.ac.id*

ABSTRACT

Accurate rainfall prediction is crucial for supporting the agricultural sector in Lampung Province. This research employs the Exponential Dispersion Model (EDM), a special case of the Generalized Linear Model (GLM), incorporating a Tweedie mixture distribution with Principal Component Analysis (PCA) to reduce correlated variables. Rainfall data were obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) through twelve rain observation posts (2013-2022), and supplemented with precipitation data from the General Circulation Model (GCM) obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF). The Tweedie mixture distribution was selected for its ability to handle non-normally distributed rainfall data containing zero values. The results show that the Root Mean Square Error of Prediction (RMSEP) for the Tweedie mixture-PCA model at the Gisting Atas station is 163.90, while the Normal-PCA model achieved 169.11. Therefore, the Tweedie mixture-PCA approach is more effective and recommended for improving rainfall prediction in Lampung Province, offering potential benefits for agricultural planning and resource management.

Keywords: Generalized Linear Model, Tweedie Mixture, Principal Component Analysis, Exponential Dispersion Model, Rainfall.

Received : 12-04-2025;
Revised : 14-09-2025;
Accepted : 07-10-2025;
Published : 11-11-2025;
.

DOI: https://doi.org/10.29303/emj.v8i2.280

1. Introduction

As an archipelagic nation, Indonesia experiences two distinct seasons : the dry season and the rainy season. Rainfall plays a crucial role in various sectors, particularly agriculture, irrigation, and development planning. The agricultural sector heavily relies on stable and well-distributed rainfall to ensure optimal crop yields [1].

Lampung Province, as one of the main food granaries in Indonesia, possesses 1.7 million hectares of agricultural land. According to the Lampung Province Food Security, Food Crops and Horticulture Office [2], Lampung makes a significant contribution to national agricultural production. Governor Arinal Djunaidi reported that Lampung's economic growth in the first quarter of 2023 exceeded the

average growth in Sumatra. The province has become a major producer of agricultural commodities and plays an important role in national food security. However, irregular and unpredictable rainfall patterns pose significant risks, such as crop failure and reduced productivity. Lampung's spatially variable and complex rainfall patterns pose a challenge in understanding rainfall characteristics and achieving accurate predictions for agricultural planning and resource management [3].

Rainfall data recorded by the Meteorology, Climatology and Geophysics Agency (BMKG) exhibit nonlinear, non-normally distributed, uncertain and fluctuating characteristics. This complexity necessitates the selection of appropriate models and predictions. One of the relevant predictor variables for this analysis is precipitation data generated by the General Circulation Model (GCM). However, GCM data has a low resolution, making it less than ideal for precise local climate prediction [4].

Given these limitations of GCM data, Statistical Downscaling (SD) techniques are necessary. This technique is employed to link large-scale global data with local observations, reduce the dimensionality of rainfall data, and improve modeling efficiency [5]. GCM outputs, such as precipitation, serve as predictor variables in SD models. However, the outputs of adjacent GCM grids are often correlated, violating the multicollinearity assumption in statistical modeling. Principal Component Analysis (PCA) is widely recognized as a solution to this problem by reducing the number of correlated variables and improving model efficiency [6].

Rainfall has two components: discrete and continuous. Rainfall events can occur with high intensity ($Y > 0$) or low intensity ($Y \leq 0$). When there is no rain ($Y = 0$), the distribution is discrete, whereas when rain occurs, the distribution becomes continuous. The non-normally distributed characteristics of rainfall make it unsuitable for use in standard normal distribution models. The Poisson distribution assumes that rainfall is always greater than zero, while the Gamma distribution cannot accommodate a value of zero without additional adjustment [7].

The mixed Tweedie distribution used in the Generalized Linear Model (GLM) framework effectively addresses these challenges. Since the mixed Tweedie distribution is part of the Exponential Dispersion Model (EDM) family, it combines the characteristics of the Poisson distribution (for non-rainfall events) and the Gamma distribution (for rainfall intensity), making it effective for handling rainfall data that is not normally distributed and contains a large number of zero-inflated values. In addition, the right-skewed nature of the Tweedie distribution is in line with the natural characteristics of rainfall which often exhibits a non-symmetrical distribution [8], [9].

Despite its many advantages, Statistical Downscaling (SD) modeling using GCM outputs often faces multicollinearity problems, which can be overcome by Principal Component Analysis (PCA). Studies by [10] show that PCA effectively reduces the number of correlated predictor variables, thereby improving model efficiency and accuracy. Based on this methodology, this study integrates PCA with the Tweedie mixture distribution to develop an innovative rainfall prediction model called the Tweedie mixture-PCA model. The model aims to accurately predict discrete and continuous rainfall components, including rainfall intensity and average rainfall occurrence.

This study validates the proposed model by comparing it with the standard Tweedie model, using metrics such as Root Mean Square Error of Prediction (RMSEP) and correlation. The performance of the model was evaluated using rainfall data obtained from 12 observation stations. This research is expected to contribute to better agricultural planning, optimal resource management, and climate resilience strategies in Lampung Province.

2. Research Materials and Methods

2.1. Study Design

This research employs a quantitative, model-based approach to predict rainfall in Lampung Province. Historical rainfall data and General Circulation Model (GCM) output data were analyzed within a Statistical Downscaling (SD) framework [5]. The objective of this study is to improve rainfall prediction accuracy by integrating Principal Component Analysis (PCA) with a Generalized Linear Model (GLM) using a Tweedie mixture distribution. The research design encompasses the following steps : (1) data collection and preprocessing, (2) feature extraction using PCA, (3) model building using GLM with a Tweedie mixture distribution, and (4) model validation using RMSEP and correlation metrics.

2.2. Data Sources

The data used in this study originated from two primary sources.

2.2.1. Rainfall Data

Daily rainfall data from January 2013 to December 2022 (120 months) were obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) for twelve rainfall observation posts in Lampung Province (Table 1). These observation posts were strategically selected to represent the diverse terrain types within the province. The geographical location of the rainfall data is situated between −3°S to −6°S and between 104°E to 106°E units of mm/day.

As detailed in Table 1 provides a description of the data utilized in this study, specifying both the number of observations and the variables included.

Table 1. Rainfall Data Description.

Variable	Rainfall Observation Station Locations
$y_{(n\times 1)}$	Data from twelve rainfall observation stations in Lampung Province
	1. Gisting Atas 7. Pajaresuk 2. Biha 8. Sumber Rejo 3. Krui Pasar 9. Fajar Mataram 4. Balik Bukit 10. Simpang Pematang 5. Way Tuba 11. Sukadana Hilir 6. Way Rerem 12. Way Urang
$X_{(n\times k)}$	Output data from GCM models related to precipitation

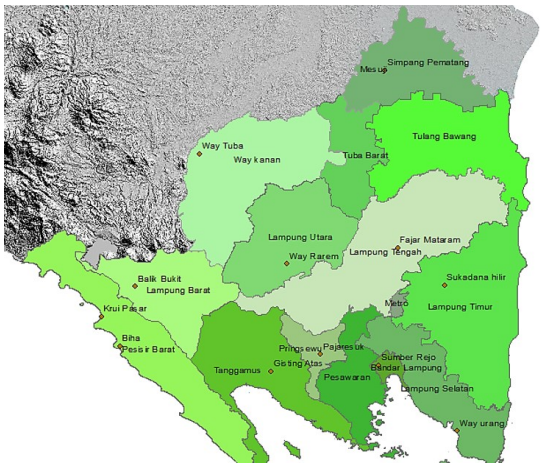

Figure 1. Location map of 12 rainfall stations of Lampung Province.

2.2.2. GCM Data

Precipitation data from the General Circulation Model (GCM) were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) in NetCDF format. This data, accessed via https://cds.climate.copernicus.eu/#!/home, comprised a 13×13 grid with a spatial resolution of $0.5° \times 0.5°$, resulting in 169 variables. These GCM data serve as predictor variables in the statistical downscaling model.

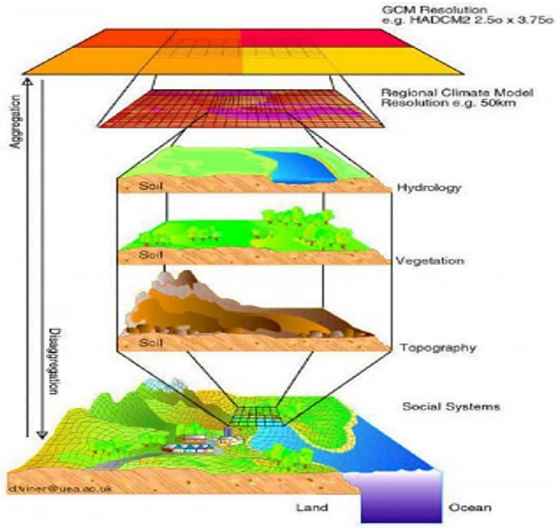

Figure 2. Schematic of the application of Statistical Downscaling techniques.

2.3. Data Preprocessing

2.3.1. Statistical Downscaling (SD)

To bridge the scale mismatch between large-scale GCM data and local rainfall observations, Statistical Downscaling (SD) techniques were applied [11]. This process involved establishing a statistical relationship between GCM outputs (precipitation) and observed local rainfall data. SD methods were implemented to correct potential biases in GCM data, enhancing its suitability for local-scale rainfall prediction. The outputs of rainfall and precipitation are used as variables in the SD model.

2.3.2. Principal Component Analysis (PCA)

Given the high dimensionality and potential multicollinearity among the 169 GCM grid variables, Principal Component Analysis (PCA) was employed for dimensionality reduction and feature extraction [12]. PCA transforms the original set of correlated variables into a smaller set of uncorrelated variables, known as principal components (PCs), while retaining most of the variance in the data. The selection of the number of PCs to retain was based on the Kaiser criterion, retaining components with standard deviation values greater than 1 [13]. In this study, the first eight principal components (PC.1 to PC.8) were selected as predictor variables for the rainfall model, capturing the dominant modes of variability in the GCM data.

2.4. Rainfall Modeling

2.4.1. Tweedie Mixture Distribution

The Tweedie mixture distribution was selected as the response distribution for the rainfall model due to its ability to accommodate both the discrete (no rain) and continuous (rainfall amount) components of rainfall data [14]. Unlike traditional distributions (e.g., Normal, Gamma, Poisson), the

Tweedie distribution can effectively handle non-normally distributed rainfall data containing a significant number of zero values, which is a common characteristic in the study area. This distribution belongs to the exponential dispersion model (EDM) family.

2.4.2. Generalized Linear Model (GLM)

Rainfall modeling was conducted using a Generalized Linear Model (GLM) with a Tweedie mixture response distribution [15]. The GLM framework allows for modeling the relationship between the Tweedie-distributed rainfall data and the selected principal components (PC.1 to PC.8) as predictor variables. The regression model used can be formulated as

$$\log(\mu) = \beta_0 + \beta^T x \tag{1}$$

where $\log(\mu)$ is the link function connecting the expected value (μ) with the linear combination of predictors, β_0 is the intercept, β^T is the vector of regression coefficients, and x represents the vector of predictor variables (PC.1 to PC.8).

2.5. Parameter Estimation

Model parameters, including the Tweedie index parameter (p), dispersion parameter (φ), and regression coefficients (β), were estimated using the maximum likelihood estimation (MLE) method. Specifically, the profile likelihood method was used with the Tweedie package in R, utilizing the tweedie.profile() function, to estimate the index parameter (p). The data analysis was performed using R software, employing the statmod and tweedie packages.

2.6. Model Evaluation

The performance of the developed Tweedie-GLM-PCA model was evaluated using several metrics, including

- Root Mean Squared Error of Prediction (RMSEP) : A measure of the average magnitude of the errors in the predictions,
- Correlation : A measure of the strength and direction of the linear relationship between predicted and observed rainfall values.

To ensure robust model evaluation, the dataset was partitioned into training and testing datasets. The training data (e.g., 80% of the data) was used to estimate model parameters, while the testing data (e.g., 20% of the data) was used to assess the model's predictive performance on unseen data.

2.7. Mathematical Representation of the Tweedie Distribution

The probability density function of the Tweedie distribution is expressed as [16]

$$f_y(y \mid \theta, \phi) = a(y, \phi) \exp\left(\frac{1}{\phi}\left[y\theta - k(\theta)\right]\right) \tag{2}$$

where θ functions as the canonical parameter, while ϕ serves as the dispersion parameter, linked to the variance via a dispersion constant. Additionally, $a(y, \phi)$ represents the normalization constant. The Tweedie distribution is a versatile statistical framework capable of modeling diverse data features, including those exhibiting varying levels of skewness and kurtosis factor which is scalar and independent of the parameters θ [17].

The Tweedie model is a member of this distribution family, with its density function $\alpha(y, \phi)$, being contingent upon these parameters. Notably, the Tweedie distribution encompasses several widely recognized distributions, such as the normal distribution (when $p = 0$), the Poisson distribution (when

$p = 1$), and the Gamma distribution (when $p = 2$). Consequently, it provides an adaptable structure for analyzing diverse data types [18]. When applied to rainfall data, N_p signifies the aggregate monthly rainfall, N_r indicates the count of rainfall events per month, and Y_r denotes the rainfall observed during the t-th event. The corresponding mathematical expression is

$$P(N = n) = \frac{e^{-\lambda}\lambda^n}{n!}, \quad \forall n \in \mathbb{N}_t, \tag{3}$$

$$N = \sum_{t \geq 1} \mathbf{1}_{[t,\infty)}(t). \tag{4}$$

The overall rainfall Y is defined as the cumulative amount of rainfall from each individual event. If $N = 0$ then Y is zero, whereas if $N > 0$, Y is calculated as the sum Σy_i. The probability density function for Y, under the condition that $N > 0$, is presented in [19]

$$\begin{cases} \mu = \lambda\,\alpha\,\gamma, \\ p = \dfrac{\alpha+2}{\alpha+1}, \\ \varphi = \dfrac{\lambda^{1-p}\,(\alpha\gamma)^{2-p}}{2-p}, \end{cases} \quad \text{parameterized by} \quad \begin{cases} \lambda = \dfrac{\mu^{2-p}}{\varphi\,(2-p)}, \\ \alpha = \dfrac{2-p}{p-1}, \\ \gamma = \varphi\,(p-1)\,\mu^{p-1} \end{cases} . \tag{5}$$

As stated in [15], the likelihood of no rainfall can be determined using the following formula

$$\pi = P(Y = 0) = e^{-\lambda} = \exp\left(-\frac{\mu^{2-p}}{\varphi(2-p)}\right). \tag{6}$$

This expression is equivalent to the subsequent equation

$$P(Y, N = n \mid \lambda, \alpha, \gamma) = d_0(y)\, e^{-\lambda} \left(\frac{y^{n\alpha-1}e^{-y/\beta}}{\beta^{n\alpha}\,\Gamma(n\alpha)}\right)\left(\frac{\lambda^n e^{-\lambda}}{n!}\right). \tag{7}$$

With $d_0(y)$ denoting the Dirac delta function at zero, the joint distribution $P(Y, N = n \mid \lambda, \alpha, \gamma)$, can attain a closed-form expression. As detailed in [20], this is accomplished by substituting Equation 5 into Equation 7. Consequently, the joint density function, characterized by $\{\mu, \phi, p\}$, is given by

$$P(Y, N = n \mid \mu, \phi, p) = \left[\exp\left(-\frac{\mu^{2-p}}{\phi(2-p)}\right)\right]^{n=0} \times \left[\exp\left(n\left(-\frac{\log(\phi)}{p-1} + \frac{2-p}{p-1}\log\left(\frac{y}{p-1}\right) - \log(2-p)\right) - \log\Gamma(n+1)\right)\right]^{n=0} \left[-\frac{1}{\phi}\left(\frac{\mu^{1-p}y}{p-1} + \frac{\mu^{2-p}}{2-p}\right) - \log\Gamma\left(\frac{2-p}{p-1}n\right) - \log(y)\right].$$

2.8. Data Analysis Step

The data analysis steps are as follows.

2.8.1. Visualizations of Data Distribution Characteristics (Density Plot, Histogram, Boxplot)

This step involves descriptive analysis to understand the nature of the rainfall data.

- Density Plot (Kernel Density Estimation/KDE).
  The general equation for KDE is

$$f(x) = \frac{1}{nh}\Sigma_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{8}$$

  where $f(x)$ is estimated probability density at point x, n is number of data points, h is badwidth (smoothig parameter), K is kernel function (e.g., Gaussian, Epanechnikov), and x_i is the i-th data point [21].
- Histogram.
  - There isn't a single equation, but it involves counting the frequency of data within each bin (interval).
  - Frequency = number of data in bin/ bin width
- Boxplot.
  Displays quartiles and outliers. Calculations involve :
  - Q1 (Quartile 1) : The value at the 25^{th} percentile.
  - Q2 (Median) : The value 50^{th} percentile.
  - Q3 (Quartile 3) The value at the 75^{th} percentile.
  - IQR (Interquartile Range) : $Q3 - Q1$.
  - Upper Bound : $Q3 + 1.5 \times IQR$.
  - Lower Bound : $Q1 - 1.5 \times IQR$.
  - Outlier : Data outside the upper or lower bounds [22].

2.8.2. Identification of the Tweedie Distribution and Index Parameter (p)

- The Tweedie distribution is an exponential family distribution with three parameters: the index parameter p, the scale parameter ϕ (phi), and the location parameter μ (mu). The Probability Density Function (PDF) of the Tweedie distribution does not have a closed form in general but is defined through its characteristic function .
- The index parameter p determines the type of Tweedie distribution. For rainfall data, values $1 < p < 2$ are often used, which corresponds to a compound Poisson-Gamma distribution.

2.8.3. Estimation of phi (ϕ) and Index Parameter with Tweedie.profile()

- The Tweedie.profile() function likely uses the profile likelihood method to find the values of [ϕ] and p that maximize the likelihood function.
- The likelihood function for the Tweedie distribution (in general) is very complex and involves integrals that are difficult to calculate analytically.
- Numerical optimization methods (e.g., Newton-Raphson) are used to find the optimal parameter values.

2.8.4. Application of the Generalized Linear Model (GLM) with Tweedie Response

- The GLM connects the expected value of the response variable (rainfall) with a linear combination of predictors through a link function [23]. The general equation for GLM is

$$g(X) = X\beta$$

where μ is expected value of the response variable $E(Y)$ and $g()$ is link function (e.g., log, identity, inverse). The choice of link function depends on the nature of the data and the Tweedie distribution used, X is predictor matrix, and β is vector of regression coefficients.

- Since the response is Tweedie, the Tweedie distribution is used in the likelihood function for parameter estimation.

2.8.5. Rainfall Prediction

After the Tweedie GLM is estimated, prediction are made using

$$\hat{Y} = g^{-1}(X\beta)$$

where $\hat{Y}$ is predicted rainfall value and $g^{-1}()$ is inverse of the link function.

2.8.6. Model Performance Evalution (RMSEP)

RMSEP (Root Mean Squared Error of Prediction) measures the prediction accuracy of the model, that is

$$RMSEP = \sqrt{\frac{(Y_i - \hat{Y}_i)^2}{n}}$$

where Y_i is observed rainfall value, $\hat{Y}$ is predicted rainfall value, and n is number of observations [24].

3. Result and Discussion

3.1. Data Distribution Characteristics (Visualization)

To understand the characteristics of the rainfall data, visualizations were performed using density plots, histograms, and boxplots.

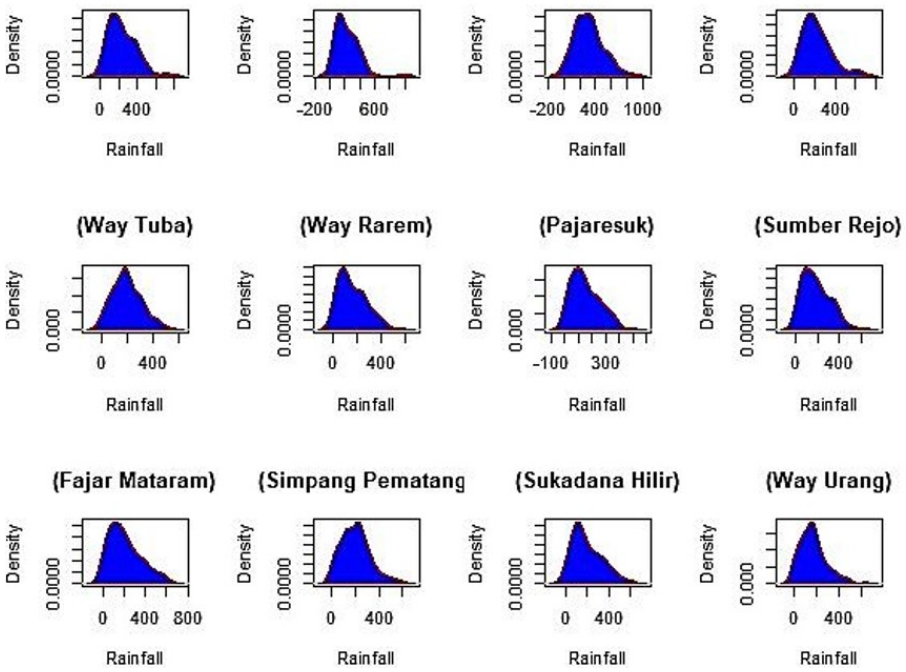

Figure 3. Data Density Plot of 12 Rain Stations in Lampung Province.

In Figure 3, the density plot of the rainfall data is shown. It is observed that the data distribution is right-skewed, indicating that extreme rainfall events occur more frequently than low rainfall events.

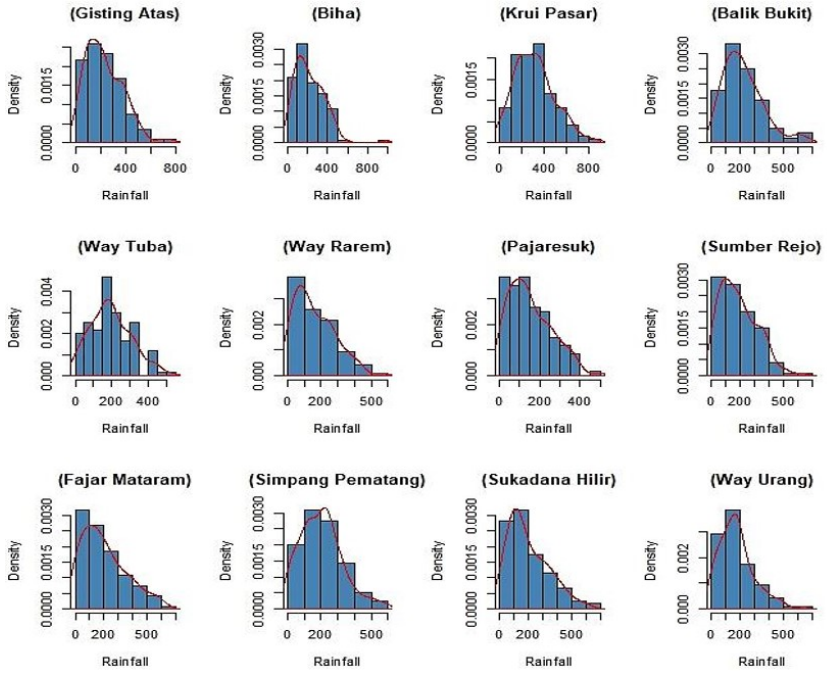

Figure 4. Histogram Plot of Data of 12 rainfall stations in Lampung Province.

In Figure 4 displays the histogram of the rainfall data. The histogram shows that the data has two main components : zero values representing (no rain) and continuous positive values indicating (occurrence of rainfall), which validates the use of a mixed distribution.

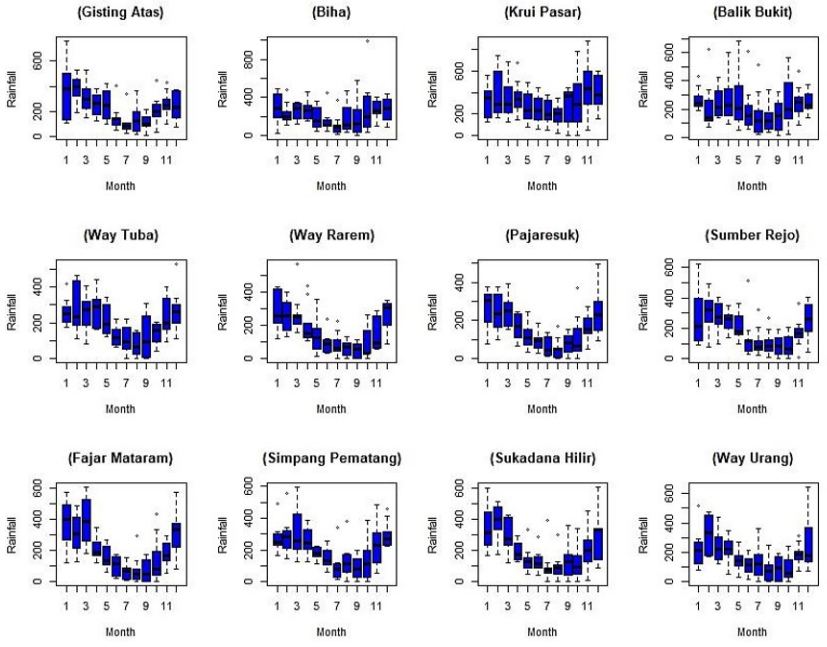

Figure 5. Boxplot of 12 rainfall stations in Lampung Province.

In Figure 5, the data visualization using Box-plot illustrates that the rainfall pattern at 12 stations shows a tendency to follow the monsoon rainfall pattern, which is characterized by a shape resembling the letter "U". This pattern indicates that rainfall intensity is at its lowest during the June to September period, which is generally the dry season. The results of this visualization support the mixed Tweedie distribution in the modeling, especially with the index parameter in the range $1 < \rho < 2$. The estimation of the value of the index parameter ρ is done using the profile likelihood method.

3.2. Tweedie Distribution and Index Parameter (p)

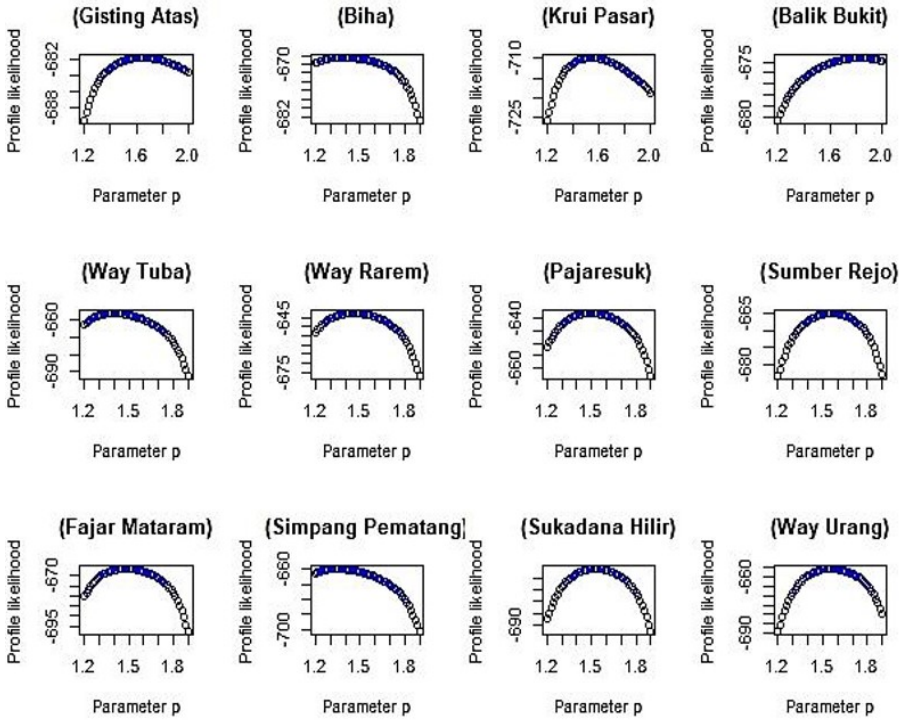

Figure 6. Profile Likelihood plot for the estimated index parameter p.

The Tweedie distribution was selected due to its ability to model zero-inflated data without *ad hoc* modifications, through its compound Poisson-Gamma structure [15]. Unlike Gamma or log-normal distributions, the Tweedie naturally handles both discrete (no rainfall) and continuous (rainfall amount) components. In Figure 6 shows the profile likelihood plot for the index parameter (p). The estimated value of p is in the range of 1.2 to 2., confirming that the rainfall data follows a mixed Tweedie distribution pattern (compound Poisson Gamma).

3.3. GLM with Tweedie Response

A Generalized Linear Model (GLM) with a Tweedie response was used to model rainfall as a function of the first eight principal components (PC1-PC8) derived from PCA. PCA addressed multi-collinearity among GCM grid variables, retaining the PCs with the highest standard deviations (Table 2 shows the standard deviantions for each of the first ten PCs). The GLM indicated a significant positive relationship between PC1 and rainfall ($p < 0.005$), suggesting that variance in rainfall explained by PC1 contributes positively to rainfall amounts.

Table 2. Standard deviation values for GCM predictor variables.

Principal Component	Standard Deviation
PC.1	11.25
PC.2	3.31
PC.3	3.16
PC.4	2.16
PC.5	1.78
PC.6	1.47
PC.7	1.22
PC.8	1.00
PC.9	0.90
PC.10	0.82

3.4. Rainfall Prediction

The trained GLM was used to predict monthly rainfall at 12 observation stations. Figure 7 illustrates predicted versus observed rainfall at Gisting Atas; overall model accuracy at this and other stations is summarized in Table 3.

Table 3 shows summarizes the comparison of model performance between the Tweedie-GLM-PCA and Normal-PCA models, showing that the Tweedie-GLM-PCA model consistently yielded prediction coser to the actual observed rainfall across all stations.

Table 3. Comparison of Actual Data & Predicted Data Prediction Results.

No	Rain Gauge Station	Actual	Tweedie Mixed-PCA	Normal-PCA
1	Gisting Atas	241.4	291.9	322.0
2	Biha	206.8	253.4	263.8
3	Krui Pasar	184.2	282.7	245.9
4	Balik Bukit	312.1	294.9	302.0
5	Way Tuba	208.3	231.1	208.4
6	Way Rerem	157.5	142.1	124.9
7	Pajaresuk	152.9	241.3	262.9
8	Sumber Rejo	148.3	133.7	116.0
9	Fajar Mataram	36.1	132.9	119.0
10	Simpang Pematang	192.7	142.6	125.0
11	Sukadana Hilir	165.0	283.6	279.0
12	Way Urang	198.6	324.5	349.1

Figure 7 shows that the model with the Tweedie-PCA distribution exhibits a consistent prediction pattern and closely matches the actual data, so the Tweedie-PCA method has better performance than the Normal-PCA model. The trend evident in Figure 7, with the Tweedie-PCA model exhibiting a closer fit to the actual data, is further substantiated by the data presented in Table 3. Specifically, Table 3 demonstrates that the RMSEP for the Gisting Atas station is lower for the Tweedie-PCA model than for the Normal-PCA model.

In Figure 7, it can be seen that the model with the Tweedie mixture-PCA distribution shows a consistent prediction pattern and is close to the actual data, so overall the Tweedie-PCA method has a better ability than the Normal-PCA model so that the Tweedie mixture modeling is good enough to be used in modeling rainfall in Lampung Province.

3.5. Model Performance Evaluation

Model performance was evaluated using RMSEP and correlation. Table 4 shows the RMSEP and correlation values for the Tweedie-GLM-PCA model and the Normal-PCA model. The Tweedie-GLM-PCA model exhibits a lower RMSEP and higher correlation compared to the Normal-PCA model at

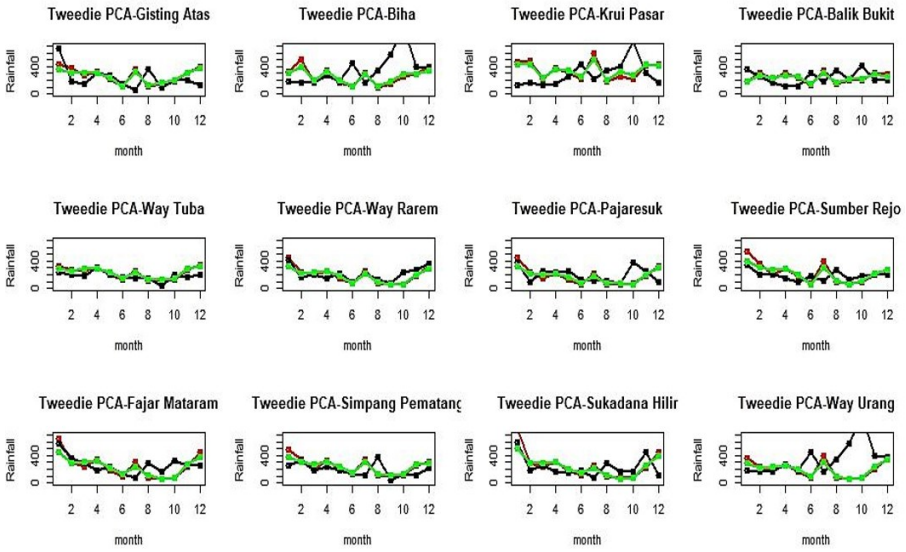

Figure 7. Plot of prediction and actual data for 2022 using Tweedie mixed-PCA method .

most stations. Overall, the Tweedie-GLM-PCA model demonstrates superior performance in capturing the rainfall patterns across the majority of stations in Lampung Province.

Table 4. Comparison of RMSEP and Correlation of twelve stations.

Rain Gauge Station	Tweedie-GLM-PCA	Normal-PCA
Gisting Atas	163.90 (0.06)	169.11 (0.17)
Biha	243.21 (0.83)	275.45 (-0.11)
Krui Pasar	103.73 (0.91)	241.61 (-0.44)
Balik Bukit	95.14 (0.37)	129.62 (-0.61)
Way Tuba	71.53 (0.38)	73.79 (0.61)
Way Rerem	103.10 (0.52)	75.82 (0.70)
Pajaresuk	99.73 (0.59)	123.78 (0.16)
Sumber Rejo	71.94 (0.22)	107.89 (0.28)
Fajar Mataram	129.54 (0.48)	125.76 (0.47)
Simpang Pematang	117.46 (-0.13)	118.16 (0.25)
Sukadana Hilir	141.34 (0.35)	142.31 (0.45)
Way Urang	264.01 (0.76)	337.52 (-0.63)

4. Conclusions

This study demonstrates that the Tweedie-GLM-PCA model provides a more effective approach for rainfall prediction in Lampung Province compared to the Normal-PCA model. By effectively accommodating both zero values and continuous positive values characteristic of rainfall data, the Tweedie-GLM-PCA model achieved lower RMSEP and higher correlation at most stations, indicating improved accuracy and reliability. This improved rainfall prediction has significant potential for supporting climate change adaptation efforts and enabling more optimal water resource management in Lampung Province. Future studies should explore further development suggestions for subsequent research, namely x (GCM precipitation) longitude latitude divided based on regency.

References

[1] F. A. Gusmayanti, E. Evi, and J. Sudrajat, "Pengaruh perubahan curah hujan terhadap produktivitas padi sawah di kalimantan barat," *Jurnal Ilmu Lingkungan*, vol. 19, no. 2, pp. 237–246, 2021. https://doi.org/10.14710/jil.19.2.237-246.

[2] Dinas Ketahanan Pangan, Tanaman Pangan dan Hortikultura Provinsi Lampung, "Kontribusi komoditas unggulan lampung tingkatkan perekonomian nasional," 2023. https://dinastph.lampungprov.go.id/detail-post/kontribusi-komoditas-unggulan-lampung-tingkatkan-perekonomian-nasional.

[3] B. Gwambene, E. Liwenga, and M. C, "Climate change and variability impacts on agricultural production and food security for the smallholder farmers in rungwe, tanzania.," *Environmental management*, vol. 71, no. 1, pp. 3–14, 2023. https://doi.org/10.1007/s00267-022-01628-5.

[4] R. N. Rachmawati, I. Sungkawa, and A. Rahayu, "Extreme rainfall prediction using bayesian quantile regression in statistical downscaling modeling," *Procedia Computer Science*, vol. 157, pp. 406–413, 2019. https://doi.org/10.1016/j.procs.2019.08.232.

[5] A. A. Keller, K. L. Garner, N. Rao, E. Knipping, and J. Thomas, "Downscaling approaches of climate change projections for watershed modeling: Review of theoretical and practical considerations," *PLOS Water*, vol. 1, pp. 1–20, 09 2022. https://doi.org/10.1371/journal.pwat.0000046.

[6] Z. Wan, I. Lopez-Gomez, R. Carver, T. Schneider, J. Anderson, F. Sha, and L. Zepeda-Núñez, "Statistical downscaling via high-dimensional distribution matching with generative models," 12 2024. https://doi.org/10.48550/arXiv.2412.08079.

[7] N. C. Dzupire, P. Ngare, and L. Odongo, "A poisson-gamma model for zero inflated rainfall data," *Journal of Probability and Statistics*, vol. 2018, no. 1, p. 1012647, 2018. https://doi.org/10.1155/2018/1012647.

[8] K. Burak and A. Kashlak, "Bootstrapping generalized linear models to accommodate overdispersed count data," *Statistical Papers*, vol. 65, pp. 1–20, 03 2024. https://doi.org/10.1007/s00362-024-01534-4.

[9] R. Jiang, X. Zhan, and T. Wang, "A flexible zero-inflated poisson-gamma model with application to microbiome sequence count data," *Journal of the American Statistical Association*, vol. 118, no. 542, pp. 792–804, 2023. https://doi.org/10.1080/01621459.2022.2151447.

[10] M. Hayati and R. Permatasari, "Comparison of generalized linear model between gamma and tweedie compound response for rainfall prediction in lampung province," *Asian Journal of Probability and Statistics*, vol. 26, p. 41–49, Jan. 2024. https://doi.org/10.9734/ajpas/2024/v26i1583.

[11] R. L. Wilby and T. M. L. Wigley, "Precipitation predictors for downscaling: observed and general circulation model relationships," *International Journal of Climatology*, vol. 20, no. 6, pp. 641–661, 2000. https://doi.org/10.1002/(SICI)1097-0088(200005)20:6%3C641::AID-JOC501%3E3.0.CO;2-1.

[12] R. E. Benestad, D. Chen, A. Mezghani, L. Fan, and K. Parding, "On using principal components to represent stations in empirical–statistical downscaling," *Tellus A: Dynamic Meteorology and Oceanography*, vol. 67, no. 1, p. 28326, 2015. https://doi.org/10.3402/tellusa.v67.28326.

[13] H. F. Kaiser, "A second generation little jiffy," *Psychometrika*, vol. 35, no. 4, p. 401–415, 1970. https://doi.org/10.1007/BF02291817.

[14] M. M. Hasan and P. K. Dunn, "A simple poisson–gamma model for modelling rainfall occurrence and amount simultaneously," *Agricultural and Forest Meteorology*, vol. 150, no. 10, pp. 1319–1330, 2010. https://doi.org/10.1016/j.agrformet.2010.06.002.

[15] P. Dunn and G. Smyth, "Series evaluation of tweedie exponential dispersion densities," *Statistics and Computing*, vol. 15, pp. 267–280, 10 2005. https://doi.org/10.1007/s11222-005-4070-y.

[16] R. Altman, O. Harari, N. Moisseeva, and D. Steyn, "Statistical modelling of the annual rainfall pattern in guanacaste, costa rica," *Water*, vol. 15, p. 700, 02 2023. https://doi.org/10.3390/w15040700?urlappend=%3Futm_source%3Dresearchgate.

[17] P. K. Dunn, "Occurrence and quantity of precipitation can be modelled simultaneously," *International Journal of Climatology*, vol. 24, no. 10, pp. 1231–1239, 2004. https://doi.org/10.1002/joc.1063.

[18] F.-Y. Chen, S. S. Yang, and H.-C. Huang, "Modeling pandemic mortality risk and its application to mortality-linked security pricing," *Insurance: Mathematics and Economics*, vol. 106, pp. 341–363, 2022. https://doi.org/10.1016/j.insmatheco.2022.06.002.

[19] R. Ma and B. Jørgensen, "Nested generalized linear mixed models: An orthodox best linear unbiased predictor approach," *Journal of the Royal Statistical Society Series B*, vol. 69, pp. 625–641, 09 2007. https://doi.org/10.1111/j.1467-9868.2007.00603.x.

[20] M. Hayati, A. H. Wigena, A. Djuraidah, and A. Kurnia, "A new approach to statistical downscaling using tweedie compound poisson gamma response and lasso regularization," *Communications in Mathematical Biology and Neuroscience*, vol. 21, pp. 1–16, May 2021. https://doi.org/10.28919/cmbn/5936.

[21] T. Peng, Y. Wu, J. Zhao, C. Wang, J. Wang, and J. Cai, "Ultrasound prostate segmentation using adaptive selection principal curve and smooth mathematical model," *Journal of Digital Imaging*, vol. 36, pp. 947–963, 2023. https://doi.org/10.1007/s10278-023-00783-3.

[22] R. Dawson, "How significant is a boxplot outlier?," *Journal of Statistics Education*, vol. 19, no. 2, 2011. https://doi.org/10.1080/10691898.2011.11889610.

[23] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. 01 2019. https://doi.org/10.1201/9780203753736.

[24] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006. https://doi.org/10.1016/j.ijforecast.2006.03.001.