Journal of Agrometeorology

(A publication of Association of Agrometeorologists)

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 27 (4) : 487-493 (December - 2025)
https://doi.org/10.54386/jam.v27i4.3182
https://journal.agrimetassociation.org/index.php/jam

Research Paper

Bayesian Tweedie Compound Poisson Gamma (TCPG) modeling for statistical downscaling of rainfall in West Java, Indonesia

SITI ROHMAH ROHIMAH[1,2], ANIK DJURAIDAH[1*], MUHAMMAD NUR AIDI[1], and BAGUS SARTONO[1]

[1]*School of Data Science, Mathematics, and Informatics, IPB University, Bogor, West Java, Indonesia*
[2]*Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, Jakarta, Indonesia*
**Corresponding Author: anikdjuraidah@apps.ipb.ac.id*

ABSTRACT

Global climate models (GCM) are effective in representing climate processes at the global scale; however, they often exhibit biases and limited accuracy at the local scale. This limitation is particularly critical in monsoon-dominated regions such as West Java, where statistical downscaling (SD) provides an appropriate approach. This research aims to predict monthly rainfall in West Java using the Bayesian Tweedie Compound Poisson Gamma (TCPG) model with combined scenarios of bias correction and dummy variables. Bias correction used empirical quantile mapping (EQM) with CHIRPS data. Monthly rainfall as the response variable was modelled using a Bayesian TCPG regression, with parameter estimation performed through Bayesian Markov chain Monte Carlo (MCMC) using the Metropolis Hastings algorithm. The best model scenario was achieved using dummy variables without bias correction, with CNRM-ESM2-1 identified as the most effective Decadal Climate Prediction Project (DCPP) model. These findings enhance rainfall prediction accuracy in tropical monsoon regions and support agricultural and water resource planning in West Java.

***Keywords*:** Bayesian, Bias correction, Dummy variables, Rainfall, Statistical downscaling, Tweedie Compound Poisson Gamma (TCPG)

Rainfall prediction is highly uncertain due to complex topography and ocean–land–atmosphere interactions, requiring accurate local-scale forecasting supported by global circulation information from global climate model (GCM) outputs. Statistical downscaling (SD) enables local rainfall prediction from GCM data, but the outputs are high-dimensional and suffer from multicollinearity. To address this, previous studies applied methods such as Least Absolute Shrinkage and Selection Operator (LASSO), principal component analysis (Soleh *et al.,* 2015; Yunus *et al.,* 2020), principal component regression, and latent root regression (Sahriman and Yulianti, 2023).

Accurate rainfall prediction requires models that can simultaneously capture all rainfall components. Rainfall data, often zero-inflated and right-skewed, are well-described by the Tweedie Compound Poisson Gamma (TCPG) distribution. TCPG-based modeling has been applied using various methods: maximum likelihood (Dunn, 2004; Hasan and Dunn, 2010; Yunus *et al.,* 2017; Dzupire *et al.,* 2018), blockwise majorization iteratively reweighted least square (Qian *et al.,* 2016; Dewanti *et al.,* 2024), quasi and pseudolikelihood (Bonat and Kokonendji, 2017), and iteratively reweighted least squares (Hayati *et al.,* 2021). TCPG models improve prediction accuracy for rainfall intensity, mean occurrence, and probability of no rainfall (Yunus *et al.,* 2017). Given the spatial-temporal complexity and high-dimensional predictors in statistical downscaling, a Bayesian approach is preferred for its flexibility and ability to incorporate prior information.

Previous studies have addressed bias correction (Shweta *et al.,* 2020; Dewanti *et al.,* 2024) and the use of dummy variables (Annisa *et al.,* 2023) separately. Previous studies have not integrated bias correction and dummy variables within the Bayesian TCPG framework, especially for tropical monsoon regions such as West Java. To fill this gap, the present study combines bias correction with dummy variables to improve local rainfall prediction. In general, West Java has four tropical seasons: the rainy season, the transition from rainy to dry season, the dry season, and the transition from dry to wet season. Therefore, to improve prediction accuracy,

***Article info - DOI*:** *https://doi.org/10.54386/jam.v27i4.3182*
Received:7 September 2025; Accepted: 7 October 2025; Published online : 1 December 2025

Table 1: Description of rainfall stations used in the study

Station	Coordinates	Altitude (m a.s.l)	Monthly rainfall (mm/ month)	Type of terrain
Cibukamanah	-6.57° S, 107.53° E	122	221.9	Lowland
Krangkeng	-6.50° S, 108.48° E	19	110.3	Lowland
Kawali	-7.19° S, 108.37° E	380	253.5	Midland
Katulampa	-6.60° S,106.80° E	262	339.6	Midland
Cibeureum	-7.04° S, 107.50° E	797	175.6	Highland
Perk. Gunung Mas	-6.71° S, 106.97° E	1130	303.0	Highland

in this study, the rainfall was divided into four groups, with three dummy variables added as explanatory variables. The novelty lies in combining TCPG with Empirical Quantile Mapping (EQM) bias correction and dummy variables to capture monsoon seasonality, thereby improving local scale rainfall prediction in West Java. In addition, improving the accuracy of rainfall prediction is closely related to the sustainable development goals (SDGs), particularly SDGs 13 on climate action.

MATERIALS AND METHODS

Study location

West Java Province of Indonesia has a varied topography and geologically rich. This region has combination of mountains, hills, highlands, midlands, lowlands, and coastlines. This study used six rainfall stations in West Java: Krangkeng and Cibukamanah (lowlands), Kawali and Katulampa (midlands), and Cibeureum and Gunung Mas (highlands). The detailed characteristics of the six stations are shown in Table 1.

Data

This research uses three types of secondary data from January 1991 to December 2020. The first is GCM couple model intercomparison project phase 6 (CMIP6) monthly rainfall data, a predictor variable in a 5×8 grid. The type of GCM used is decadal climate prediction project *(*DCPP) model, obtained from the page https://esgf-node.ipsl.upmc.fr/search/cmip6-ipsl/. The DCPP models used CNRM-ESM2, MIROC6, MRI-ESM2-0, and MPI-ESM1-2-HR (Dewanti *et al.,* 2024; Sativa *et al.,* 2025). The second is climate hazards group infrared precipitation with stations (CHIRPS) monthly rainfall data obtained from the page https://iridl.ldeo.columbia.edu/SOURCES/.UCSB/.CHIRPS/. CHIRPS locations correspond to six rain stations with a grid size of 20 x 20 and a resolution of 0.05°x 0.05° each. This data will be used to correct bias in the GCM output data. The third is monthly rainfall intensity data (mm/month) from the Indonesia Agency for Meteorology Climatology and Geophysics (BMKG).

Empirical quantile mapping

Empirical Quantile Mapping (EQM) is a bias correction technique that aligns the quantile distribution of model outputs with observations by mapping model values to the corresponding observational quantiles. In addition, EQM does not assume any specific distribution for precipitation and method capable of effectively reducing biases in mean, variance, quantiles, and wet day frequency. Bias correction via EQM involves (1) Calculating empirical percentiles for both predicted and observed data; (2) Obtain the cumulative distribution function for each predicted and observed data from empirical percentiles. The values that fall between the given percentiles are computed through linear interpolation; (3) Perform bias correction with the following equation (Gudmundsson *et al.,* 2012):

$$P_{cpd} = F_o^{-1}\big(F_d(P_d)\big) \tag{1}$$

where F_d is prediction data, is CDF of P_d, F_o^{-1} , CDF inverse of *Po, Po*, is observation data and P_{cpd} is corrected prediction data. Bias correction aims to establish the relationship between predicted data and actual data using a transfer function F_o^{-1} *(Fd (.))*.

Bayesian TCPG regression model

The TCPG distribution belongs to the exponential family and is characterized by two parameters: the mean μ dan dispersion parameter $\phi > 0$ denoted as $Tw_p(\mu,\phi)$, where the power index parameter p lies in the range $1<p<2$. According to Bonat and Kokonendji (2017), the link function for the response variable with TCPG distribution is the logarithmic link function, namely Suppose $\{(y_i,x_i),i=1,\dots,n\}$ is a pair of response variables and explanatory variables where stochastically free identical to the sample size n and $Y_i \sim T(W_p)\ (\mu_i,\phi)$. The TCPG based generalized linear model (GLM) with dummy variables can be expressed as:

$$log(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji} + \gamma_1 D_{1i} + \gamma_2 D_{2i} + \gamma_3 D_{3i} \tag{2}$$

where X_{ji} represents the j-th principal component obtained from the PCA of DCPP predictors for the i-th observation, and β are the regression coefficients for the GCM predictors, and γ are the regression coefficients corresponding to the three seasonal dummy variables compared to the wet season. Specifically, D_{1i} denotes the dummy for the transition from dry to wet season, D_{2i} for the wet season, and D_{3i} for the transition from wet to dry season. $(Y_i)=\mu_i$ and Var $(Y_i) = \phi\mu_i^p$. Parameter estimation for this model was conducted using the Markov Chain Monte Carlo (MCMC) with the Metropolis Hastings algorithm. The prior structure was adopted by Zhang (2013), specified as: $N\ \beta \sim N(0,\sigma_\beta^2 I) = \phi \sim U(0, 100)$, and $p \sim U\ (1,2)$.

Analysis procedure

The data analysis procedures carried out in this research

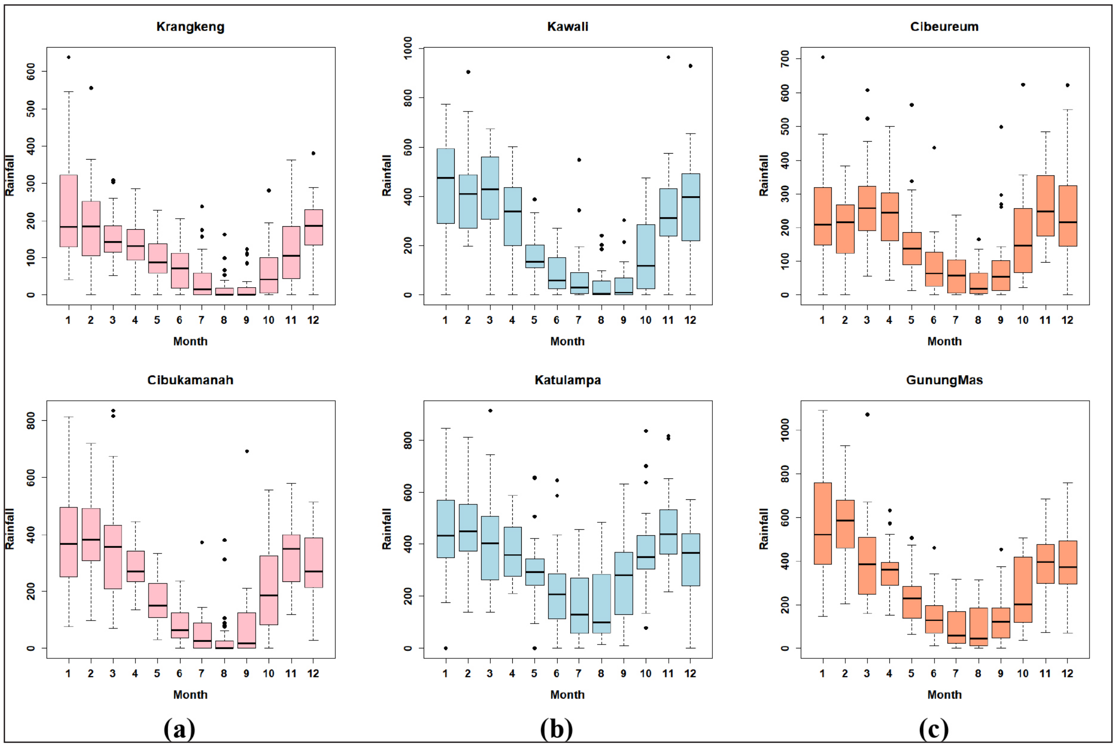

Fig. 1: Boxplot of rainfall at six rainfall stations in West Java Province in 1991-2020; (a) lowlands, (b) midlands, and (c) highlands

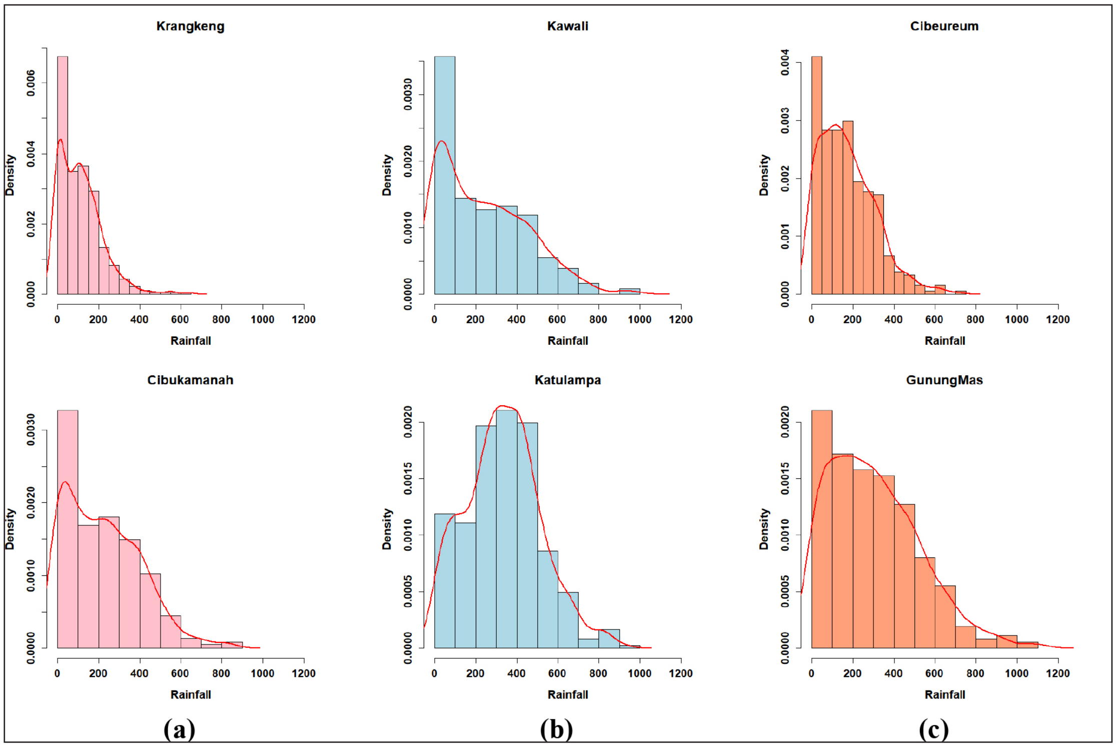

Fig. 2: Histogram and density plot of rainfall at six rainfall stations in West Java Province in 1991-2020; (a) lowlands, (b) midlands, and (c) highlands

are as follows: (1) Conducting data exploration to determine the characteristics of rainfall data; (2) Splitting the data into training sets (1991–2017) and testing sets (2018-2020); (3) Performing DCPP data bias correction on CHIRPS data used EQM; (4) Checking the distribution of rainfall data; (5) Reducing multicollinearity used principal component analysis (PCA); (6) Doing SD rainfall modelling with Bayesian TCPG regression used four scenarios (without bias correction used with and without dummy; and with bias correction used with and without dummy) for each DCPP and stations; (7) Evaluating the model by calculating

Table 2: RMSEP and correlation value of each station based on model scenario

Station	RMSEP				Correlation			
	Without bias correction		Bias correction		Without bias correction		Bias correction	
	Without dummy	With dummy	Without dummy	With dummy	Without dummy	With dummy	Without dummy	With dummy
Cibukamanah	126.29	109.37	134.30	120.69	0.61	0.77	0.61	0.71
Krangkeng	72.36	73.63	71.00	73.74	0.57	0.61	0.59	0.61
Katulampa	140.90	131.61	139.22	131.93	0.55	0.64	0.57	0.63
Kawali	180.28	162.67	181.54	170.4	0.69	0.76	0.67	0.72
Cibeureum	88.28	81.96	89.19	87.01	0.58	0.67	0.57	0.64
Perk. Gunung Mas	142.16	123.66	140.51	123.28	0.70	0.80	0.69	0.80
Average	125.05	113.82	125.96	117.84	0.62	0.71	0.62	0.69
Standard deviation	39.30	33.01	39.80	34.33	0.06	0.08	0.05	0.07

Table 3: RMSEP and correlation value of each DCPP based on model scenario

DCPP	RMSEP				Correlation			
	Without bias correction		Bias correction		Without bias correction		Bias correction	
	Without dummy	With dummy	Without dummy	With dummy	Without dummy	With dummy	Without dummy	With dummy
CNRM-ESM2-1	112.27	109.39	111.91	108.57	0.72	0.74	0.73	0.74
MIROC6	140.70	118.60	141.50	119.50	0.48	0.66	0.47	0.66
MRI-ESM2-0	118.76	112.83	131.05	132.10	0.69	0.72	0.60	0.62
MPI-ESM1-2-HR	128.46	114.44	119.39	111.19	0.60	0.70	0.66	0.72
Average	125.05	113.82	125.96	117.84	0.62	0.71	0.62	0.69
Standard deviation	12.38	3.82	13.01	10.58	0.11	0.03	0.11	0.06

$$\text{RMSEP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

and correlation coefficient (r) between actual data and predicted data with formula and;

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}};$$

(8) Comparing models for four scenarios used root mean square error of prediction (RMSEP) and correlation coefficient.

RESULTS AND DISCUSSION

Exploration of rainfall data

Rainfall at six stations distributed across West Java Province exhibited a monsoonal pattern. Rainfall is lowest from July to September (<100 mm/month), representing the dry season. Rainfall rose from October to March, peaking in the wet season (>200 mm/month), then declined from April to June as the transition to the dry season. The wet–dry month classification followed Tasiyah *et al.,* (2024) and was used to construct dummy categories. The average rainfall pattern of six stations is shown in Fig. 1.

Rainfall at Cibukamanah and Krangkeng tends to be low, below 200 mm/month with rare extreme events, while Kawali, Cibeureum, and Gunung Mas have long tails of distribution, indicating a higher frequency of extreme rainfall and greater variation. Katulampa Station showed more dispersed distribution with several small peaks, reflecting a varied rainfall pattern. In general, all locations showed a positive asymmetric distribution with most rainfall in the range of 0-300 mm/month and a right-tailed tail, supporting the use of the TCPG distribution. The rainfall distribution is illustrated in Fig. 2.

Comparison of model performance for each station and DCPP

The Bayesian TCPG regression model showed different predictive performance across stations. Table 2 presents the RMSEP values and correlation coefficients for each station under different model scenarios, namely with and without bias correction as well as with and without dummy variables. On average, the inclusion of dummy variables reduced RMSEP and increased correlation compared to scenarios without dummy variables.

Dummy variables notably improved model accuracy, while bias correction had only minor effects. The highest correlation (0.8) occurred at Perkebunan Gunung Mas with dummy variables, the lowest RMSEP (71) at Krangkeng with bias correction without dummy, and higher RMSEP (>130) at Kawali and Katulampa despite moderate correlations (0.63–0.76), reflecting complex rainfall patterns.

The CNRM-ESM2-1 model showed the most consistent

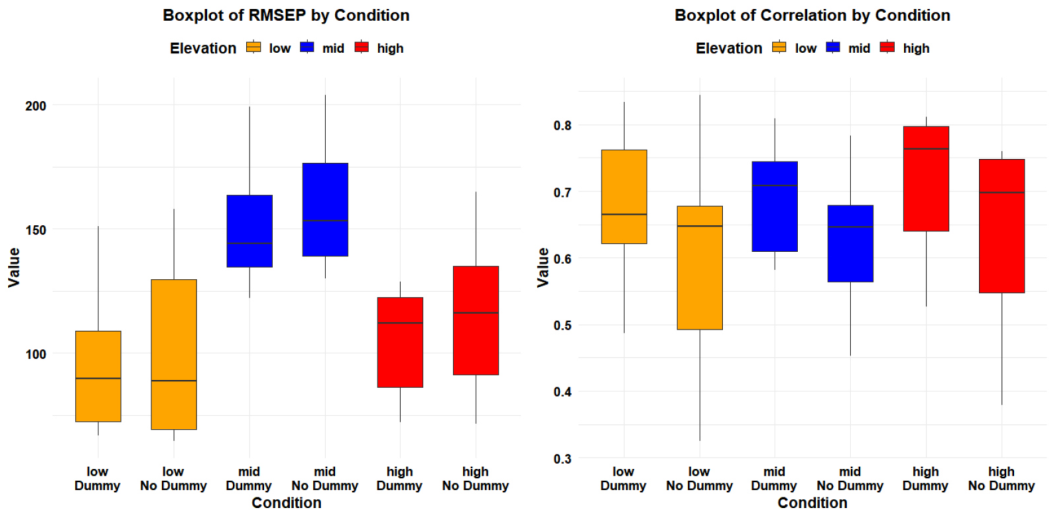

Fig. 3: Boxplot of RMSEP and correlation value for dummy and without dummy scenarios

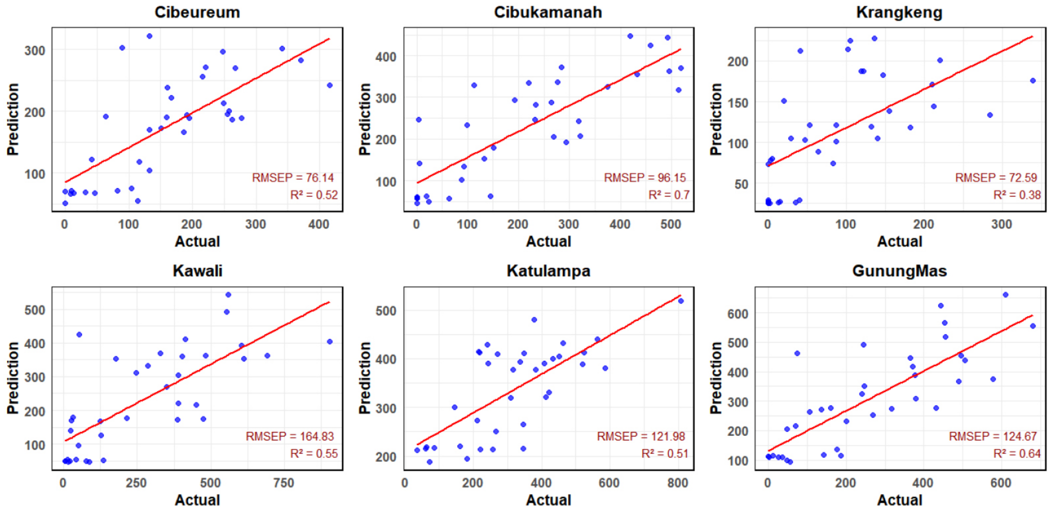

Fig. 4: Comparison plot of actual data with predicted data from the best scenario

Table 4: RMSEP and correlation value of terrain types

Terrain	RMSEP				Correlation			
	Without bias correction		Bias correction		Without bias correction		Bias correction	
	Without dummy	With dummy	Without dummy	With dummy	Without dummy	With dummy	Without dummy	With dummy
Lowland	99.33	91.50	102.65	97.22	0.59	0.69	0.60	0.66
Midland	160.59	147.14	160.38	151.17	0.56	0.70	0.62	0.68
Highland	115.22	102.81	128.71	105.15	0.64	0.74	0.63	0.72

performance with lower prediction errors and higher correlations, whereas MIROC6 demonstrated the weakest performance after bias correction. The use of dummy variables tends to help maintain correlation stability compared to models without dummy variables. Overall, the evaluation results based on Table 3 indicated that bias correction did not necessarily improve model performance. This occurred because the model was strongly affected by seasonal variation (dummy variables).

The addition of dummy variables improved model accuracy, both with and without bias correction. Because GCM predictors are continuous and often overlap between seasons, they cannot explicitly capture categorical seasonal effects. Dummy variables allowed the model to represent seasonal heterogeneity more precisely, providing advantages in identifying distinct seasonal influences.

The application of bias correction actually tended to degrade model performance, given the data and methods used. The Bayesian approach naturally corrected for bias through resampling to obtain the posterior distribution. The estimation process already incorporated natural bias correction through the posterior distribution, making frequentist bias correction irrelevant.

Comparison of Model Performance between Terrain Types

The predictive performance of the Bayesian TCPG regression model varied across terrains. The best model performance was observed in the lowland and highland areas when dummy variables were included without bias correction, with RMSEP values of 91.5 and 102.81, respectively. This is presented in Table 4.

According to Table 4, the model was capable of predicting rainfall in lowland and highland areas with relatively small errors. In contrast, the midland area showed the lowest performance, with the highest RMSEP value of 160.59. This was consistent with the research by Dewanti *et al.,* (2024), which explained that in the Stacking-RF ensemble model, bias correction did not improve model performance either overall or by terrain types. The limitation of Dewanti *et al.,* (2024) was that they did not include dummy variables as predictors.

The inclusion of dummy variables as a novelty in this study improved the model's performance in predicting rainfall across different elevations. This finding was consistent with the study by Annisa *et al.,* (2023), which reported that the inclusion of dummy variables in the model improved the estimation of rainfall data. Fig. 3 indicates that dummy variables reduced RMSEP and increased correlation, especially in lowland and highland areas.

Best model performance

Overall, that the model performed best in each station and terrain used CNRM-ESM2-1 with dummy and without bias correction. The model showed varying performance across stations, with the best results at Cibukamanah ($R^2 = 0.70$, RMSEP = 96.15) and Gunung Mas ($R^2 = 0.64$, RMSEP = 124.67), while the lowest accuracy was observed at Krangkeng ($R^2 = 0.38$, RMSEP = 72.59). Overall, the model adequately captured rainfall variability, although prediction accuracy differed among stations. This is illustrated in Fig. 4.

Based on the best model performance was obtained the index parameter p at all stations consistently ranged between $1 < p < 2$ (Katulampa: 1.317; Krangkeng: 1.357; Perkebunan Gunung Mas: 1.403; Cibukamanah: 1.384; Cibeureum: 1.438; Kawali: 1.471). This confirmed that rainfall data followed a compound Poisson process with a Gamma component, consistent with the TCPG distribution and supporting previous findings (Dunn, 2004; Dzupire *et al.,* 2018; Hayati *et al.,* 2021).

CONCLUSION

Bayesian TCPG regression with principal component analysis effectively handles rainfall data and multicollinearity, with the best results achieved using the DCPP CNRM-ESM2-1 model with dummy variables and without bias correction. This indicated that the model has a high capability in capturing local and seasonal climate dynamics in West Java, Indonesia. This highlights the importance of grouping through dummy variables in enhancing model performance. This model performance was reflected at almost all stations. In addition, the model performed best in lowland and highland areas. From a practical perspective, this approach can support agricultural management, water resource planning, and disaster mitigation. For future research, the application of hierarchical Bayesian structures is recommended to better capture spatial variability and enhance model robustness.

ACKNOWLEDGEMENT

The authors would like to thank the Indonesia Agency for Meteorology Climatology and Geophysics (BMKG) and the University of California and the Earth System Grid Federation (ESGF) website for providing the necessary data and the IPB University for supporting the research.

Conflict of Interests: The authors declare that there is no conflict of interest related to this article.

Funding: This research was funded by the Saudi Fund for Development (SFD) Batch 2 Universitas Negeri Jakarta Year 2020.

Data availability: DCPP and CHIRPS can be obtained from the link provided. Rainfall data is available upon request.

Authors contribution: **S. R. Rohimah*:*** Data collection, Analysis, Interpretation, Writing-original draft and Review; **A. Djuraidah*:*** Conceptualization, Methodology, Review, Supervision; **M. N. Aidi*:*** Conceptualization, Methodology, Supervision; **B. Sartono*:*** Conceptualization, Methodology, Supervision

***Disclaimer*:** The contents, opinions, and views expressed in the research article published in the Journal of Agrometeorology are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

***Publisher's Note*:** The periodical remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

Annisa, F., Raupong, and Sahriman S. (2023). Performa Model Statistical Downscaling dengan Peubah Dummy Berdasarkan K-Means dan Average Linkage. ESTIMASI: *J. Stat. Appl.,* 4(2): 165-175. https:// doi.org/10.20956/ejsa.v4i2.12658

Bonat, W. H., and Kokonendji, C. C. (2017). Flexible Tweedie regression models for continuous data. *J. Stat. Comput. Simul.*, 87(11): 2138-2152. https://doi.org/10.1080/00949655.2017.1318876

Dewanti, D., Djuraidah, A., Sartono, B., and Sopaheluwakan, A.

(2024). Bias correction and ensemble techniques in statistical downscaling model for rainfall prediction using Tweedie-LASSO in West Java, Indonesia. *J. Agrometeorol.,* 26(3): 324-330. https://doi.org/10.54386/jam.v26i3.2614

Dunn, P. K. (2004). Occurrence and quantity of precipitation can be modelled simultaneously. *Int. J. Climatol.,* 24(10): 1231-1239. https://doi.org/10.1002/joc.1063

Dzupire, N. C., Ngare, P., and Odongo, L. (2018). A Poisson-Gamma model for zero inflated rainfall data. *J. Probab. Stat.,* (1012647): 1-12. https://doi.org/10.1155/2018/1012647

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Skaugen, T. E. (2012). Technical note: downscaling RCM precipitation to the station scale using quantile mapping - a comparison of methods. *Hydrol. Earth Syst. Sci. Discuss.,* 9: 6185-6201. https://doi.org/10.5194/hessd-9-6185-2012

Hasan, M.M. and Dunn, P.K. (2010). A simple Poisson-gamma model for modelling rainfall occurrence and amount simultaneously. *Agric. Forest Meteorol.,* 150(10): 1319-1330.

Hayati, M., Wigena, A. H., Djuraidah, A., and Kurnia, A. (2021). A new approach to statistical downscaling using Tweedie compound Poisson Gamma respone and lasso regularization. *Commun. Math. Biol. Neurosci.,* 2021(60): 1-16. https://doi.org/10.28919/cmbn/5936

Qian, W., Yang, Y., and Zou, H. (2016). Tweedie's compound Poisson model with grouped elastic net. *J. Comput. Graph. Stat.*, 25(2): 606-625. https://doi.org/10.1080/10618600.2015.1005213

Sahriman, S. and Yulianti, AS. (2023). Statistical Downscaling Model with Principal Component Regression and Latent Root Regression to Forecast Rainfall in Pangkep Regency. *BAREKENG: J. Math. App.,* 17(1): 401-410.

Sativa, O., Djuraidah, A., and Notodiputro, K.A. (2025). Ensemble Methods in Statistical Downscaling with Gamma-LASSO Regression for Rainfall Prediction in West Java. *J. Math., Comp. Stat.,* 8(1): 244-258.

Shweta Panjwani, S. Naresh Kumar, and Laxmi Ahuja. (2020). Simulation performance of selected global and regional climate models for temperature and rainfall in some locations in India. *J. Agrometeorol.,* 22(4): 407-418. https://doi.org/10.54386/jam.v22i4.443

Soleh, A.M., Wigena, A.H., Djuraidah, A., and Saefuddin, A. (2015). Statistical downscaling to predict monthly rainfall using generalized linier model with gamma distribution. *Informatika Pertanian,* 24(2): 215-222.

Tasiyah, L.A., Sutriono, R., and, Silawibawa, I.P. (2024). Analisis Tipe Iklim Berdasarkan Curah Hujan Pada Beberapa Kecamatan di Kabupaten Lombok Barat. *J. Soil Qual. Manag.,* 1(1): 67-72.

Yunus, M., Saefuddin, A., and Soleh, A.M. (2020). Pemodelan Statistical Downscaling dengan LASSO dan Group LASSO untuk Pendugaan Curah Hujan. Indonesian *J. Stat. Appl.,* 4(4): 649-660.

Yunus, R.M., Hasan, M.M., and Zubairi, Y.Z. (2017). Fitting monthly Peninsula Malaysian rainfall using Tweedie distribution. *J. Phys. Conf. Ser.,* 890 (1). https://doi.org/10.1088/1742-6596/890/1/012164

Zhang, Y. (2013). Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. *Stat Comp.,* 23(6):743-757. https://doi.org/10.1007/s11222-012-9343-7.