



Daily Rainfall-State Forecasting in Medan Using a First-Order Markov Chain

Septhia Eka Nurviranthi Panjaitan^{1*}

¹Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sumatera Utara, Medan, North Sumatra, Indonesia, nurviranthi03@gmail.com

ABSTRACT

This study examines daily rainfall-state transitions in Medan using a first-order Markov chain fitted to historical rainfall observations from the Belawan station of Indonesia's Meteorology, Climatology, and Geophysics Agency. The observation period extends from 1 January 2020 to 31 December 2024. Rainfall amounts were assigned to three states: zero rainfall, 0.1–0.99 mm, and at least 1 mm. These categories are rainfall-based proxies and do not independently describe cloud cover. Of 1,827 calendar days, 1,725 were classified and 102 records marked 8888 were excluded. The reported transition matrix indicates persistence probabilities of 0.600, 0.172, and 0.487 for the three states, respectively. The stationary distribution is approximately 0.477, 0.138, and 0.385. In the seven-day illustration, the zero-rainfall state has the largest forecast probability at every horizon, although its probability declines toward 0.477. Comparison with six valid observations yields three correct classifications, equivalent to 50% accuracy; all three observed days with rainfall of at least 1 mm are missed. The model provides an interpretable local example of rainfall-state dynamics, but this short evaluation does not establish operational forecast skill. Reproducible transition estimation, confirmed temporal separation of estimation and evaluation data, and longer benchmark comparisons are needed before practical forecasting use.

Keywords: rainfall states; Markov chain; transition probabilities; stationary distribution; Medan

Received: 2 January 2026

Revised: 4 February 2026

Published: 30 March 2026

INTRODUCTION

Daily rainfall records can be used to describe how wet and dry conditions follow one another. Markov models summarize this dependence through transition probabilities, making their assumptions and outputs relatively easy to examine. Stochastic weather generation has a long history, including joint simulation of precipitation and other meteorological variables (Richardson, 1981). However, reproducing rainfall statistics and predicting the state of a particular future day are different objectives. Work on seasonal totals and long-term variability shows why performance must be assessed at the temporal scale relevant to the application (Chowdhury et al., 2017; Gregory et al., 1993).

Indonesian applications include rainfall modeling in West Java, simulation at Kertajati, and daily rainfall-state analysis in Bogor and Karo (Azizah et al., 2019; Fikri & Mutaqin, 2022; Hardianti, 2024; Wahyudi et al., 2021). Studies in Makassar and at Djalaluddin station also use Markov models, but their monthly classifications are not directly interchangeable with daily rainfall categories (Ihsan et al., 2020; Nasib et al., 2022). Hybrid fuzzy time-series methods have been applied to rainfall forecasting for rice-cropping schedules (Laily et al., 2023). A comparison of fuzzy time-series methods in Medan further indicates that the performance of a Markov-based method must be assessed against alternatives rather than presumed from its mathematical structure (Arnita et al., 2020).

International studies address daily rainfall transitions in Paraíba, Brazil, and rainfall variability in Bangladesh (Chowdhury et al., 2019; Jale et al., 2019). More elaborate models couple occurrence processes with rainfall events, represent dependence among multiple sites, or use weather-radar fields to generate continuous rainfall sequences (Andersen et al., 2024; Gao et al., 2020; Gao et al., 2021). These developments provide context for a simple local model; they do not establish that the same level of performance can be achieved using a single pooled transition matrix.

This study describes a first-order Markov model based on daily rainfall observations from Belawan station during 2020–2024. It aims to characterize transitions among three rainfall-based states, interpret their stationary probabilities, and examine a seven-day forecast illustration against the supplied observations. Its contribution is a local, interpretable case study rather than a new forecasting method. The analysis distinguishes state persistence, long-run model behavior, and short-term classification performance, while identifying the information needed for a reproducible evaluation.

METHODS

The study uses secondary daily rainfall observations from the Belawan station of the Meteorology, Climatology, and Geophysics Agency (BMKG), covering 1 January 2020 to 31 December 2024. No human participants were involved. Figure 1 summarizes the reported workflow, from data preparation to comparison with observations. The matrix-validation stage refers to checking mathematical constraints; it does not by itself establish the first-order Markov assumption or predictive validity.

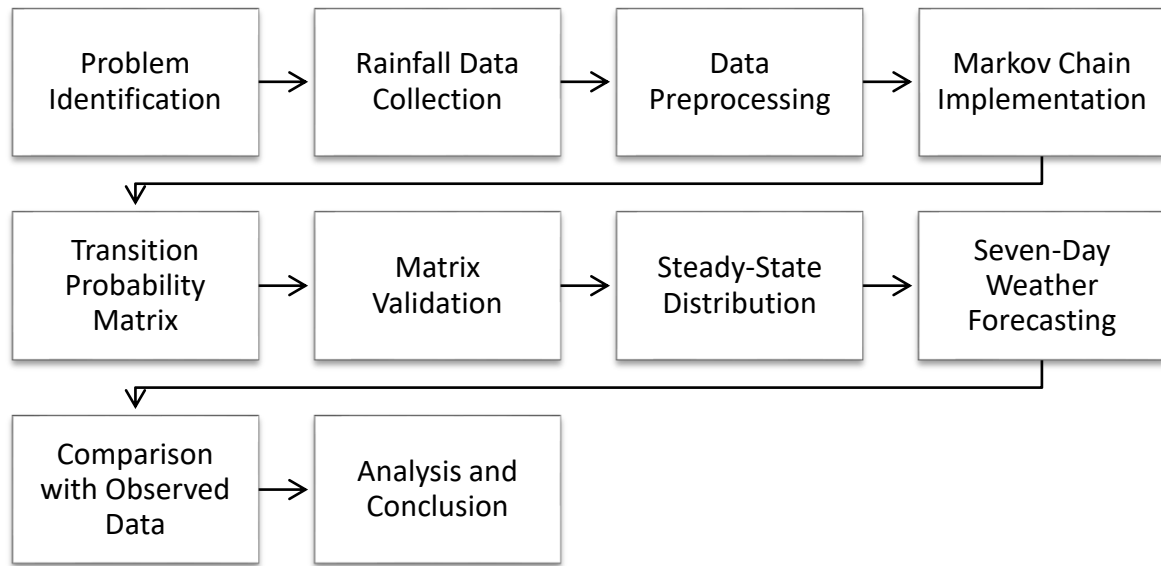


Figure 1. Research workflow

Rainfall was classified using the thresholds in Table 1. The original labels S (“sunny”), C (“cloudy”), and R (“rainy”) are retained to maintain correspondence with the figures. Throughout the analysis, they denote rainfall categories only: rainfall measurements alone do not establish sunshine duration or cloud cover. The thresholds are treated as study-specific definitions, since their status as an official station classification has not been established.

Table 1. Rainfall-based state definitions

Code	Original figure label	Daily rainfall	Interpretation
S	Sunny / Cerah	0 mm	Zero rainfall
C	Cloudy / Mendung	0.1–0.99 mm	Small rainfall amount
R	Rainy / Hujan	≥ 1 mm	Rainfall of at least 1 mm

Note. Original labels are retained for figure correspondence; S and C do not establish sunshine or cloud cover.

The reported data preparation excluded 102 records marked 8888, leaving 1,725 classified days. Excluded records must not be interpreted as zero rainfall. The treatment of missing observations also matters when estimating a daily transition matrix: two observations separated by an unmeasured day are not an observed one-day transition. The available results do not specify whether such gaps were retained when transition pairs were counted. Rainfall measurement precision is also unspecified; values between 0 and 0.1 mm or between 0.99 and 1 mm would require an explicit classification rule if they occur.

Let the observed state sequence be x and let F contain transition counts, with rows denoting the current state and columns the next state, both ordered S, C, R. For a first-order, time-homogeneous model, the next-state distribution depends on the current state, and the same transition probabilities are used throughout the modeled period. The state vector and count matrix are written as:

$$x = (x_1, x_2, \dots, x_n)^T, \quad F = (f_{ij}), \quad i, j = 1, 2, 3 \quad (1)$$

For a row with at least one observed transition, the estimated probability is the corresponding transition count divided by the total count in that row:

$$p_{ij} = \frac{f_{ij}}{\sum_{k=1}^3 f_{ik}}, \quad P = (p_{ij}) \quad (2)$$

Here, f_{ij} counts observed transitions from state i to state j , and the denominator sums over destination states k . Each unrounded row must contain nonnegative probabilities summing to one. This construction is consistent with occurrence-based rainfall models (Azizah et al., 2019; Gao et al., 2020), but satisfying these constraints is a structural check, not a statistical test of the assumed memory length or time homogeneity. The daily sequence and transition counts are not available with the reported results, so the estimated probabilities cannot be independently reconstructed from annual totals.

The stationary row vector π describes the model's limiting state proportions and satisfies:

$$\pi = \pi P, \quad \pi_S + \pi_C + \pi_R = 1 \quad (3)$$

Stationary distributions have also been used to characterize long-term drought-state probabilities (Azimi et al., 2020). Their interpretation depends on the fitted transition system and does not imply that observed climate conditions are constant. The reported matrix is rounded to three decimal places; its R row sums to 0.999. For a numerical consistency check only, each displayed row was divided by its row sum before solving for π . This rounding adjustment is not a replacement for estimating the matrix from the original daily records.

For a fixed forecast origin with initial row vector q_0 , the distribution at horizon h is:

$$q_h = q_0 P^h, \quad h = 1, \dots, 7 \quad (4)$$

The first forecast distribution in the supplied illustration is (0.600, 0.102, 0.298), corresponding to an initial S state, $q_0 = (1, 0, 0)$. The category with the largest probability is selected separately at each horizon. This fixed-origin procedure differs from rolling one-day forecasting, in which each new observed state is used to initialize the following forecast. Python in Google Colab was reported as the computational environment, but the underlying notebook is not available for reproducibility checks.

The evaluation compares the seven displayed category predictions with observations labeled 1–7 January. One observation is excluded, so accuracy is calculated as the number of correct predictions divided by six. The evaluation year and its relationship to the 2020–2024 estimation interval are unspecified; consequently, the comparison cannot be classified as an independently verified out-of-sample test. Classification accuracy describes the chosen labels, whereas probability forecasts require separate assessment, for example through proper scoring rules (Gneiting & Raftery, 2007).

RESULT AND DISCUSSION

Table 2 presents the reported annual counts. The 824 S days account for 47.8% of the 1,725 classified observations, compared with 238 C days (13.8%) and 663 R days (38.4%). The 102 excluded records represent approximately 5.6% of all 1,827 calendar days and are not included in these state percentages. The annual totals are consistent with leap years in

2020 and 2024. These counts describe state frequencies but do not determine transition probabilities without the ordered daily sequence.

Table 2. Reported daily rainfall-state counts, 2020–2024

State / record	2020	2021	2022	2023	2024	Total	Share valid
S	161	161	140	157	205	824	47.8%
C	38	46	58	64	32	238	13.8%
R	140	134	152	123	114	663	38.4%
Excluded (8888)	27	24	15	21	15	102	—
Calendar days	366	365	365	365	366	1827	—

Note. State percentages use 1,725 classified days. Excluded observations account for 102/1,827 calendar days (approximately 5.6%).

Table 3 shows the reported transition matrix, and Figure 2 presents the same values graphically. The S-to-S probability is 0.600, while the R-to-R and C-to-C probabilities are 0.487 and 0.172. Thus, S has the greatest next-day persistence among the defined states. For an observed C state, S is the most probable next state (0.429), slightly above R (0.399). For an observed R state, R remains the most probable next state. The matrix therefore does not imply that every rolling one-day prediction would be S.

Table 3. Reported one-day transition probabilities

Current state	Next S	Next C	Next R
S	0.600	0.102	0.298
C	0.429	0.172	0.399
R	0.342	0.170	0.487

Note. State order is S, C, R. Values are reported to three decimal places; the displayed R row sums to 0.999 because of rounding.

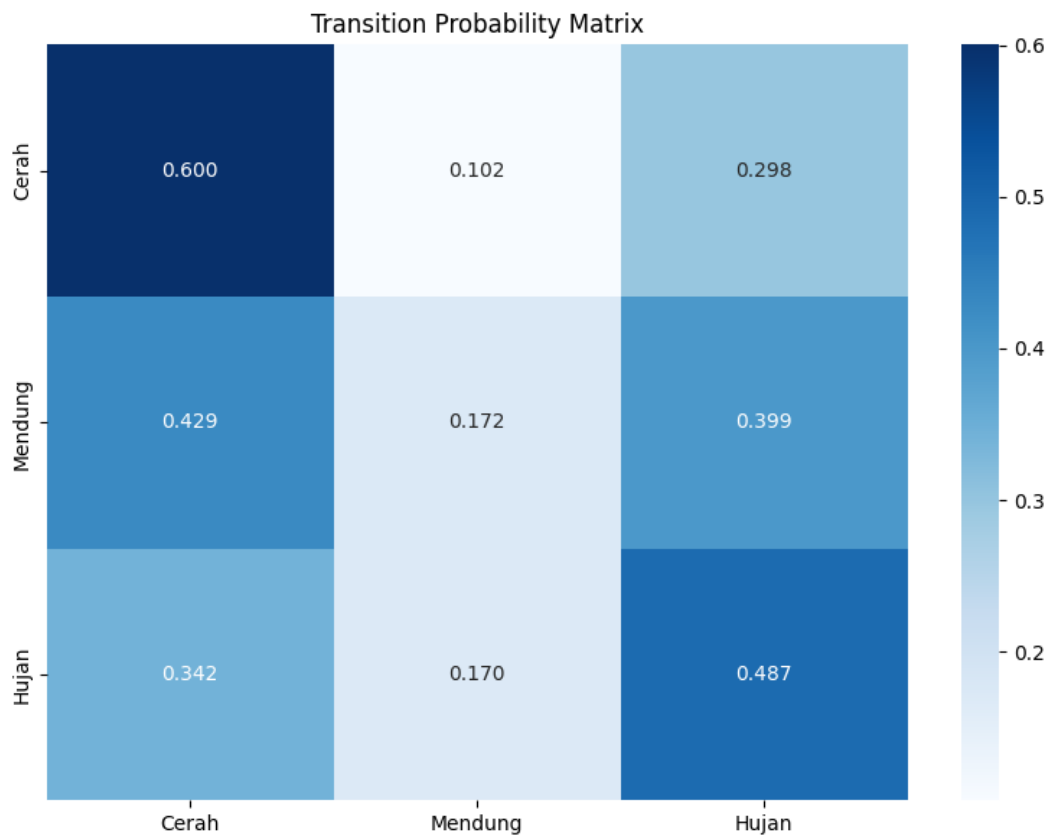


Figure 2. Reported one-day rainfall-state transition matrix

Note. Cerah = S; Mendung = C; Hujan = R. Rows denote the current state and columns the next state.

Figure 3 shows the stationary distribution rather than the diagonal persistence probabilities. The reported limiting proportions are approximately 0.477 for S, 0.138 for C, and 0.385 for R. Solving with the row-normalized displayed matrix gives approximately (0.47726, 0.13789, 0.38485), consistent with those rounded values. They are also close to the classified sample proportions in Table 2. This agreement is descriptive and is not independent evidence of forecast skill.

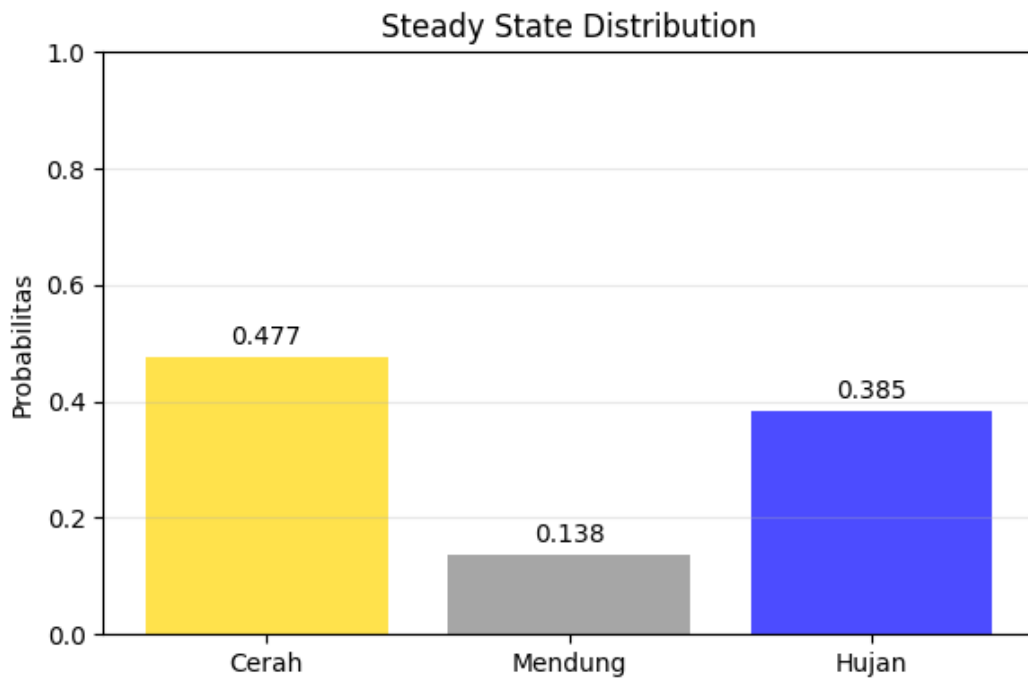


Figure 3. Stationary probabilities of the three rainfall states

Note. The bars represent long-run model probabilities, not next-day persistence.

Every entry of the reported matrix is positive. A corresponding stochastic matrix is therefore irreducible and aperiodic, with a unique stationary distribution toward which repeated multiplication converges. Figure 4 illustrates this behavior from an initial S state. The first-day probabilities are (0.600, 0.102, 0.298); by the seventh step, the rounding-adjusted consistency calculation gives approximately (0.477, 0.138, 0.385). The initial condition has little influence at that horizon under this fitted model.

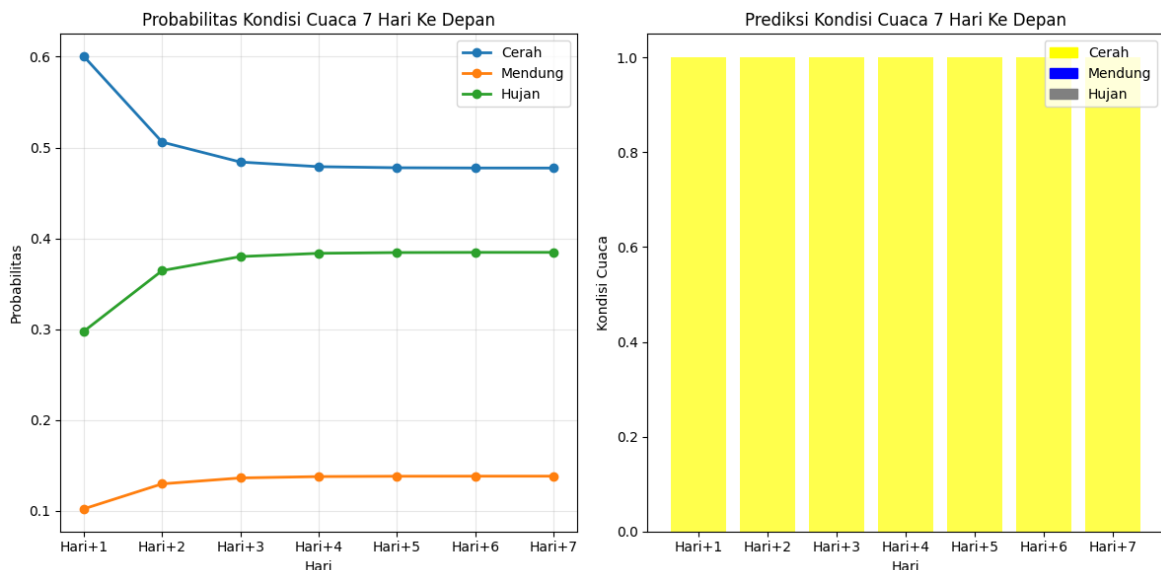


Figure 4. Seven-day forecast probabilities and selected categories

Note. The right panel shows categorical selections; bar height 1 does not indicate probability 1.

S remains the largest component of the forecast vector throughout the seven-day illustration, so all seven selected labels are S. This does not mean a 100% probability of zero

rainfall. Near the seventh horizon, the probability assigned to S is below one half, while the probability assigned to R is approximately 38.5%. The categorical panel in Figure 4 displays the selected labels; its bar height should not be read as the model’s probability of S.

Figure 5 compares these labels with the displayed observations. The observed states are S on 1, 2, and 4 January and R on 3, 6, and 7 January. The record for 5 January is excluded. Consequently, the valid evaluation contains three S days and three R days, with no C days. The difference between the all-S forecast and this mixed sequence shows the loss of information that occurs when probability vectors are reduced to a single category.

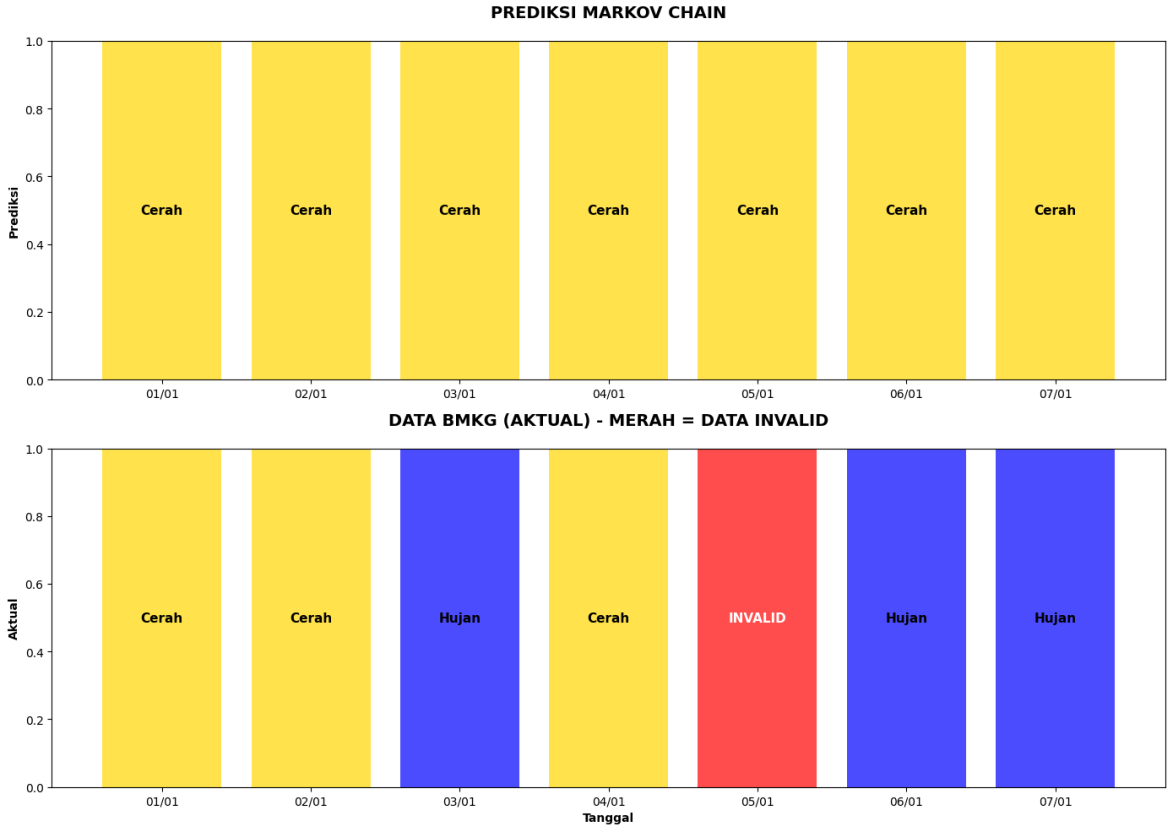


Figure 5. Forecast categories and observations for 1–7 January

Note. Yellow = S; blue = R; red = excluded record. The evaluation year is unspecified.

The confusion matrix in Figure 6 records three correctly classified S days and three R days incorrectly classified as S. Accuracy is therefore exactly 3/6, or 50%, for the displayed valid observations. S recall is 100%, while R recall is 0%; C performance cannot be evaluated because no C observations occur. A constant-S forecast would make the same predictions on this sequence. The comparison therefore does not demonstrate an improvement over that simple baseline.

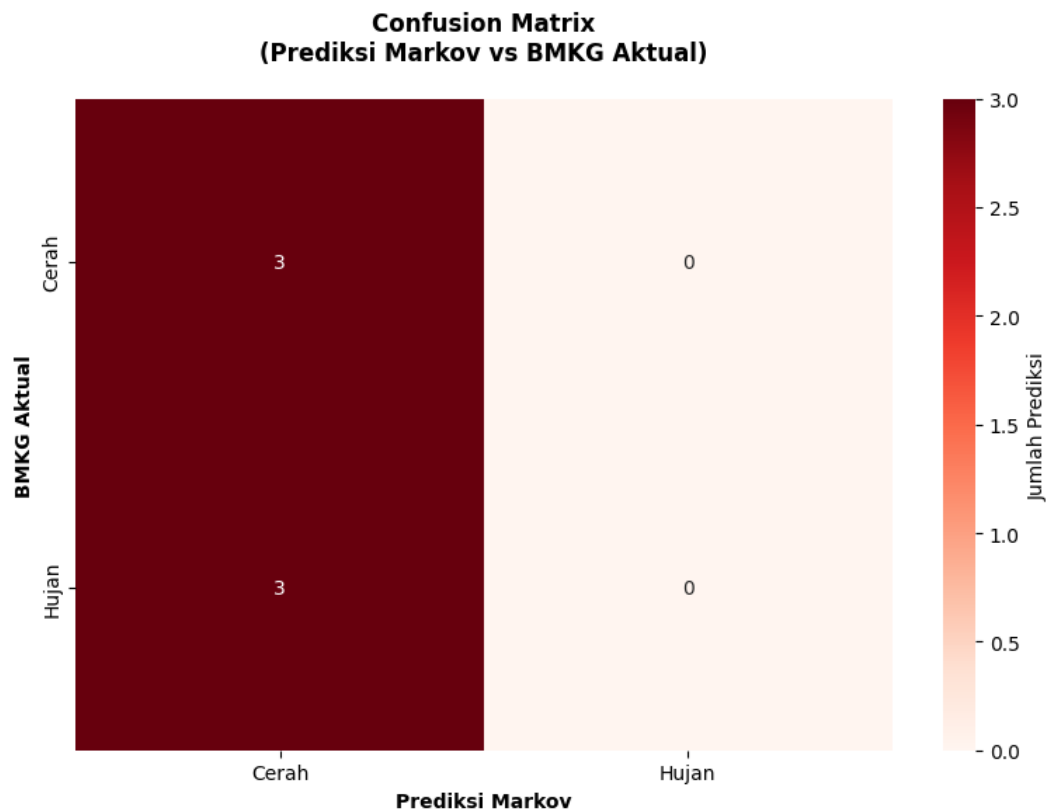


Figure 6. Confusion matrix for the six valid evaluation days

Note. Rows show observed states; columns show predicted states. No C observations occur.

These findings support a limited interpretation: the model summarizes reported rainfall-state persistence and explains how fixed-origin probabilities approach their stationary distribution. They do not establish reliable short-term weather forecasting. Research evaluating stochastic rainfall models for hydrological use likewise emphasizes that satisfactory rainfall statistics do not automatically imply satisfactory performance for another target variable or application (Nguyen et al., 2023). In the present case, agreement between stationary and empirical proportions cannot substitute for validation of daily forecasts.

Pooling all five years into a single matrix may conceal seasonal or interannual differences. Work on long-term rainfall variability and seasonal hidden Markov models provides a basis for examining time-varying alternatives (Chowdhury et al., 2017; Gobet et al., 2025). The Bogor and Karo applications also motivate explicit treatment of temporal structure and rainfall-state definitions (Hardianti, 2024; Wahyudi et al., 2021). Such extensions would need to be estimated and tested on the Belawan records; their benefits cannot be assumed from results in another location.

A stronger evaluation would preserve calendar-day adjacency during preprocessing, disclose transition counts and the unrounded matrix, specify the forecast origin and evaluation year, and use a longer chronologically separated test period. It should compare persistence and climatological baselines, report performance by state, and evaluate the probability vectors as well as modal labels (Gneiting & Raftery, 2007). Multi-site rainfall-event models or radar-based generators address richer spatial information (Andersen et

al., 2024; Gao et al., 2021), but they require inputs beyond the single-station state sequence used here.

For educational use, the example makes the distinction between conditional persistence, stationary proportions, and categorical prediction visible. No community intervention or measured effect on planning decisions was evaluated. Any application to community weather information would therefore require further validation and careful communication of uncertainty before operational use.

CONCLUSION

The reported first-order Markov model describes three rainfall-based states at Belawan station using observations from 2020–2024. Among 1,725 classified days, the zero-rainfall state is most frequent. Its persistence probability is 0.600, compared with 0.172 for the 0.1–0.99 mm state and 0.487 for the state with at least 1 mm of rainfall. The stationary distribution is approximately (0.477, 0.138, 0.385), which describes the long-run behavior of the fitted model rather than verified long-term weather conditions.

In the seven-day fixed-origin illustration, S is the most probable category at every horizon, but its probability approaches 0.477 rather than 1.000. Three of the six valid observed categories are predicted correctly, yielding 50% accuracy, and all three R observations are missed. This small comparison provides no demonstrated advantage over a constant-S forecast. The model is therefore useful as an interpretable local example, while operational forecasting claims remain unsupported. Confirming the data-processing rules, reconstructing transition counts, identifying the evaluation year, and conducting longer baseline-based validation are necessary next steps.

BIBLIOGRAPHY

- Andersen, C. B., Wright, D. B., & Thorndahl, S. (2024). CON-SST-RAIN: Continuous stochastic space–time rainfall generation based on Markov chains and transposition of weather radar data. *Journal of Hydrology*, 637, Article 131385. <https://doi.org/10.1016/j.jhydrol.2024.131385>
- Arnita, Afnisah, N., & Marpaung, F. (2020). A comparison of the fuzzy time series methods of Chen, Cheng and Markov chain in predicting rainfall in Medan. *Journal of Physics: Conference Series*, 1462(1), 012044. <https://doi.org/10.1088/1742-6596/1462/1/012044>
- Azimi, S., Hassannayebi, E., Boroun, M., & Tahmoures, M. (2020). Probabilistic analysis of long-term climate drought using steady-state Markov chain approach. *Water Resources Management*, 34(15), 4703–4724. <https://doi.org/10.1007/s11269-020-02683-5>
- Azizah, A., Welastika, R., Falah, A. N., Ruchjana, B. N., & Abdullah, A. S. (2019). An application of Markov chain for predicting rainfall data at West Java using data mining approach. *IOP Conference Series: Earth and Environmental Science*, 303(1), 012026. <https://doi.org/10.1088/1755-1315/303/1/012026>

- Chowdhury, A. F. M. K., Kar, K. K., Shahid, S., Chowdhury, R., & Rashid, M. M. (2019). Evaluation of spatio-temporal rainfall variability and performance of a stochastic rainfall model in Bangladesh. *International Journal of Climatology*, 39(11), 4256–4273. <https://doi.org/10.1002/joc.6071>
- Chowdhury, A. F. M. K., Lockart, N., Willgoose, G., Kuczera, G., Kiem, A. S., & Parana Manage, N. (2017). Development and evaluation of a stochastic daily rainfall model with long-term variability. *Hydrology and Earth System Sciences*, 21(12), 6541–6558. <https://doi.org/10.5194/hess-21-6541-2017>
- Fikri, A. M., & Mutaqin, A. K. (2022). Penerapan model pembangkit curah hujan stokastik untuk simulasi curah hujan harian di stasiun badan meteorologi klimatologi dan geofisika (BMKG) Kertajati Jawa Barat. *Bandung Conference Series: Statistics*, 2(2), 79–86. <https://doi.org/10.29313/bcss.v2i2.3220>
- Gao, C., Booij, M. J., & Xu, Y.-P. (2020). Development and hydrometeorological evaluation of a new stochastic daily rainfall model: Coupling Markov chain with rainfall event model. *Journal of Hydrology*, 589, Article 125337. <https://doi.org/10.1016/j.jhydrol.2020.125337>
- Gao, C., Guan, X., Booij, M. J., Meng, Y., & Xu, Y.-P. (2021). A new framework for a multi-site stochastic daily rainfall model: Coupling a univariate Markov chain model with a multi-site rainfall event model. *Journal of Hydrology*, 598, Article 126478. <https://doi.org/10.1016/j.jhydrol.2021.126478>
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>
- Gobet, E., Métivier, D., & Parey, S. (2025). Interpretable seasonal multisite hidden Markov model for stochastic rain generation in France. *Advances in Statistical Climatology, Meteorology and Oceanography*, 11(2), 159–201. <https://doi.org/10.5194/ascmo-11-159-2025>
- Gregory, J. M., Wigley, T. M. L., & Jones, P. D. (1993). Application of Markov models to area-average daily precipitation series and interannual variability in seasonal totals. *Climate Dynamics*, 8(6), 299–310. <https://doi.org/10.1007/bf00209669>
- Hardianti, W. T. (2024). Penentuan pola distribusi curah hujan harian kabupaten Bogor dengan model rantai Markov. *Journal of Sciencetech Research and Development*, 6(1), 1600–1609. <https://doi.org/10.56670/jsrd.v6i1.486>
- Ihsan, H., Sanusi, W., & Hasriani, H. (2020). Peramalan pola curah hujan di kota Makassar menggunakan model rantai Markov. *Journal of Mathematics, Computations, and Statistics*, 2(1), 19–30. <https://doi.org/10.35580/jmathcos.v2i1.12448>
- Jale, J. D. S., Xavier Júnior, S. F. A., Xavier, É. F. M., Stošić, T., Stošić, B., & Ferreira, T. A. E. (2019). Application of Markov chain on daily rainfall data in Paraíba-Brazil from 1995-2015. *Acta Scientiarum. Technology*, 41(1), e37186. <https://doi.org/10.4025/actascitechnol.v41i1.37186>

- Laily, Y. H., Rakhmawati, F., & Husein, I. (2023). Penerapan metode fuzzy time series-Markov chain dalam peramalan curah hujan sebagai jadwal tanaman padi. *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, 4(1), 162–174. <https://doi.org/10.46306/lb.v4i1.235>
- Nasib, S. K., Nurwan, Yanuari, E. D. D., & Macmud, T. (2022). Karakteristik rantai Markov pada data curah hujan bulanan stasiun Djalaluddin. *JMPM: Jurnal Matematika dan Pendidikan Matematika*, 7(2), 81–89. <https://doi.org/10.26594/jmpm.v7i2.2654>
- Nguyen, T. H. T., Bennett, B., & Leonard, M. (2023). Evaluating stochastic rainfall models for hydrological modelling. *Journal of Hydrology*, 627, Article 130381. <https://doi.org/10.1016/j.jhydrol.2023.130381>
- Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1), 182–190. <https://doi.org/10.1029/wr017i001p00182>
- Wahyudi, S. R., Sari, R. F., & Widiasari, R. (2021). Penentuan pola penyebaran curah hujan harian kabupaten Karo dengan menggunakan rantai Markov orde-n. *Jurnal Riset dan Aplikasi Matematika*, 5(2), 144–157. <https://doi.org/10.26740/jram.v5n2.p144-157>