Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 388~397
e-ISSN: 2548-6861

Hybrid Rainfall Analysis in Semarang by Integrating SARIMA Predictions with Meteorological Association Rules

Kristina Agustin [1]*, Ika Novita Dewi [2]*
* Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang
112202206904@mhs.dinus.ac.id[1], ikadewi@dsn.dinus.ac.id[2]

Article Info

Article history:

Received 2025-12-12
Revised 2026-01-02
Accepted 2026-01-13

Keyword:

Hydrometeorological Analysis,
Rainfall Forecasting,
SARIMA,
Association Rule Mining,
Apriori Algorithm.

ABSTRACT

Climate variability necessitates advanced analytical approaches to understand irregular rainfall patterns, particularly in coastal cities like Semarang, Central Java. This research employs a dual-analysis framework combining the Seasonal Autoregressive Integrated Moving Average (SARIMA) model and the Apriori algorithm to forecast rainfall and uncover hidden meteorological associations. Analyzing BMKG monthly climatological data from January 2020 to December 2024, the research addresses both temporal trends and variable dependencies. The SARIMA $(1,0,0)(2,1,0)_{12}$ model projected rainfall dynamics for 2025, identifying critical wet periods (January-March, November-December) and dry intervals (July-September), achieving a MAPE of 44.97%. To complement temporal forecasting, the Apriori algorithm was applied with 50% minimum support and 50% confidence, generating association rules from daily discretized meteorological data. Results reveal that the combination of low temperature (Tx_Low, Tn_Low) and moderate wind speed (FFx_Medium) exhibits the strongest correlation with heavy rainfall events Lift Ratio 12.34, indicating a 12-fold increased risk compared to random conditions. By synergizing temporal forecasting with the identification of meteorological triggers, this research offers a robust basis for early warning systems, supporting flood mitigation and water resource management strategies in Semarang.

I. INTRODUCTION

Global climate change in recent years has caused weather patterns to become increasingly irregular and difficult to predict. These uncertain conditions have amplified the frequency and intensity of extreme weather events, particularly rainfall [1]. As a critical climatic element measured in millimeters (mm) over specific periods, rainfall variability serves as a primary indicator of hydrological dynamics [2]. However, rainfall patterns are showing significant anomalies due to global warming. This instability is exacerbated by periodic phenomena such as La Niña and El Niño, which disrupt rainfall distribution in Indonesia, especially during the transitional seasons [3]. Furthermore, the Indian Ocean Dipole (IOD) adds another layer of complexity; a positive IOD can worsen drought during El Niño, while a negative IOD occurring simultaneously with La Niña can trigger extreme rainfall events [4].

The uncertainty of these patterns has severely impacted various regions in Indonesia, particularly Semarang City, the capital of Central Java. Located on the northern coast of Java Island, Semarang experiences high rainfall intensity, often resulting in severe flooding [5]. Despite the recurring risks, existing prediction systems remain unable to provide adequate early warning information. Consequently, there is an urgent need for a system that goes beyond merely projecting future rainfall quantities. An effective solution must systematically identify seasonal patterns, trends, and the triggering meteorological variables to serve as a robust basis for risk mitigation.

Anticipatory action relies on accurate forecasting and a deep understanding of the factors driving rainfall. Time series analysis and data mining are highly relevant approaches to addressing this challenge [6]. To capture future rainfall patterns, the Seasonal Autoregressive Integrated Moving Average (SARIMA) model offers an effective solution.

SARIMA is statistically designed to analyze time series data exhibiting periodic seasonal patterns, accommodating both long-term trends and random fluctuations [7]. By utilizing univariate monthly rainfall data, SARIMA can estimate periods of peak and minimum rainfall. Thus, it serves as a viable foundation for an early warning system in Semarang City.

However, predicting the "when" is not enough; understanding the "why" is equally crucial. Data mining offers techniques to explore hidden relationships within complex meteorological data [8]. One such method is the Apriori Association Rule Mining algorithm, which identifies meaningful association rules and item combinations [9]. This method is employed to uncover patterns between climatological variables, such as temperature (minimum, maximum, average), humidity, and wind speed and rainfall events. The Apriori algorithm utilizes support, confidence, and lift ratio parameters to evaluate the strength of these relationships [10]. By revealing hidden patterns undetectable by standard statistical analysis, Apriori provides a deeper explanation of the meteorological conditions triggering rain.

Previous researches have demonstrated the efficacy of SARIMA in rainfall forecasting with varying results. Ramli et al. (2023) achieved a prediction accuracy of 80.5% (MAPE 19.5%) in Aceh using a SARIMA model $(0,0,1)(0,0,1)_{12}$ [11]. Similarly, Adams et al. (2020) successfully forecasted a 10% increase in rainfall in Abuja, Nigeria, using a SARIMA $(0,0,2)(0,1,2)_{12}$ model after confirming stationarity via the Augmented Dickey-Fuller test [12]. Furthermore, Kabbilawsh et al. (2022) applied SARIMA to 29 stations in India, finding that seasonal components were dominant in long-term rainfall data [13]. These researches confirm that SARIMA is a consistent and reliable tool for capturing seasonal hydrological cycles.

Parallel to this, the Apriori algorithm has proven effective in identifying meteorological associations. Gunawan et al. (2023) used Apriori in Tegal City, producing the highest accuracy of 78.68%. All association rules had a lift ratio >1, indicating significant and reliable power for predicting rainfall [14]. Coulibaly et al. (2021) also applied association rule learning to weather prediction, identifying temperature, humidity, and wind speed as the most frequent antecedents for rainfall events [15]. These findings underscore the Apriori algorithm's ability to explain the specific weather conditions that trigger precipitation.

Despite these advances, most researches focus either solely on forecasting the time of rainfall or solely on the causal variables. Few attempts have been made to integrate these perspectives. This research addresses this gap by combining two algorithms typically used separately SARIMA and Apriori. Using monthly climate data from the Meteorology, Climatology, and Geophysics Agency (BMKG) Semarang City (January 2020–December 2024), this research aims to predict monthly rainfall patterns for 2025 while simultaneously identifying the interrelationships between meteorological variables that trigger these events. This dual-analysis approach integrates temporal forecasting with pattern discovery to provide a comprehensive tool for hydrometeorological disaster mitigation.

II. METHOD

The methodology adopts a dual-analysis approach using dataset from Semarang City. The Seasonal Autoregressive Integrated Moving Average (SARIMA) is used to capture temporal seasonality and forecast future rainfall, while the Apriori algorithm is employed to identify meteorological triggers through association rule mining. Figure 1 presents the comprehensive flowchart guiding this research process.

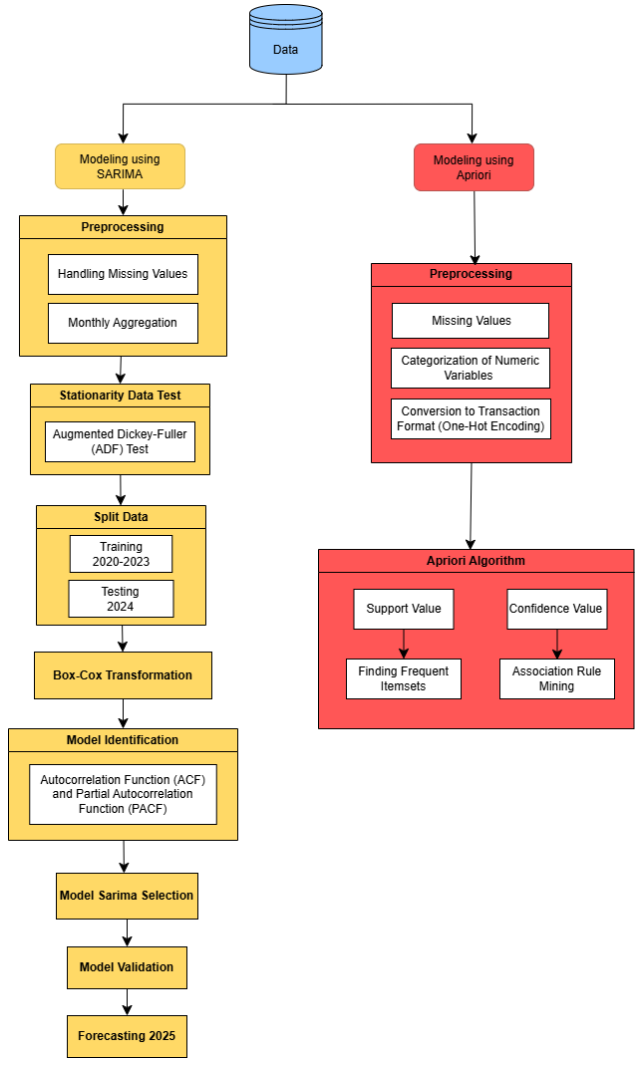

Figure 1. Research flow

A. Dataset

The research utilizes meteorological data acquired from BMKG [16], covering the Semarang City region. The dataset spans a five-year period from January 2020 to December 2024 and comprises seven key variables: rainfall, average temperature, minimum temperature, maximum temperature, average humidity, average wind speed, and maximum wind

speed. This dataset supports two distinct but complementary analytical approaches. First, the SARIMA method employs the univariate monthly rainfall series to forecast precipitation levels for 2025 and identify seasonal peaks and troughs. Second, the Apriori algorithm leverages the complete multivariate dataset to perform association rule mining, aiming to uncover latent relationships between climatological variables and rainfall occurrences.

B. Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA is a time series statistical model designed to analyze data that shows trends and seasonal patterns simultaneously. This model is a refinement of the traditional ARIMA model, but with a seasonal component that can detect recurring patterns at certain time intervals. SARIMA combines three main components, namely autoregressive (AR) which models the relationship between current observations and previous observations, differencing (I) which is used to create stationary data, and moving average (MA) which models the relationship between observations and past errors [16]. The SARIMA model is generally denoted as $ARIMA(p,d,q)(P,D,Q)_s$, representing both non-seasonal and seasonal components. In this notation, the non-seasonal part is defined by the parameters p for the autoregressive (AR) order, d for the differencing order to achieve stationarity, and q for the moving average (MA) order. The seasonal characteristics of the model are captured by the parameters $P, D,$ dan Q which represent the seasonal autoregressive, seasonal differencing, and seasonal moving average orders, respectively. Furthermore, the subscript s denotes the specific number of seasonal periods, which in this research corresponds to the 12-month annual hydrological cycle. In general, the form of the SARIMA [17] presents in equation (1)

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (1)$$

where:

$\phi_p(B)$: non-seasonal AR coefficient with order p
$\Phi_P(B^S)$: seasonal AR coefficient with order p
$(1-B)^d$: operator for difference of order d
Y_t : observation value at time -t,
$\theta_q(B)$: non-seasonal MA coefficient with order q
$\Theta_Q(B^s)$: seasonal MA coefficient with order q
ε_t : random error (white noise).

1) Preprocessing

Data preprocessing is critical to ensure dataset integrity prior to modeling. This phase involved two primary procedures to prepare the rainfall data for time series analysis. First, Linear Interpolation was employed to address data gaps caused by recording anomalies. This method estimates missing values based on the slope between adjacent known data points, thereby preserving the temporal continuity essential for time series analysis [18]. Subsequently, daily rainfall records were aggregated into monthly totals to align the data scale with the SARIMA model's capacity to detect medium-to-long-term periodic seasonality. This temporal aggregation transforms the granular daily observations into a format more suitable for capturing the broader seasonal patterns inherent in rainfall data.

2) Stationarity Data Test

Stationarity data test is a statistical test to determine whether data has a constant mean, variance and autocorrelation over time [19]. The Augmented Dickey-Fuller (ADF) test is used as a formal testing method to detect the presence of a unit root which indicates non-stationarity. The null hypothesis (H_0) states that the data has a unit root (not stationary), while the alternative hypothesis (H_1) states that the data is stationary. If the p-value of the ADF test is smaller than the significance level of 0.05, then H_0 is rejected and the data is considered stationary. Data that does not meet stationary conditions requires transformation or differencing before SARIMA modelling.

3) Split Data

Split data is the process of dividing a dataset into two subsets, namely training data and testing data for the purposes of training and evaluating models [20]. This process aims to prevent overfitting and validate the generalization ability of the model on data that has not been seen during training. The training data covers historical records from January 2020 to December 2023, for a total of 48 months, which is considered sufficient to capture seasonal variations with a period of s=12. For evaluation purposes, the testing data covers the period from January to December 2024.

This research uses an 80:20 split ratio, which is common practice in time series modelling to balance training and validation needs. The model's performance on the testing data is then measured using MAE, RMSE, and MAPE metrics as indicators of future forecasting reliability

.

4) Box-Cox Transformation

Box-Cox transformation is a statistical technique used to stabilize variance and normalize the distribution of time series data [21]. This transformation is necessary when the data shows heteroscedasticity (variance is not constant) or a non-normal distribution, so as to better meet the assumptions of the SARIMA model. The lambda parameter (λ) of the Box-Cox transformation is estimated only on training data to prevent data leakage. The parameters λ that have been obtained are then applied to the testing data without refitting, ensuring that the model does not have information from future data during training. The inverse transformation process uses the same λ applied to the prediction results to return the values to the original scale by ensuring the confidence interval does not produce negative values.

5) Model identification (ACF/PACF)

Model identification is the stage of determining the optimal parameters of SARIMA through Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analysis [22]. ACF measures the correlation between observations at time t and observations at time t-k for various lag k values, which is used to identify the order of the moving average (q and Q). PACF measures the correlation after removing the influence of the intermediate lag, which is used to identify the autoregressive order (p and P). The cut-off pattern on the ACF chart indicates the MA order, while the cut-off pattern on the PACF indicates the AR order. The significant spike at lag multiple s=12 shows a strong seasonal component, indicating the need for seasonal parameters (P, D, Q). The combination of ACF and PACF analysis provides initial guidance for optimal SARIMA grid search parameters.

6) Model Sarima Selection

Model selection is the process of selecting the best model from various combinations of SARIMA parameters based on statistical criteria. Grid search is carried out by trying various combinations of parameters (p, d, q) $(P, D, Q)_s$, and comparing performance using information criteria. The model was chosen with the AIC value because it indicated the optimal balance between goodness of fit and model complexity, as well as the lowest MAPE for prediction accuracy. The model with the lowest AIC value that meets statistical tests is selected as the optimal model [23].

7) Model Validation

Model validation is a verification stage to ensure the SARIMA model meets statistical assumptions and produces accurate predictions. Model validation includes checking the residuals to ensure that the residuals are white noise (uncorrelated), normally distributed, and have constant variance (homoscedasticity). The Ljung-Box test was carried out to detect residual autocorrelation, with a p-value > 0.05 indicating independent residuals. The Jarque-Bera test evaluates the normality of the residuals, where p-value > 0.05 indicates the residuals are normally distributed [24].

8) Forecasting

Future rainfall values are predicted by applying a validated SARIMA model to the complete historical dataset from 2020 to 2024, thereby maximizing the capture of temporal and seasonal patterns. Forecasting results include a point forecast (single predicted value) and a 95% confidence interval which shows the range of prediction uncertainty. The inverse Box-Cox transformation is applied to the method prediction results to ensure that the prediction value and confidence interval are not negative.

C. Apriori Algorithm

The Apriori algorithm is a basic method in data mining that is used to identify frequent itemset and generate association rules in transactional data sets. This algorithm uses a bottom-up approach, in which frequently occurring individual items are systematically identified and repeatedly developed into larger k-itemsets as long as they meet the minimum support threshold [25]. In this meteorological research, Apriori was used to reveal hidden associations between variables such as temperature, humidity, wind speed, and rainfall through three main evaluation metrics: support, confidence, and lift.

Before the mining process began, an important preliminary stage of data preprocessing was carried out to convert raw meteorological data into a suitable format. This stage included handling missing data through linear interpolation to maintain temporal continuity, followed by categorizing numerical variables into discrete labels to facilitate pattern recognition. The results of handling missing values are shown in Table 1.

TABLE I
SAMPLE OF PREPROCESSED METEOROLOGICAL DATA

Date	**TAVG**	**RH_AVG**	**RR**	**...**	**FF_AVG**
01–01-2020	27.0	88.0	9.6	…	4.0
02-01-2020	27.4	87.0	16.7	…	3.0
03-01-2020	28.4	84.0	2.0	…	3.0
04-01-2020	27.1	90.0	36.6	…	2.0
05-01-2020	26.8	92.0	3.7	…	2.0
…	…	…	…	…	...
27-05-2024	28.3	82.0	0.0	…	2.0
28-05-2024	28.1	86.0	0.0	…	1.0
29-05-2024	27.1	88.0	29.8	…	1.0
30-05-2024	28.4	80.0	0.4	…	3.0
31-05-2024	28.2	77.0	1.4	…	3.0

Categorization of numeric variables or discretization is the process of converting continuous numerical data into discrete categorical data by dividing a range of values into certain intervals [26]. This process is very important because Apriori algorithm is designed to work with categorical or transactional data. The categorization process helps simplify data, reduce noise, and make patterns easier to interpret. This research implements two discretization approaches, namely manual rule-based binning which applies a fixed threshold based on BMKG meteorological standards [27], and quantile-based binning with parameter q=3 (tertile) which divides the data based on statistical distribution.

After evaluation, the manual rule-based binning approach was chosen because it produces categories that are easy to interpret, as shown in Table 2.

TABLE II
CATEGORIZATION OF NUMERIC VARIABLES

Code	Variable	Category	Range
RR	Rainfall	No Rain Light Rain Medium Rain Heavy Rain Very Heavy Rain	0 mm 0.1-20mm 20-50mm 50-100mm >100mm
Tn	Minimum temperature	Tn_ Low Tn_Medium Tn_ High	$<$ 24℃ 24 - 26℃ $>$ 26℃

Tx	Maximum temperature	Tx_Low Tx_ Medium Tx_ High	$< 32℃$ $32 - 34℃$ $> 34℃$
Tavg	Average temperature	Tavg_Low Tavg_ Medium Tavg_ High	$< 28℃$ 28 - 30℃ $> 30℃$
RH_avg	Average humidity	RH_Dry RH_Normal RH_ High	< 70% 70 – 85% > 85%
FF_x	Maximum wind speed	FFx_Weak FFx_Medium FFx_ Fast	< 4m/s 4 – 7m/s > 7m/s
FF_avg	Average wind speed	FFavg_Weak FFavg_Medium FFavg_ Fast	< 2m/s 2 – 3m/s > 3m/s

The final stage of data processing involves conversion to transaction format, which is the process of transforming categorical data into transactional format using the one-hot encoding technique. This technique converts each category of each variable into a binary column (0 or 1), where a value of 1 indicates the presence of an item in the transaction and 0 indicates its absence [16]. This transformation process enables the Apriori algorithm to identify patterns of co-occurrence of various meteorological conditions that occur simultaneously, with the results of the one-hot encoding conversion presented in Table 3.

TABLE III
CONVERSION TO TRANSACTION FORMAT

FFavg_ Fast	FFx_ Medium	Very Heavy Rain	....	Medium Rain
1	1	0	....	0
0	1	0	....	0
0	1	1	....	0
0	0	0	....	1
0	0	1	....	0

As the core stage of the association rule mining process, Apriori algorithm aims to find frequent itemsets and generate strong association rules. To ensure model robustness, a grid search evaluation was conducted across various support and confidence thresholds. This process aims to determine the optimal sensitivity for detecting both frequent seasonal patterns and rare, high-impact meteorological anomalies. The strength of the resulting association rules is then evaluated through three main metrics, starting from the support value, confidence, and lift ratio.

Support value is a threshold parameter that determines how often an itemset must appear in the dataset to be considered frequent or significant. The minimum support threshold is set before the algorithm is run and functions as a filter to eliminate itemsets that rarely appear. Itemsets that have a support value above or equal to the minimum support threshold will be considered frequent itemsets and retained for the next iteration, while itemsets with support below the threshold will be discarded. Choosing the right minimum support value is very important: a value that is too high can result in the loss of interesting patterns, while a value that is too low can result in too many meaningless rules.

In the context of meteorological data, support value shows how often a certain combination of weather conditions occurs within the observation period. The support calculation is in the equation (2).

$$Support\ (A) = \frac{The\ number\ of\ transactions\ is\ A}{Total\ transactions} \times 100\% \quad (2)$$

The support value of the 2 items is obtained based on equation calculations (3).

$$Support\ (A,B) = \frac{The\ number\ of\ transactions\ is\ A\ and\ B}{Total\ transactions} \times 100\% \quad (3)$$

Confidence as a metric that measures how often item B appears in transactions containing item A, or in other words, the conditional probability that the consequent will occur if the antecedent occurs. Once frequent itemsets are found, the algorithm generates association rules of the form "If A then B" (A → B), where A is the antecedent and B is the consequent. Confidence value shows the percentage of transactions that contain antecedent A that also contain consequent B. The minimum confidence threshold is used to filter weak association rules, and only rules with confidence above the threshold will be retained as strong association rules. A high confidence value indicates that the rule is reliable and can be trusted to predict the emergence of consequences based on antecedents. Confidence calculations use calculations in the equation (4)

$$Confidence\ P(B/A) = \frac{The\ number\ of\ transactions\ is\ A\ and\ B}{Total\ transactions} \times 100\% \quad (4)$$

The lift ratio is a metric that measures how strong the relationship between antecedents and consequences is compared to if the two items were independent. The lift ratio shows whether items A and B appear together more often than expected if they were independent. The results of the lift ratio value can be used to assess the validity or strength of the rules formed. The lift ratio calculation is shown in the equation (5) and (6)

$$Lift\ Ratio = \frac{Confidence\ (A,B)}{Support\ (B)} \quad (5)$$

Or

$$Lift\ Ratio = \frac{Support(A \cup B)}{(Support(A) \times Support(B))} \quad (6)$$

III. RESULT AND DISCUSSION

A. Result of SARIMA Modeling

1) Identification of Annual Rainfall Time Series Plots

The initial stage of SARIMA modeling is identifying data characteristics through time series graphic visualization. The data used is monthly rainfall data collected from daily BMKG Semarang City data for the period January 2020 to December 2024, with a total of 60 monthly observations. The time series

visualization in Figure 2 aims to identify the existence of seasonal patterns, trends and data variability, which are the basis for determining the SARIMA model specifications.

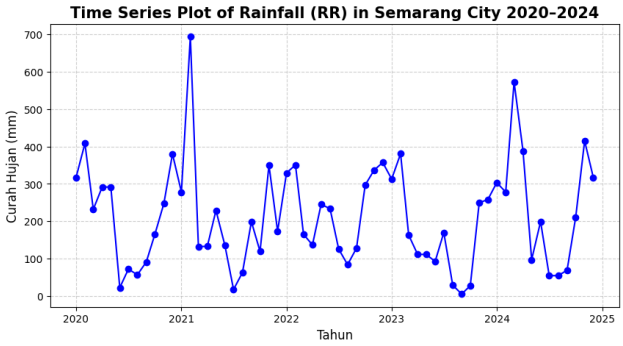

Figure 2. Time series plot of rainfall in Semarang city (2020-2024)

Figure 2 shows a plot of monthly rainfall time series in Semarang City. A clear seasonal pattern can be seen, with high rainfall recurring in the period November-March (rainy season) and low rainfall in the period June-September (dry season). The data shows high variability with several periods experiencing extreme rainfall >600 mm/month, particularly in early 2021 and 2024. Furthermore, formal testing using the Augmented Dickey-Fuller (ADF) test will be conducted to confirm the statistical stationarity of the data.

TABLE IV
STASIONARY TEST

Statistics	Value
ADF p-value	-3.433522
p-value	0.009868
Critical Value (1%)	-3.568486
Critical Value (5%)	-2.921360
Critical Value (10%)	-2.598662

Table 4 shows the results of the ADF test which produces a p-value of 0.009868 (<0.05) with a statistical ADF value of -3.433522 which is smaller than the critical value of 5% (-2.921360), so the null hypothesis is rejected. This result confirms that the data is stationary in the mean, so it does not require non-seasonal differencing (d=0).

2) *Box-Cox Transformation and Parameter Identification*

Even though the data is stationary, further evaluation of distribution normality and homoscedasticity is needed to comprehensively meet the assumptions of the SARIMA model. Data is divided into training set (48 months) and testing set (12 months) with a ratio of 80:20 before transformation to prevent data leakage. The Jarque-Bera test on the training set produced a p-value of 0.0006 (<0.05) with a skewness of 0.9731, indicating a non-normal distribution. The variance ratio is 4.67x (>3x) indicating heteroscedasticity. This condition can affect the accuracy of parameter estimates and the reliability of confidence intervals. Box-Cox transformation with parameter λ=0.3156 (estimated only from the training set) succeeded in normalizing the distribution (Jarque-Bera p-value increased to 0.9695) and stabilized the variance (variance ratio decreased to 3.01x), as shown in Table 5.

TABLE V
COMPARISON OF ORIGINAL STATIONARY DATA TEST AND BOX-COX TRANSFORMATION

Metric	Original	Box-Cox
Jarque-Bera *p-value*	0.0006	0.9695
Skewness	0.9731	-0.0731
Kurtosis	1.8886	-0.0978
ADF p-value	0.0099	0.0449
Variance Ratio	4.67x	3.01x

After the data is transformed, ACF and PACF analysis is carried out on the training set to identify the optimal SARIMA model parameter order. Visual parameter identification was conducted using ACF and PACF plots on the transformed training data. The ACF plot (Figure 3) exhibits a slow decay in non-seasonal lags but displays significant spikes at multiples of lag 12 (12, 24, 36). This pattern confirms a strong seasonal component, necessitating the inclusion of a seasonal differencing parameter (D=1). Conversely, the PACF plot (Figure 4) demonstrates a distinct cut-off after lag 1, providing a strong indication for a non-seasonal Autoregressive term of order 1 (p=1). Based on these visual diagnostics, the grid search parameter space was constrained to prioritize seasonal components and low-order autoregressive terms.

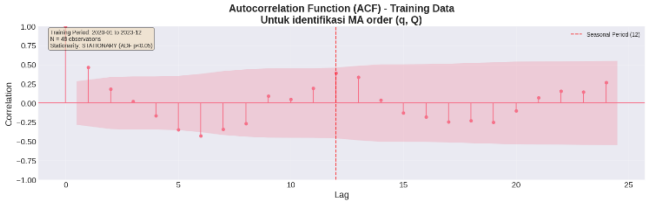

Figure 3. ACF Plot

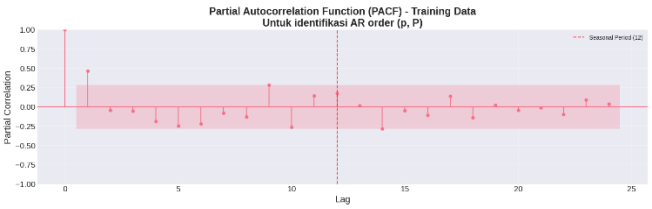

Figure 4. PACF Plot

3) *SARIMA Model Selection*

Following the parameter identification phase, a comprehensive grid search was executed to evaluate 81 distinct parameter combinations based on the range $p, q \in [0,2]$, $d \in [0,1]$, and seasonal parameters $P, Q \in [0,2]$, $D \in [0,1]$. The selection criteria prioritized minimizing the AIC while ensuring predictive accuracy (MAPE) and satisfying residual diagnostic assumptions. Table 6 presents a comparative summary of the top-performing models against representative alternative candidates.

TABLE VI
SARIMA MODEL SELECTION

SARIMA Model	AIC	MAPE	LB(p)	JB(p)
$(1,0,0)(2,1,0)_{12}$	86.05	44.97	0.158	0.607
$(1,0,0)(2,1,1)_{12}$	87.97	59.19	0.174	0.619
$(1,0,0)(2,1,2)_{12}$	89.90	67.25	0.192	0.633
$(0,1,1)(0,1,1)_{12}$	157.21	55.88	0.236	0.513
$(0,1,1)(0,0,2)_{12}$	165.84	62.34	0.104	0.222

The SARIMA $(1,0,0)(2,1,0)_{12}$ model was identified as the optimal structure, achieving the lowest AIC (86.05) and MAPE (44.97%). The table highlights a significant performance divergence: while the top three models mentioned AIC scores below 90, alternative candidates (e.g., Rank 4 and 5) exhibited a sharp increase in AIC values (>157). This substantial gap statistically confirms that the selected model structure provides a significantly superior fit compared to other potential combinations. Furthermore, the selected model demonstrated robust statistical validity. As shown in the diagnostic columns of Table VI, the Ljung-Box test yielded a p-value of 0.158 (> 0.05), confirming the absence of autocorrelation in residuals. Similarly, the Jarque-Bera test resulted in a p-value of 0.607 (> 0.05), validating that the residuals follow a normal distribution. Consequently, this model is adopted for forecasting as it offers the best balance between statistical efficiency and validity.

4) *Performance Evaluation on Testing Data*

The SARIMA $(1,0,0)(2,1,0)_{12}$) model which had been statistically validated was then evaluated for its generalization ability on the testing set for the period January to December 2024 which the model had never seen during training. Evaluation is carried out by comparing the predicted value of the Box-Cox inverse transformation results with the actual value using several standard performance metrics in time series forecasting.

TABLE VII
SUMMARY OF FORECASTING ERRORS FOR THE 2024 TEST DATA

Metric	Value
MAE	123.21 mm
RMSE	171.35 mm
MAPE	44.97%

Based on Table 7, the evaluation yielded an MAE of 123.21 mm, an RMSE of 171.35 mm, and a MAPE of 44.97%. While this value indicates a moderate level of prediction deviation, it represents the optimal achievable accuracy for this specific dataset. As previously demonstrated in the model selection phase Table 6, alternative structural models yielded significantly higher error rates, ranging from 55% to over 100%. This performance underscores the inherent complexity of modeling Semarang's stochastic tropical weather using univariate time-series data alone. The deviations are largely driven by extreme rainfall anomalies in early 2024, which extend beyond historical seasonal trends. Consequently, to address these irregularities, a supplementary analysis of meteorological associations is required to uncover the specific variable interactions triggering these events, thereby complementing the seasonal baseline established by SARIMA.

5) *Rainfall Forecasting in 2025*

After the model was validated on testing data, the model was then retrained using all data (2020-2024) to produce rainfall predictions for 2025. Table 8 displays the results of monthly rainfall forecasting for Semarang City throughout 2025.

TABLE VIII
MONTHLY RAINFALL FORECASTING IN 2025

Month	Prediction (mm)
January	323.58
February	334.43
March	306.49
April	213.39
May	124.84
June	156.42
July	107.27
August	47.28
September	41.39
October	127.74
November	330.06
December	299.35

Table 8 shows clear characteristics based on the BMKG's deterministic monthly rainfall classification. Based on this standard, January to March and November to December are classified as high rainfall due to their values (300-500 mm/month), while July to September is categorized as the dry season due to their values (0-100 mm/month)..

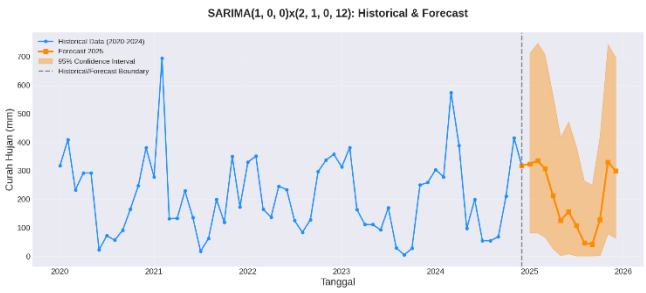

Figure 5. Rainfall Prediction For 2025

Figure 5 visualizes the forecast, showing a consistent seasonal pattern where peak rainfall is expected in January (323.58 mm) and February (334.43 mm), while the dry season is projected to occur from July to September (<110 mm/month). It is important to interpret these projections within the context of data limitations. The 95% Confidence Interval (shaded area in Figure 5) widens significantly during the peak rainy season, reflecting the higher variance and uncertainty inherent in the limited five-year historical dataset. While the SARIMA model successfully captures the recurring seasonal periodicity, the magnitude of extreme rainfall events may deviate from the point forecast. Therefore, these predictions

should be utilized as a baseline trend indicator for flood mitigation planning, with real-time adjustments made based on short-term meteorological alerts.

B. Results of the Apriori Algorithm

1) *Frequent Itemset and Parameter Robustness Analysis*

The implementation of the Apriori algorithm begins with the identification of frequent itemsets, defined as combinations of meteorological variab The Apriori algorithm is usedles such as Tn, Tx, Tavg, RH_avg, ff_x, ff_avg that co-occur in the dataset with a frequency exceeding a certain threshold. For association analysis, continuous daily weather data are discretized into categorical labels as detailed in the methodology section. This research applies a grid search evaluation to test the robustness and sensitivity of the pattern extraction process to variations in parameter thresholds. Figure 6 visualizes the distribution of rules formed at various ranges of Minimum Support (10%–50%) and Confidence (50%–70%).

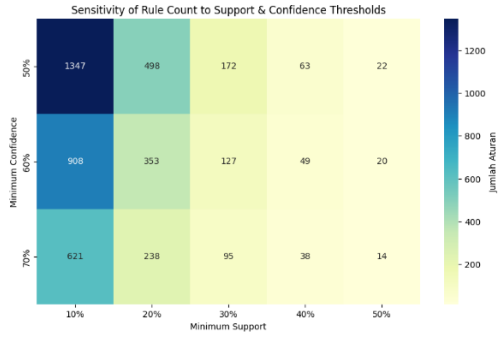

Figure 6. Heatmap Visualization of Parameter Sensitivity

Parameter sensitivity and robustness analyses were used to ensure model robustness. Grid search evaluation was performed to assess the Apriori algorithm against various thresholds. Figure 6 visualizes the sensitivity of the resulting rules. The heatmap shows that at higher support levels (>10%), the algorithm is unable to capture the details of complex weather patterns. Because extreme rainfall is a rare anomaly (less than 3% of the total data), the analysis is generally less successful. Therefore, based on this robustness test, the analysis was expanded to a lower support threshold of 0.5% to detect potential flood-causing events.

2) *Association rules Mining*

Association rules are expressed in the form "If A then B" (A →B), where the strength of the rule is measured using three main metrics: support which shows the frequency of occurrence of the combination A and B, confidence which shows the conditional probability of B appearing if A occurs, and lift ratio which shows the degree of correlation between A and B compared to if they were independent. Unlike the initial exploration, which only measured the volume of rules generated, this stage focused on extracting meaningful meteorological patterns. Based on previous robustness tests, the mining process was carried out using two targeted strategies, validating consistency during the rainy season and identifying early warning indicators for extreme events. To implement the first strategy, the analysis was specifically segmented to the rainy season (November–March) to validate seasonal consistency and maximize the detection of flood-triggering patterns. Table 9 presents the top association rules extracted from this high-risk period, ordered by the highest lift ratio.

TABLE IX
TOP 5 SEASONAL ASSOCIATION RULES (NOV–MAR)

Antecedents and Consequent	Support	Confidence	Lift Ratio
RH_High, Tavg_Low, Tx_Low → Medium Rain	5.0	15.2	1.51
FFx_Fast, Tn_Medium → Light Rain	5.5	73.7	1.50
RH_High, Tavg_Low → Medium Rain	5.2	14.6	1.45
RH_High, Tx_Low → Medium Rain	5.2	14.4	1.43
Tavg_Low, Tx_Low → Medium Rain	6.2	14.1	1.40

While the SARIMA model utilizes a full 12-month dataset to capture annual trends, the Apriori analysis was strategically segmented specifically for the rainy season (November–March). This targeted approach was chosen to validate seasonal consistency and maximize the sensitivity of flood trigger pattern detection, which in annual analyses is often obscured by the large amount of non-rainfall data. The results of this segmented analysis successfully identified a specific meteorological pattern: a combination of High Humidity (RH_High) and Low Temperature (Tavg_Low, Tx_Low) significantly increases the probability of Moderate Rain, with a Lift Ratio reaching 1.51. This confirms that high humidity accompanied by a decrease in temperature is the dominant cause of rainfall during this period. As a complement to seasonal analysis, the second strategy focuses on detecting extreme rainfall events. Although rare, these phenomena have a major impact, making their detection crucial for establishing early warning mechanisms. Since extreme rainfall is a statistical anomaly, the minimum support is adjusted to 0.5% so that this critical pattern can be detected. Table 10 presents the rules extracted for Heavy Rainfall, which reveal the most significant findings of this research.

TABLE X
TOP 5 SEASONAL ASSOCIATION RULES FOR HEAVY RAINFALL DETECTION

Antecedents and Consequent	Support	Confidence	Lift Ratio
Tx_Low, Tn_Low, FFx_Med → Heavy Rain	0.6	29.7	12.34
FFavg_Med, Tn_Low, Tx_Low → Heavy Rain	0.7	27.1	11.24

FFavg_Med,Tn_Low, Tavg_Low→ Heavy Rain	0.6	26.2	10.87
RH_High, Tn_Low, Tavg_Low → Heavy Rain	0.5	26.2	10.38
Tn_Low, Tavg_Low, FFx_Med→ Heavy Rain	0.7	24.5	10.17

The results reveal a strong early warning signal for hydrometeorological disasters. The top rule indicates that a specific combination of Low Temperature (Tx, Tn, Tavg_Low) and Moderate Wind Speed (FFx_Medium) significantly increase the probability of Heavy Rain, with a Lift Ratio of 12.34. Statistically, this figure indicates that the occurrence of this specific weather pattern increases the risk of heavy rain by up to 12 times compared to random conditions. This finding confirms that the phenomenon of temperature drops accompanied by specific wind dynamics is a valid quantitative parameter as a basis for an early warning system.

To clarify the scientific contribution of this study, Table 11 provides a comparison between the proposed hybrid SARIMA-Apriori framework and previous works. Unlike earlier studies that typically separate temporal forecasting from causal variable analysis, this research successfully integrates both to provide operational validation of flood risks in Semarang City.

TABLE XI
COMPARISON RESULT WITH PREVIOUS STUDIES

References	Method	Results
[11] 2023.	SARIMA	SARIMA $(0,0,1)(0,0,1)_{12}$ model achieved 80.5% accuracy with MAPE 19.5% for forecasting rainfall in Aceh. SARIMA
[12] 2020	SARIMA	SARIMA model reliably detects seasonal pattern changes caused by climate change.
[13] 2022	SARIMA	SARIMA is consistent for multi-location forecasting with different climate characteristics and reliable for strategic water resource planning
[14] 2023	Apriori	Apriori algorithm is effective for predicting rainfall in Tegal City with the highest accuracy of 78.68%. All resulting association rules have a lift ratio value greater than 1, indicating a significant and reliable level of strength for predicting rainfall.
[15] 2021	Apriori	The results indicate that wind speed, wind direction, temperature, humidity, and global radiation are important factors in rainfall formation, making the a priori method effective for improving the
Ours (2025	SARIMA & Apriori	Successfully projected 2025 rainfall dynamics with optimal accuracy (MAPE 44.97%) and identified critical flood-risk periods during January–March and November–December. The proposed approach provides operational validation through the extraction of specific meteorological signatures, where a combination of low temperatures and moderate wind speeds was found to increase heavy rainfall risk by 12.34 times (Lift Ratio 12.34)

IV. CONCLUSION

This research successfully integrated the dual analysis framework of the SARIMA model with the Apriori algorithm to optimize rainfall forecasting and disaster mitigation in Semarang City. The optimized SARIMA model was able to effectively map the hydrological cycle in 2025 and indicate the timing of extreme rainfall events requiring early preparedness. In this scheme, the SARIMA model plays a role in providing numerical estimates of rainfall volume, while the Apriori algorithm serves as an operational validation mechanism through the extraction of significant weather patterns. The main findings of this research indicate that a specific combination of low temperature and moderate wind speed is a strong indicator of heavy rainfall triggers, with a Lift Ratio value reaching 12.34. The combination of these two methods transforms static numerical predictions into a dynamic early warning system. In this system, peak rainfall projections are validated through identified events. Thus, the results of this research provide a stronger foundation for future flood risk management.

To further enhance the precision of this early warning system and address the limitations of statistical error, future research should transition from monthly to daily data aggregation to capture short-term fluctuations. Furthermore, subsequent studies are recommended to adopt hybrid modeling approaches, such as integrating SARIMA with Deep Learning (e.g., LSTM) or multivariate SARIMAX, which can directly incorporate the causal variables identified by Apriori to significantly reduce prediction error and improve resilience against climate anomalies.

REFERENCES

[1] F. Lubis and I. J. A. Saragih, "Performance of probabilistic forecast of the onset of the rainy season over Java Island based on the application of Constructed Analogue (CA) method on Climate Forecast System Version 2 (CFSV2) model output," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Nov. 2021. doi: 10.1088/1755-1315/893/1/012037.

[2] G. Ayu Windari *et al.*, "Mekanisme Terjadinya Hujan dan Pengaruhnya Terhadap Lingkungan," *Jurnal Teknologi Lingkungan UNMUL*, vol. 8, no. 2, 2024.

[3] M. Yustiana, M. Zainuri, D. N. Sugianto, M. P. N. Batubara, and A. M. Hidayat, "Dampak Variabilitas Iklim Inter-Annual (El Niño, La Niña) Terhadap Curah Hujan dan Anomali Tinggi Muka Laut di Pantai Utara Jawa Tengah," *Buletin Oseanografi Marina*, vol. 12, no. 1, pp. 109–124, Feb. 2023, doi: 10.14710/buloma.v12i1.48377.

[4] P. K. Pothapakula, C. Primo, S. Sørland, and B. Ahrens, "The Synergistic Impact of ENSO and IOD on Indian Summer Monsoon Rainfall in Observations and Climate Simulations-an Information Theory Perspective," *Earth System Dynamics*, vol. 11, no. 4, pp. 903–923, Nov. 2020, doi: 10.5194/esd-11-903-2020.

[5] E. Hermawan *et al.*, "Large-Scale Meteorological Drivers of Extreme Precipitation Event and Devastating Floods of Early February 2021 in Semarang, Indonesia," 2022. [Online]. Available: https://sharaku.eorc.jaxa.jp/GSMaP/

[6] F. Wang, M. Li, Y. Mei, and W. Li, "Time Series Data Mining: A Case Study with Big Data Analytics Approach," *IEEE Access*, vol. 8, pp. 14322–14328, 2020, doi: 10.1109/ACCESS.2020.2966553.

[7] H. Mohammed and A.-M. Al-Sharif, "Libyan Journal of Medical and Applied Sciences LJMAS Analysis and Evaluation of ARIMA and SARIMA Models Performance in Time Series Forecasting: An Applied Study," 2025.

[8] L. Zeng, Q. Chen, and M. Huang, "RSFD: A Rough Set-Based Feature Discretization Method For Meteorological Data," *Front Environ Sci*, vol. 10, Sep. 2022, doi: 10.3389/fenvs.2022.1013811.

[9] M. H. Santoso, "Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom," *Brilliance: Research of Artificial Intelligence*, vol. 1, no. 2, pp. 54–66, Dec. 2021, doi: 10.47709/brilliance.v1i2.1228.

[10] H. Chen, M. Yang, and X. Tang, "Association Rule Mining of Aircraft Event Causes Based on The Apriori Algorithm," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-64360-6.

[11] I. Ramli, S. Rusdiana, A. Achmad, Azizah, and M. E. Yolanda, "Forecasting of Rainfall Using Seasonal Autoregreressive Integrated Moving Average (SARIMA) Aceh, Indonesia," *Mathematical Modelling of Engineering Problems*, vol. 10, no. 2, pp. 501–508, Apr. 2023, doi: 10.18280/mmep.100216.

[12] S. O. Adams and M. Ardo Bamanga, "Modelling and Forecasting Seasonal Behavior of Rainfall in Abuja, Nigeria; A SARIMA Approach," *American Journal of Mathematics and Statistics*, vol. 2020, no. 1, pp. 10–19, 2020, doi: 10.5923/j.ajms.20201001.02.

[13] P. Kabbilawsh, D. S. Kumar, and N. R. Chithra, "Forecasting long-term monthly precipitation using SARIMA models," *Journal of Earth System Science*, vol. 131, no. 3, Sep. 2022, doi: 10.1007/s12040-022-01927-9.

[14] Gunawan, W. Andriani, and F. Z. Hidayatullah, "Penerapan Metode Association Rule Dan Algoritma Apriori Untuk Analisis Pola Frekuensi Tinggi Prediksi Curah Hujan Di Kota Tegal," *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, vol. 11, no. 2, pp. 45–53, Oct. 2023, doi: 10.21063/jtif.2023.v11.2.45-53.

[15] L. Coulibaly, B. Kamsu-Foguem, and F. Tangara, "Explainability with Association Rule Learning for Weather Forecast," *SN Comput Sci*, vol. 2, no. 2, Apr. 2021, doi: 10.1007/s42979-021-00525-8.

[16] D. Mircetic, S. Nikolicic, M. Maslaric, N. Ralevic, and B. Debelic, "Development of S-ARIMA Model for Forecasting Demand in a Beverage Supply Chain," *Open Engineering*, vol. 6, no. 1, pp. 407–411, 2016, doi: 10.1515/eng-2016-0056.

[17] G. Christie, D. Hatidja, and R. Tumilaar, "Penerapan Metode SARIMA dalam Model Intervensi Fungsi Step untuk Memprediksi Jumlah Pegunjung Objek Wisata Londa (Application of the SARIMA Method in the Step Function Intervention to Predict the Number of Visitors at Londa Tourism Object)," *JURNAL ILMIAH SAINS*, vol. 22, no. 2, p. 96, Aug. 2022, doi: 10.35799/jis.v22i2.40961.

[18] A. S. AlSalehy and M. Bailey, "Improving Time Series Data Quality: Identifying Outliers and Handling Missing Values in a Multilocation Gas and Weather Dataset," *Smart Cities*, vol. 8, no. 3, Jun. 2025, doi: 10.3390/smartcities8030082.

[19] T. M. Wanjuki, A. Wagala, and D. K. Muriithi, "Evaluating the Predictive Ability of Seasonal Autoregressive Integrated Moving Average (SARIMA) Models using Food and Beverages Price Index in Kenya," *European Journal of Mathematics and Statistics*, vol. 3, no. 2, pp. 28–38, Apr. 2022, doi: 10.24018/ejmath.2022.3.2.100.

[20] M. Faizan Tahir, K. Mehmood, M. Aamir, A. Wali Khan University Mardan, and P. Rizwan Raheem Ahmed, "The comparative Analysis of SARIMA, Facebook Prophet, and LSTM for Road Traffic Injury prediction in Northeast China."

[21] C. Maulana and N. Hajarisman, "Penerapan Transformasi Box Cox untuk Mengatasi Masalah Ketidakstasioneran dan Pola Periodik dalam Data Deret Waktu pada Ekspor Bidang Pertanian di Indonesia," *Bandung Conference Series: Statistics*, vol. 3, no. 2, pp. 763–772, Aug. 2023, doi: 10.29313/bcss.v3i2.9371.

[22] M. Othman, R. Indawati, A. A. Suleiman, M. B. Qomaruddin, and R. Sokkalingam, "Model Forecasting Development for Dengue Fever Incidence in Surabaya City Using Time Series Analysis," *Processes*, vol. 10, no. 11, Nov. 2022, doi: 10.3390/pr10112454.

[23] I. Mahib Zuhair Riyanto *et al.*, "Forecasting the Number of Passengers for the Jakarta-Bandung High-Speed Rail using SARIMA and SSA Models," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[24] L. Martínez-Acosta, J. P. Medrano-Barboza, Á. López-Ramos, J. F. R. López, and Á. A. López-Lambraño, "SARIMA Approach to Generating Synthetic Monthly Rainfall in The Sinú River Watershed in Colombia," *Atmosphere (Basel)*, vol. 11, no. 6, Jun. 2020, doi: 10.3390/atmos11060602.

[25] N. S. Poli and A. S. Sikder, "Predictive Analysis of Sales Using the Apriori Algorithm: A Comprehensive Study on Sales Forecasting and Business Strategies in the Retail Industry.," *International Journal of Imminent Science & Technology.*, vol. 1, no. 1, pp. 1–16, Nov. 2023, doi: 10.70774/ijist.v1i1.1.

[26] Y. Kaya and R. Tekin, "Comparison of Discretization methods for Classifier Decision Trees and Decision Rules on Medical Data Sets," *European Journal of Science and Technology*, Mar. 2022, doi: 10.31590/ejosat.1080098.

[27] E. L. Limahelu and B. Herwanto, "Buletin Stasiun Meteorologi Umbu Mehang Kunda Sumba Timur," Jun. 2020.