American Journal of Integrated STEM Education, 2026, 10-20
https://publishingjournals.org/stem
Volume: 3 | Number: 4 (2026) October

Modeling and Forecasting Monthly Rainfall Quantities Using the SARIMA Model: An Applied Study at the Central Park Station in New York

Alaa Farhan Ahmed Azawi*
Department of Mathematics, College of Basic Education, Al-Shirqat, University of Tikrit, Iraq

***Correspondence:**
Alaa Farhan Ahmed Azawi
alaa.f@tu.edu.iq

Received: 14 Jul 2026
Accepted: 20 Aug 2026
Published: 19 Sep 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/ 4.0/
Open Access

Abstract
This study aims to model and forecast the temporal and seasonal behaviour of monthly rainfall amounts using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The study drew on 552 monthly rainfall observations from the New York City Central Park station in New York City, covering the period January 1980 to December 2025, based on NOAA/NCEI data. The analysis revealed statistically significant differences between months, with a continued need to represent the seasonal structure in the model. After examining stationarity, autocorrelation and partial autocorrelation functions, and estimating a set of candidate models, the model was selected as the final model based on information criteria, residual diagnostics and out-of-sample forecasting performance. The data were split into a training sample covering the period 1980–2022 and a test sample covering the period 2025–2023. In the test sample, the model achieved an MAE of 48.52 mm and an RMSE of 66.2 mm, whilst the MASE calculated for the seasonal perspective was 0.749 and the Ljung–Box test showed no significant autocorrelation in the residuals at lags 12, 24 and 36. The model was also used to produce monthly forecasts for the period 1/2027–12/2027.

Keywords
Seasonal Time Series, SARIMA, Rainfall Forecasting, Seasonal Difference, Autocorrelation.

Introduction
Rainfall is one of the most important climatic variables affecting ecological and economic systems and water resource management; understanding its temporal behaviour is therefore a key factor in planning and decision-making [1]. Rainfall time series are characterised by marked inter-monthly variability and the potential for recurring seasonal patterns; this makes the

development of statistical models capable of capturing both temporal dependence and seasonality a key issue in applied studies [2] [3].

The research problem lies in the difficulty of representing monthly and seasonal variations in rainfall within a single statistical model, and the need to identify a suitable structure for a seasonal model that can describe the series and be used for forecasting. The challenge lies in determining the need for normal and seasonal components, selecting the orders of the non-seasonal and seasonal components, and then verifying the model's suitability and its ability to make out-of-sample predictions.

The study aims to analyse the statistical and temporal characteristics of monthly rainfall at the Central Park station, verify the presence of seasonality and stationarity suitable for analysis, then to estimate candidate SARIMA models and select the best model based on information criteria, residual analysis and forecast accuracy measures, and to use the final model to produce future forecasts.

The study is based on the hypotheses that the monthly series contains measurable seasonal variations[4] [5], that the seasonal component may be necessary to obtain a stationary series, that a suitable SARIMA model can represent the time structure of the series and produce forecasts of acceptable accuracy[6], and that the residuals of the best model do not exhibit statistically significant autocorrelation[7].

The general form of the model

$$SARIMA(p,d,q)(P,D,Q)_s$$

The mathematical formulation of the model is therefore

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

Methods

The methodology of the study comprises a series of sequential stages, beginning with a descriptive and exploratory analysis of the data, followed by an examination of stationarity and the use of autocorrelation and partial autocorrelation functions to identify possible model structures, followed by the estimation of a set of SARIMA models and their comparison using criteria such as AIC and BIC, whilst conducting the necessary diagnostic tests on the residuals. Once the most appropriate model has been identified, its predictive ability is assessed using independent test data and metrics such as MAE, RMSE and MASE, after which the selected model is used to generate future forecasts.

The theoretical aspect of the SARIMA model

The SARIMA model is an extension of the ARIMA model designed for time series that include a seasonal component. It consists of a non-seasonal component, determined by the orders p, q and d, and a seasonal component, determined by the orders P, D and Q, whilst S determines the length of the seasonal cycle[8].

The general form of the model is

$$SARIMA(p,d,q)(P,D,Q)_s$$

Where

p: the order of the non-seasonal autoregressive (AR) component.

d: the number of times the non-seasonal difference is applied to stabilise the series.

q: the order of the non-seasonal moving average (MA) component.

P: the order of the seasonal autoregressive component.

D: the number of times the seasonal difference is applied to stabilise the series.

Q: the order of the seasonal moving average component.

s: the length of the seasonal period.

The displacement and difference effects[9].

$$\mathrm{B\ Y_t = Y_{t-1}}\ (2)$$

Where:

B: the lag term.

Y_t: the value of the series at time t.

k: the number of lags.

$$\nabla Y_t = (1 - B)Y_t$$
$$\nabla Y_t = Y_t - Y_{t-1}$$
$$\nabla_s Y_t = (1 - B^s)Y_t$$
$$\nabla_s Y_t = Y_{t-s}\ (3)$$

d represents the number of non-seasonal differences required to achieve stationarity, whilst D represents the number of seasonal differences required to remove the instability associated with the seasonal cycle. In the case of the present study, the seasonal difference is of the first order, i.e. D = 1 and S = 12; therefore:

$$(1 - B^{12})Y_t = Y_t - Y_{t-12}\ (4)$$

The autoregressive component and non-seasonal moving averages

The autoregressive component is based on the previous values of the series, and the model is of order p, given by: [10]

$$Y_t = c + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + \varepsilon_t\ (5)$$

The autocorrelation component in the form of a shift term is given by:

$$\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p\ (6)$$

Where

c: the model constant.

Q_i: the autoregressive coefficients.

p: the order of the AR model.

ε_t: the random error at time t

Whilst the moving average component depends on the current and previous errors, the MA(q) model takes the form [11]

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}\ (7)$$

It takes the form of a displacement operator as follows:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\ (8)$$

The orders q and p are selected based on the behaviour of the ACF and PACF and the other model selection criteria [12]

Autoregressive and Seasonal Moving Average Components

The seasonal component AR is based on previous values occurring at multiples of the seasonal period s, and its general form is as follows: [13].

$$\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS} \quad (9)$$

Where:

P: Order of the seasonal AR model.

Φ_i: Seasonal AR coefficients.

B^s: Shift equal to one full season.

s: Length of the seasonal period.

The seasonal moving average component measures the influence of past errors at seasonal shifts, and its formula is as follows: [14].

$$\Theta_Q(B^S) = 1 + \Theta_1 B^S + \Theta_2 B^{2S} + \cdots + \Theta_Q B^{QS} \quad (10)$$

Where

Q: Seasonal MA order.

Φ_i: Seasonal MA coefficients.

B^s : Seasonal shift.

s : Season length.

General Form of the Model

Having identified all the components of the model, the general form of the model is as follows: [15] [16].

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D Y_t = \Theta_Q(B)^s \theta_q(B)\varepsilon_t \quad (11)$$

Where

Φ_P (B^s): seasonal autoregression.

ϕ_P (B): non-seasonal autoregression.

$(1\text{-}B)^d$: non-seasonal difference.

$(1\text{-}B^s)^D$: seasonal difference.

Θ_Q $(B)^s$: seasonal moving averages.

θ_q (B): non-seasonal moving averages.

ε_t: the random error at time t.

4. Data description and analysis methodology

The study was based on monthly precipitation data from the NY City Central Park station in New York City, covering the period from January 1980 to December 2025.

The dataset consists of 552 consecutive monthly observations, and the annual season length for the monthly data is 12 months. The data were divided into a training sample spanning January 1980 to December 2022 and a test sample spanning January 2023 to December 2025, in order to assess the model's predictive ability on out-of-sample data.

Description of the data and analysis methodology

The study was based on monthly precipitation data from the NY City Central Park station in New York City, covering the period from January 1980 to December 2025.

The dataset consists of 552 consecutive monthly observations, and the annual season length for the monthly data is 12 months. The data were divided into a training sample covering the period from January 1980 to December 2022 and a test sample covering the period from January 2023 to December 2025, in order to assess the model's predictive ability on out-of-sample data.

$$n = (2025 - 1980 + 1) \times 12 = 552 \ (12)$$

$$s = 12$$

Statistical analysis of the data

The statistical analysis included a descriptive analysis of the time series, an examination of the mean, standard deviation and variance, and an analysis of the monthly means to identify seasonal patterns. Tests were carried out to determine whether there were differences between months, and stationarity was examined using the ADF and KPSS tests after applying seasonal adjustment where necessary. Subsequently, the autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to generate candidate models, The models were then compared using AIC and BIC, the residuals were assessed using the Ljung-Box test, and finally, predictive ability was evaluated using MAE, MASE and seasonal RMSE.

Descriptive statistics

The descriptive analysis showed that the mean monthly rainfall was 105.75 mm, whilst the median was 97 mm and the standard deviation was 60.95 mm; the range was 3,714.91, with values ranging from 0.3 mm to 481.3 mm. The coefficient of variation was 57.64 per cent, indicating a significant level of variability in the monthly series.

Table 1: Descriptive statistics for the data

The Scale	Value
Number of observation	552
Mean	105.75
Median	97
Standard Deviation	60.95
Variance	3714.91
Minimum	0.3
Maximum	481.3
Coefficient of variation	57.64%

Analysis of Seasonality

The monthly averages showed a clear variation in precipitation levels; the average for February was the lowest at 79.01 mm, whilst July recorded the highest monthly average at 123.23 mm, followed by August at 118.31 mm. To verify the significance of the differences between the months, the Kruskal-Wallis test was used. The test statistic was H = 21.94 with P = 0.0248, which is

less than 0.05. This provides statistical evidence of differences between the months and justifies examining the seasonal component in the forecast model.

$$H = [12/(N(N+1))]\sum(R_j^2/n_j) - 3(N+1) \quad (13)$$

Table 2: Monthly average rainfall (mm)

The Month	Average rainfall (mm)
January	90.63
February	79.01
March	111.83
April	113.36
May	108.16
June	105.24
July	123.23
August	118.31
September	110.55
October	108.38
November	95.31
December	105.01

Stability and Autocorrelation Tests

The ADF test on the original series yielded a significant negative value of −21.538 with P < 0.001, indicating that the hypothesis of a unit root in the original series was rejected. However, a further test was conducted for seasonal variation; for the seasonally adjusted series, the ADF value was –10.158 with P < 0.001, whilst the KPSS value was approximately 0.02198 with P > 0.1, supporting the stationarity of the series following seasonal adjustment.

$$\Delta Y_t = \alpha + \beta_t + \gamma Y_t - 1 + \sum \delta_i \Delta Y_t - i + \varepsilon_t \quad (14)$$

$$H_0: \gamma = 0$$

$$H_1: \gamma < 0$$

As for the autocorrelation function, it showed a weak correlation at the first lag (0.0849), and the correlation at a 12-month seasonal lag was also very weak (0.0042). For this reason, the model specification was not determined based on the ACF and PACF alone; rather, these indicators were combined with information criteria and out-of-sample evaluation.

$$\rho k = \sum[(Y_t - \bar{Y})(Y_t - k - \bar{Y})]/\sum(Y_t - \bar{Y})^2 \quad (15)$$

Model building and selection of the best model

After examining the characteristics of the time series, a set of seasonal models of varying orders was estimated, with a focus on those models that achieve an acceptable residual fit and good predictive ability. The comparison included models with different values for P and q, with seasonal components P and Q. The need for the seasonal difference D was also examined, and the comparison showed that models with D = 1 achieved better informational and predictive fit than the main alternatives with D = 0.

$$AIC = -2\ln(L) + 2k \quad (16)$$

$$BIC = -2\ln(L) + k\ln(n) \quad (17)$$

Table 3: A set of seasonal models of various orders.

The Modle	AIC	BIC	MAE	RMSE	Liung-Box p(12)
$SARIMA(0,0,0)(0,1,1)_{12}$	5606.81	5615.26	48.62	66.23	0.094
$SARIMA(0,0,1)(0,1,1)_{12}$	5605.97	5618.64	48.52	66.2	0.465
$SARIMA(1,0,0)(0,1,1)_{12}$	5605.76	56.18.43	48.53	66.21	0.525
$SARIMA(0,0,1)(1,0,1)_{12}$	5752.26	5769.24	48.27	66.3	0.475

Although the SARIMA(1,0,0)(1,0,1)$_{12}$ model achieved a lower MAE in the test sample, the D=1 models achieved a significantly lower AIC, Furthermore, the SARIMA(0,0,1)(0,1,1)$_{12}$ model achieved the lowest RMSE amongst the main models with acceptable residuals; in addition, the Ljung-Box test did not indicate any residual autocorrelation. Based on a balance between model fit, forecast accuracy and residual analysis, the SARIMA(0,0,1)(0,1,1)$_{12}$ model was adopted as the final model.

The estimated equation for the final model

Following analysis of the study data, the SARIMA(0,0,1)(0,1,1)$_{12}$ model was selected, that is,

$$S = 12 \text{ و } Q = 1 \text{ و } D = 1 \text{ و } P = 0 \text{ و } q = 1 \text{ و } d = 0 \text{ و } p = 0$$

The equation then becomes

$$(1 - B^{12})Y_t = (1 + \theta_1 B)(1 + \Theta_{12} B^{12})\varepsilon_t \quad (18)$$

When the selected model was re-estimated using the full time series for the period 1980–2025, the value of the first-order non-seasonal moving average coefficient $\hat{\theta}_1$ was 0.0713, whilst the value of the twelfth-order seasonal moving average coefficient $\hat{\Theta}_{12}$ was −0.9984, The small value of the non-seasonal moving average coefficient indicates the limited impact of short-term shocks from the previous period, whilst the high absolute value of the seasonal coefficient reflects the significance of the annual seasonal component in the model's time structure. When estimated on the training sample, the p-value for the MA(1) coefficient was approximately 0.0598, which is slightly greater than the 5 per cent significance level, indicating that its individual significance was marginal, whilst the seasonal coefficient was statistically significant at the 5 per cent level. However, model selection was not based solely on the significance of individual coefficients, but rather on a combination of model selection criteria, out-of-sample forecasting performance, and residual diagnostics, including the Ljung-Box test.

Substituting the value of the first-order non-seasonal moving average coefficient and the twelfth-order seasonal moving average coefficient into Equation (18) yields

$$(1 - B^{12})Y_t = (1 + 0.0713\ B)(1 - 0.9984\ B^{12})\varepsilon_t \quad (19)$$

Model Diagnosis and Residuals

The residuals from the final model were examined to verify the absence of any residual temporal structure that could be modelled, using the p-values at lags 12, 24 and 36 were approximately 0.465, 0.226 and 0.326 respectively, all of which are greater than 0.05, indicating no statistical evidence of residual autocorrelation in the residuals. In contrast, the Jarque–Bera test revealed non-normality in the residuals, which is to be expected in rainfall data that contain

outliers and high values; therefore, the forecast horizons must be interpreted as approximate within the model's assumptions.

$$e_t = Y_t - \hat{Y}_t \quad (20)$$

$$Q(h) = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \quad (21)$$

$H_0: \rho_1 = \rho_2 = \cdots = \rho_h = 0$

$H_1: There\ is\ autocorrelation\ in\ the\ residuals$

Table 4: Results of the final model residual analysis

The Test	Value	Interpretation
Ljung-Box(12)	$p = 0.456$	There is no significant autocorrelation
Ljung-Box(24)	$p = 0.226$	There is no significant autocorrelation
Ljung-Box(36)	$p = 0.326$	There is no significant autocorrelation
Jarque-Bera	$p < 0.001$	The residuals are abnormal

Discussion

After finalising the model, the period 1980–2022 was used for the initial estimation and out-of-sample evaluation over the period 2023–2025, after which the model was re-estimated over the entire 1980–2025 time series to produce future forecasts. Accuracy was assessed using MAE, MASE and seasonal RMSE, with lower values of these metrics indicating improved predictive performance.

$$\hat{Y}_{t+h|t} = E(Y_{t+h}|I_t) \quad (22)$$

where

I_t: the information available up to time t

$$MAE = \frac{\sum_{t=1}^{n}|Y_t - \hat{Y}_t|}{n}$$

$$RMSE = \sqrt{(1/n)\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}$$

$$MASE = \frac{MAE}{\frac{1}{n-1}\sum_{t=2}^{n}|Y_t - Y_{t-1}|}$$

The MAE on the test sample was 48.52 mm, and the RMSE was 66.2 mm, whilst the seasonal MASE was 0.749; a MASE value of less than one indicates that the model achieves a lower mean absolute error than the seasonal reference forecast based on the value from 12 months previously.

$$\hat{Y}_{t+h} \pm 1.96\ SE(\hat{Y}_{t+h})$$

Forecasts covering a 12-month period from January 2027 to December 2027 have been produced.

Table 5: 12-month rainfall forecasts (January 2027 to December 2027) and the upper and lower limits of rainfall amounts

Year	Month	Expected value (mm)	Minimum limit (mm)	Maximum limit (mm)
2027	January	90.59	0	210.28
2027	February	79	0	198.69
2027	March	111.82	0	231.5
2027	April	113.35	0	233.03
2027	May	108.15	0	227.84
2027	June	105.22	0	224.91
2027	July	123.23	3.54	242.91
2027	August	118.31	0	237.99
2027	September	110.54	0	230.23
2027	October	108.37	0	228.06
2027	November	95.29	0	214.97
2027	December	104.95	0	224.63

These forecasts are contingent upon the persistence of the temporal and climatic patterns observed during the study period.

Conclusion

The study showed that the rainfall series at the City Central Park station for the period 1980–2025 is characterised by significant variability, with a mean forced precipitation of 105.75 mm and a coefficient of variation of 57.64 per cent.

A stationarity test indicated that the series is suitable in terms of its general characteristics, and that the first-order seasonal difference with a season length of 12 months provides an adequate representation of the seasonal structure used in the model.

The SARIMA (0,0,1)(0,1,1)_12 model was adopted following a comparison of several seasonal models, an analysis of the residuals, and an out-of-family performance assessment.

The model achieved an MAE of 48.52 mm, an RMSE of 66.2 mm, and a seasonal MASE of 0.749.

The Ljung–Box test showed no residual autocorrelation at lags 12, 24 and 36, which supports the validity of the model in terms of the residual time series structure, whilst noting that the residuals do not follow a normal distribution well.

The model produced monthly forecasts for the period January 2027 to December 2027, which can be used as a statistical baseline for planning and analysis, whilst taking into account the uncertainty in the forecast.

The researcher recommends conducting a future comparison between SARIMA and hybrid models or machine learning algorithms, using the same test period to ensure a fair comparison.

When using forecast periods in practical applications, it should be borne in mind that the limits resulting from Gaussian rounding may be negative; it is therefore recommended to use methods or transformations specific to positive variables.

References

[1] E. O'Connell, "Towards adaptation of water resource systems to climatic and socio-economic change," *Water Resources Management*, vol. 31, no. 10, pp. 2965–2984, 2017.

[2] R. Wiegels, C. Chwala, J. Polz, L. Glawion, C. Lorenz, T. C. Schober, and H. Kunstmann, "Improved seasonal precipitation forecasts for the Blue Nile Basin: A deep learning approach," *Frontiers in Climate*, vol. 8, p. 1691030, 2026.

[3] S. Ezati, M. Hashemzadeh, F. Borhani, M. Talaie, and F. Mazinani, "Hybrid model-based forecasting of temperature and precipitation changes in Iran," *Journal of Atmospheric and Solar-Terrestrial Physics*, p. 106852, 2026

[4] M. Ruth and A. C. Lin, "Regional energy demand and adaptations to climate change: Methodology and application to the state of Maryland, USA," *Energy Policy*, vol. 34, no. 17, pp. 2820–2833, 2006.

[5] S. Pezzulli, D. B. Stephenson, and A. Hannachi, "The variability of seasonality," *Journal of Climate*, vol. 18, no. 1, pp. 71–88, 2005.

[6] M. Majka, "Seasonal time series analysis: Why SARIMA outshines ARIMA," *ResearchGate*, Sep. 19, 2024.

[7] C. F. Dormann, J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. D. Kissling, I. Kühn, R. Ohlemüller, P. R. Peres-Neto, B. Reineking, B. Schröder, F. M. Schurr, and R. Wilson, "Methods to account for spatial autocorrelation in the analysis of species distributional data: A review," *Ecography*, vol. 30, no. 5, pp. 609–628, 2007.

[8] O. Ebhuoma, M. Gebreslasie, and L. Magubane, "A seasonal autoregressive integrated moving average (SARIMA) forecasting model to predict monthly malaria cases in KwaZulu-Natal, South Africa," *SAMJ: South African Medical Journal*, vol. 108, no. 7, pp. 573–578, 2018.

[9] B. Kanguzhin and B. Koshanov, "Uniqueness criteria for solving a time nonlocal problem for a high-order differential operator equation l(·)—A with a wave operator with displacement," *Symmetry*, vol. 14, no. 6, p. 1239, 2022.

[10] T. Ullrich, "On the autoregressive time series model using real and complex analysis," *Forecasting*, vol. 3, no. 4, pp. 716–728, 2021.

[11] H. R. Alsamamra, S. Salah, and J. H. Shoqeir, "Performance analysis of ARIMA model for wind speed forecasting in Jerusalem, Palestine," *Energy Exploration & Exploitation*, vol. 42, no. 5, pp. 1727–1746, 2024.

[12] H. Hassani, L. M. Mashhad, M. Royer-Carenzi, M. R. Yeganegi, and N. Komendantova, "White noise and its misapplications: Impacts on time series model adequacy and forecasting," *Forecasting*, vol. 7, no. 1, p. 8, 2025.

[13] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," *International Journal of Forecasting*, vol. 36, no. 1, pp. 75–85, 2020.

[14] U. H. Perez-Guerra, R. Macedo, Y. P. Manrique, E. A. Condori, H. I. Gonzáles, E. Fernández, and M. García-Herreros, "Seasonal autoregressive integrated moving average (SARIMA)

time-series model for milk production forecasting in pasture-based dairy cows in the Andean highlands," *PLoS ONE*, vol. 18, no. 11, e0288849, 2023.

[15] S. Ray, S. S. Das, P. Mishra, and A. M. G. Al Khatib, "Time series SARIMA modelling and forecasting of monthly rainfall and temperature in the South Asian countries," *Earth Systems and Environment*, vol. 5, no. 3, pp. 531–546, 2021.

[16] A. C. Gabriel, "A SARIMA and adjusted SARIMA models in a seasonal nonstationary time series; evidence of Enugu monthly rainfall," *European Journal of Mathematics and Statistics*, vol. 2, no. 1, pp. 13–18, 2021.