Time Series Analysis Course

Cambodia Rainfall Forecasting with Focus on Pursat Province

ARIMA / SARIMA Time Series Modeling

Group 01

Name	ID
1. CHEA Piseth	e20210871
2. HENG Sopanha	e20210931
3. KHUN Sithanut	e20211527
4. KOSAL Chansothay	e20210574

Instructor: Dr. SIM Tepmony

Date: 24 January 2026

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia

Contents

List of Figures

List of Tables

Abstract

Cambodia's agriculture is strongly dependent on monsoon rainfall, and fluctuations in seasonal precipitation directly affect crop yields, food security and water resources. This project conducts an end–to–end time series analysis to model and forecast national monthly rainfall for Cambodia using classical autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) models. The analysis uses CHIRPS–based subnational rainfall indicators from the Humanitarian Data Exchange (HDX), spanning 1981–2026, and aggregates daily/dekadal provincial data to a national monthly series.

Exploratory analysis, seasonal decomposition and stationarity testing show a strong annual monsoon cycle and seasonal non–stationarity. Seasonal differencing at lag 12 is applied, followed by autocorrelation (ACF) and partial autocorrelation (PACF) analysis to guide ARIMA/SARIMA order selection. A rigorous evaluation framework is adopted: the last 60 months are held out as a test set, and all models are evaluated with rolling one–step–ahead forecasts.

Baseline AR, MA, ARMA and non–seasonal ARIMA models are compared with seasonal SARIMA specifications. A SARIMA$(0, 0, 1) \times (1, 1, 1)_{12}$ model achieves the best performance, with RMSE of approximately 1246 mm, MAE around 922 mm, and R^2 about 0.88 on the test horizon, while residual diagnostics indicate no remaining autocorrelation. The selected SARIMA model is finally refitted on the full series and used to generate a 24–month ahead forecast with 95% prediction intervals, preserving the monsoon pattern and providing information useful for medium–term agricultural planning.

1 Introduction

Cambodia lies in mainland Southeast Asia and experiences a tropical monsoon climate with a pronounced wet and dry season. The majority of cropland is rain–fed, so interannual variability in rainfall directly influences planting dates, yields and rural livelihoods. Increasing climate variability and extreme events have raised concerns about agricultural resilience and water security.

On seasonal to annual horizons, statistical time series models remain attractive because they are interpretable, require modest data and are computationally efficient. Classical ARIMA and SARIMA models, introduced in the Box–Jenkins framework, are widely used for stationary and seasonally stationary series, and they provide a strong baseline even in the era of deep learning. [1, 2]

This report develops and evaluates univariate time–series models for national monthly rainfall in Cambodia. The workflow includes: data acquisition from HDX, spatio–temporal aggregation, exploratory data analysis, stationarity testing, model identification, parameter estimation, out–of–sample validation via rolling forecasts, residual diagnostics, and multi–step forecasting.

1.1 Objectives

The main objectives are:

- To construct a high–quality national monthly rainfall time series for Cambodia from CHIRPS–based subnational data.
- To characterise the temporal structure (trend, seasonality, variability) of this series.
- To develop and compare AR, MA, ARMA, ARIMA and SARIMA models under a rigorous train–test protocol with no data leakage.
- To select a parsimonious SARIMA model with good predictive accuracy and well–behaved residuals.
- To generate 24–month ahead forecasts with uncertainty bands for potential use in agricultural and water–resource planning.

1.2 Report Structure

The report is organised as follows. Section 2 reviews relevant literature on ARIMA/SARIMA modeling, graphical–model extensions, time–series preprocessing and climate applications, including references on ARIMA vs. LSTM. Section 3 describes the Cambodia rainfall dataset, preprocessing steps and construction of the national monthly series. Section 4 presents theoretical background on ARIMA/SARIMA and the Box–Jenkins methodology. [1, 2] Section 5 reports exploratory data analysis, decomposition and stationarity testing. Section 6 explains the modeling strategy, train–test split and rolling forecast

evaluation, while Section 7 presents model comparison and residual diagnostics. Section 9 presents the 24–month forecast, and Section 11 discusses conclusions and future research directions.

2 Literature Review

2.1 Classical ARIMA and SARIMA Modeling

Cryer and Chan provide a comprehensive treatment of time series analysis with applications in R, covering AR, MA, ARMA, ARIMA and SARIMA models, the Box–Jenkins identification–estimation–diagnostic cycle, and model selection via AIC and BIC. [1] They emphasise covariance stationarity, invertibility, differencing strategies, and the use of ACF and PACF in distinguishing AR and MA components. [1]

Velicer and Fava review time–series methods for psychological research, focusing on ARIMA models, dependency and autocorrelation, interrupted time–series intervention analysis, and pooled time–series for generalisation. [2] They discuss model identification difficulties, the importance of adequate sample size (e.g. $n > 50$ for stable autocorrelations), and alternative approaches such as general transformation matrices to handle dependence without full ARIMA identification. [2]

Both sources support the use of SARIMA$(p, d, q) \times (P, D, Q)_s$ for strongly seasonal monthly climate variables, recommending seasonal differencing and low–order seasonal AR and MA terms. [1, 2]

2.2 Graphical–Model Extensions of ARMA / ARIMA

Thiesson et al. express ARMA time–series models as directed graphical models, revealing the conditional independence structure between observations and noise terms and enabling Bayesian network inference for forecasting and parameter estimation with missing data. [3] They show that deterministic relations in standard ARMA make EM infeasible, and propose a stochastic ARMA variant (ARMA$^{\epsilon}$) where deterministic equations are replaced by Gaussian distributions with small variance, allowing EM–based learning and improving predictive accuracy via smoothing. [3]

Their ARMA$^{\text{xp}}$ model introduces cross–predictors from related series and arbitrary non–temporal covariates, generalising vector ARMA while permitting series–specific AR/MA orders and selective cross–lag structure. [3] This framework suggests future extensions for Cambodian rainfall, such as modeling multiple provinces with cross–predictors or incorporating climate indices, while keeping an ARIMA core.

2.3 Time–Series Data Preparation and Feature Engineering

Baumann et al. discuss an introductory approach to time–series data preparation and analysis, focused on prognostics and health management (PHM) for vehicles but broadly applicable. [4] They emphasise common issues in real–world time–series data such as unordered or inconsistent timestamps, sampling–rate mismatches, missing values, noise, out–of–range data, and type shifts, and they advocate a structured pipeline including data cleaning, transformation, synchronization and feature selection. [4]

Their taxonomy classifies signals as constant, binary, low–state, mid–state, high–state, diagnostic and utility, advising that mid/high–state signals typically carry most information for complex models, while constants and diagnostics serve as metadata. [4] They demonstrate the use of low–pass filters for noise reduction, IQR and 3σ rules for outlier detection, interpolation for alignment, entropy–based feature selection, and spectral transforms such as the FFT for additional features. [4]

Although their examples come from vehicle sensor data, the principles apply to rainfall: verifying time consistency, aggregating to a suitable temporal resolution (months), examining outliers, handling missing values carefully, and using exploratory statistics to guide model building.

2.4 Climate and Environmental Time Series

Herho and Firdaus present a pilot study on time–series analysis and statistical forecasting of daily rainfall in Kupang, East Nusa Tenggara, Indonesia. [6] They fill missing values using piecewise cubic Hermite interpolation (PCHIP), analyse monthly patterns and periodicity using continuous wavelet transform (CWT), compute a rainfall anomaly index (RAI) as a drought/wetness indicator, and then model and forecast daily RAI using the Prophet algorithm (a GAM–based model). [6]

Their CWT analysis reveals strong annual periodicity dominated by monsoon cycles and weaker interannual signals potentially related to Madden–Julian Oscillation (MJO), while ENSO signals at 2–8 year scales are less apparent in daily data due to annual dominance. [6] Prophet captures a linear trend toward increasing drought and reproduces annual seasonality but struggles with longer–scale variability, highlighting trade–offs between flexible additive models and targeted seasonal ARIMA structures. [6]

Another study on temperature forecasting using SARIMA demonstrates that seasonal ARIMA models with yearly period can successfully capture annual cycles and trends in temperature data, offering a close analogue to rainfall and reinforcing the choice of SARIMA for hydro–climatic series. [7]

Herho and colleagues also illustrate how R and Python workflows (PyGMT for mapping, PyCWT for CWT, “precintcon” for rainfall indices, Prophet for forecasting) can be combined with statistical modeling, suggesting practical multi–tool pipelines similar

to the present Cambodia project. [6]

2.5 ARIMA, LSTM and Hybrid Forecasting

Liu compares ARIMA and LSTM models for stock price forecasting, reviewing classical linear models (AR, MA, ARMA, ARIMA, ARCH/GARCH) and noting their assumptions of linearity and stationarity. [5] She highlights that modern financial series often display nonlinearity and heteroskedasticity, motivating neural networks such as ANN and LSTM, which can model complex, long–term dependencies and handle noisy sequences. [5]

Using Walmart stock prices from 2020–2021, Liu shows that a customised ARIMA(1,1,10) outperforms basic and multi–step LSTM in RMSE for short–horizon forecasts, while LSTM models better capture multi–day trend direction. [5] The study concludes that ARIMA remains competitive and more interpretable for short–term predictions, whereas LSTM offers advantages for longer–term trend forecasting but requires more complex tuning and is prone to overfitting. [5]

More generally, other authors (e.g. Siami–Namini et al., cited in Liu) find that LSTM often surpasses ARIMA on highly nonlinear series, yet ARIMA/SARIMA remain useful baselines and are easier to diagnose via residual analysis. [5] For Cambodia rainfall, this literature supports using SARIMA as a transparent baseline, with future work possibly exploring LSTM or hybrid SARIMA–LSTM models once a solid classical benchmark is established.

2.6 Additional AR and Time–Series References

An article on autoregressive time–series modeling using real data discusses AR(p) estimation, order selection and forecasting, stressing that even simple AR models require careful diagnostics and that higher orders rarely improve performance beyond $p = 3$ or 4. [8] A chapter by Velicer and Fava also covers multivariate time–series extensions such as cross–lagged correlations, dynamic factor models and structural–equation–modeling–based time–series, and discusses generalisability, pooled designs and meta–analysis for single–subject series. [2]

These broader references highlight common themes: the importance of choosing an appropriate sampling interval, addressing missing data and dependence correctly, avoiding overfitting by keeping models simple, and prioritising interpretability, all of which are reflected in this project's design. [2, 1]

3 Data Description and Preprocessing

3.1 Data Source

The rainfall data used in this study come from the Humanitarian Data Exchange (HDX) dataset titled *"Cambodia: Rainfall Indicators at Sub–national Level"*, derived from the CHIRPS precipitation product. CHIRPS combines satellite–based estimates with station observations to provide quasi–global rainfall fields at fine spatial and temporal resolution.

The HDX dataset includes, among others, the following fields:

- `date`: observation date (daily or dekadal).
- `admlevel`: administrative level (1 = province).
- `admid`, `PCODE`: administrative identifiers.
- `npixels`: number of contributing grid cells.
- `rfh`: rainfall amount in mm.
- `rfhavg`, `r1h`, `r3h`, `rfq`, `r1q`, `r3q`: intensity and quantile indicators.
- `version`: data version flag.

The raw dataset contains 358,241 rows and 15 columns, covering all 25 provinces over multiple decades.

3.2 Cleaning and Selection

The `date` column is converted to a proper `datetime` type, and the dataset is filtered to administrative level 1 (`admlevel = 1`) to focus on provincial aggregates. Dataset information (types, non–null counts, memory usage) and summary statistics are computed to check for missing values and plausible numeric ranges across variables such as `rfh`, `r1h`, `r3h` and quantile fields.

For certain exploratory tasks, a representative province (with the most complete data) is selected, but the main modeling is based on a national aggregate constructed from all provinces. This national focus aligns with the project objective of country–level rainfall forecasting for policy and planning.

3.3 Handling Missing Values

Missing values occur sporadically in some rainfall and intensity fields due to observational gaps or processing artefacts. For the provincial series used to construct the national aggregate, missing rainfall values are imputed using forward–fill followed by backward–fill, which is suitable given short gaps and the non–negative nature of rainfall. This simple imputation strategy, similar in spirit to approaches in Baumann et al., avoids introducing unrealistic extremes while maintaining temporal continuity. [4]

After imputation, the count of missing values in the key rainfall series is zero, indicating a complete input for monthly aggregation and modeling.

3.4 Temporal Aggregation and National Series

To match medium–term planning horizons and capture seasonal structure, daily/dekadal provincial rainfall is aggregated to monthly totals. Let $Y_{p,d}$ denote rainfall for province p on date d; the national total on date d is

$$R_d = \sum_{p=1}^{P} Y_{p,d},$$

and monthly rainfall for month m is

$$R_m = \sum_{d \in m} R_d.$$

This yields a national monthly rainfall series spanning January 1981 to January 2026, with 541 observations at month–start frequency (`MS`). The series is stored as a Pandas `Series` with a `DatetimeIndex`, denoted $\{R_t\}_{t=1}^{T}$ with $T = 541$.

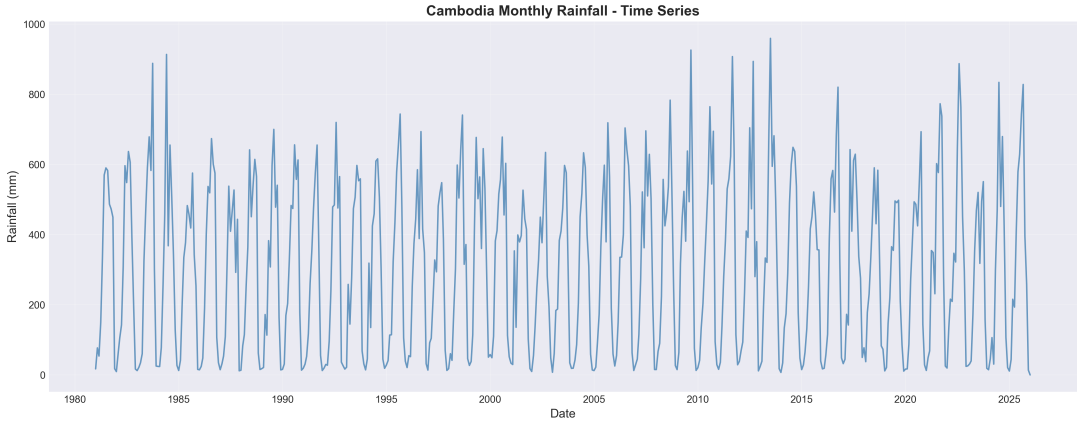

Figure 1: National monthly rainfall series for Cambodia, 1981–2026.

Basic descriptive statistics, including mean, median, standard deviation, quartiles, range, coefficient of variation, skewness and kurtosis, are computed to characterise overall variability.

4 Theoretical Background

4.1 Time Series and Stationarity

A (discrete) time series is a sequence of random variables $\{Y_t\}$ indexed by time $t = 1, 2, \ldots, T$. It is often conceptualised as

$$Y_t = f(t) + \varepsilon_t,$$

where $f(t)$ represents systematic components (trend, seasonality) and ε_t is a random noise term. [1] For ARIMA modeling, covariance stationarity is usually required: constant mean, constant variance and autocovariance depending only on lag h. [1]

The autocovariance function at lag h is

$$\gamma(h) = \mathrm{Cov}(Y_t, Y_{t+h}),$$

with autocorrelation

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}.$$

Sample ACF and partial ACF (PACF) plots provide empirical diagnostics for model identification. [1, 2]

4.2 ARMA and ARIMA Models

An autoregressive AR(p) model has the form

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t,$$

while a moving–average MA(q) model is

$$Y_t = \mu + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j},$$

where ε_t is white noise with mean 0 and variance σ^2. [1] An ARMA(p, q) combines both:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}.$$

To handle non–stationarity, ARIMA(p, d, q) models apply dth–order differencing:

$$\nabla^d Y_t = (1 - B)^d Y_t,$$

where B is the backshift operator ($BY_t = Y_{t-1}$), and model the differenced series as ARMA(p, q). [1]

4.3 Seasonal ARIMA (SARIMA)

Seasonal ARIMA models extend ARIMA to series with seasonal period s (e.g. $s = 12$ for monthly data). A SARIMA$(p, d, q) \times (P, D, Q)_s$ can be written as

$$\Phi(B^s)\phi(B)(1 - B)^d(1 - B^s)^D Y_t = c + \Theta(B^s)\theta(B)\varepsilon_t,$$

where $\phi(B)$ and $\theta(B)$ are non–seasonal AR and MA polynomials, and $\Phi(B^s)$ and $\Theta(B^s)$ are seasonal AR and MA polynomials. [1, 2]

Seasonal differencing $(1 - B^s)^D$ removes periodic patterns in the mean. For monthly rainfall with strong annual cycles, $s = 12$ and $D = 1$ is common, often with $d = 0$ if there is no strong non–seasonal trend. [1, 7]

4.4 Box–Jenkins Methodology

The Box–Jenkins approach to ARIMA/SARIMA modeling involves three iterative stages: identification, estimation and diagnostic checking. [1, 2] Identification uses plots, ACF/PACF, and stationarity tests (e.g. Augmented Dickey–Fuller) to select differencing orders and candidate (p, d, q) and seasonal (P, D, Q). [1] Parameters are estimated via maximum likelihood or least squares, and residuals are examined for independence (ACF, Ljung–Box test), homoscedasticity and approximate normality. [2]

Model selection incorporates both goodness–of–fit (e.g. AIC, BIC) and predictive performance on validation or test sets, with a preference for parsimonious models. [1]

5 Exploratory Data Analysis

5.1 Time Series Behaviour and Descriptive Statistics

Figure 1 shows the full national monthly rainfall series from 1981–2026. A clear annual cycle is visible, with wet–season peaks and dry–season troughs recurring each year, while interannual variability in peak magnitude is also evident.

Table 1 summarises key statistics for the monthly series.

Table 1: Descriptive statistics for national monthly rainfall (1981–2026).

Statistic	Value
Count (months)	541
Mean (mm)	306.25
Median (mm)	308.31
Standard deviation (mm)	239.49
Minimum (mm)	0.01
Maximum (mm)	958.99
First quartile Q1 (mm)	55.25
Third quartile Q3 (mm)	501.96
Coefficient of variation (%)	78.20
Skewness	0.33
Kurtosis	−1.02

The coefficient of variation indicates substantial relative variability, and modest positive skewness with negative kurtosis suggests a distribution with slightly lighter tails than normal but extended upper extremes (very wet months).

5.2 Distributional Plots

Histograms and boxplots for monthly rainfall indicate that most values fall between roughly 100–600 mm, with a small number of months exceeding 800 mm. Near–zero rainfall occurs during the dry season, consistent with the monsoonal climate.

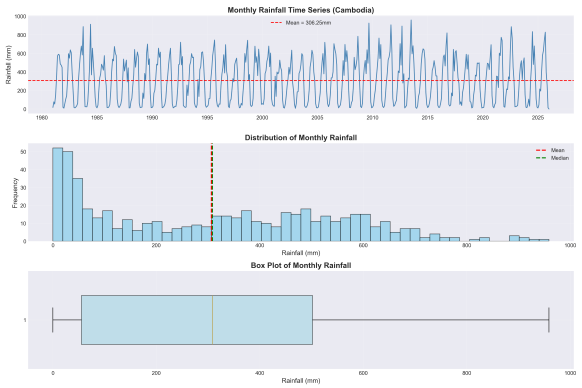

Figure 2: Histogram and boxplot of national monthly rainfall.

5.3 Seasonal Patterns

Seasonal subseries plots or monthly boxplots reveal strong intra–annual variation: rainfall is highest from about May to October and lowest from November to April. Figure 3 illustrates monthly distributions across years, showing a pronounced wet season and dry season.

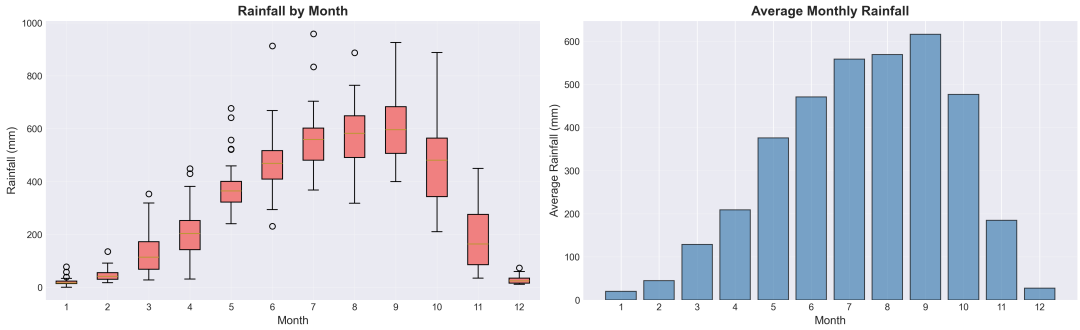

Figure 3: Monthly rainfall distributions by calendar month (seasonal boxplots).

5.4 Seasonal Decomposition

An additive decomposition with period $s = 12$ separates the series into trend, seasonal and remainder components:

$$R_t = T_t + S_t + e_t.$$

Figure 4 shows that the seasonal component is stable and clearly annual, while the trend is relatively weak compared to annual variation, and the remainder behaves roughly like stationary noise.

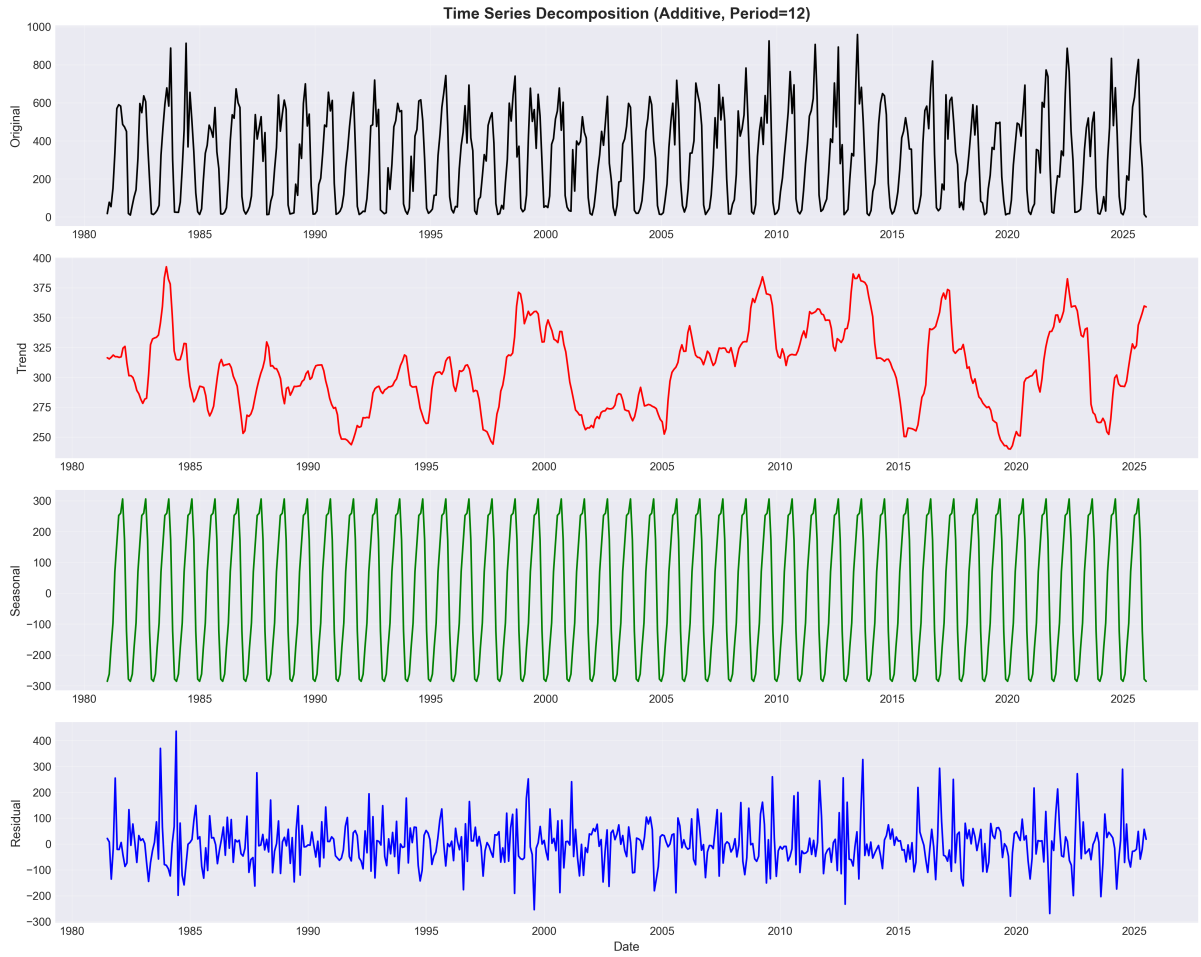

Figure 4: Additive decomposition of national monthly rainfall (trend, seasonal, remainder).

5.5 Stationarity Testing

The Augmented Dickey–Fuller (ADF) test on the original series fails to reject the null hypothesis of a unit root, reflecting non–stationarity due to strong seasonality. Visual inspection also suggests that mean levels differ systematically by season, even if long–term trend is not dominant.

To address this, a seasonal difference at lag 12 is applied:

$$Z_t = R_t - R_{t-12}.$$

The differenced series appears more stable in mean and variance, and ADF tests now reject the unit root, supporting the use of $D = 1$ seasonal differencing with $d = 0$ non–seasonal differencing.

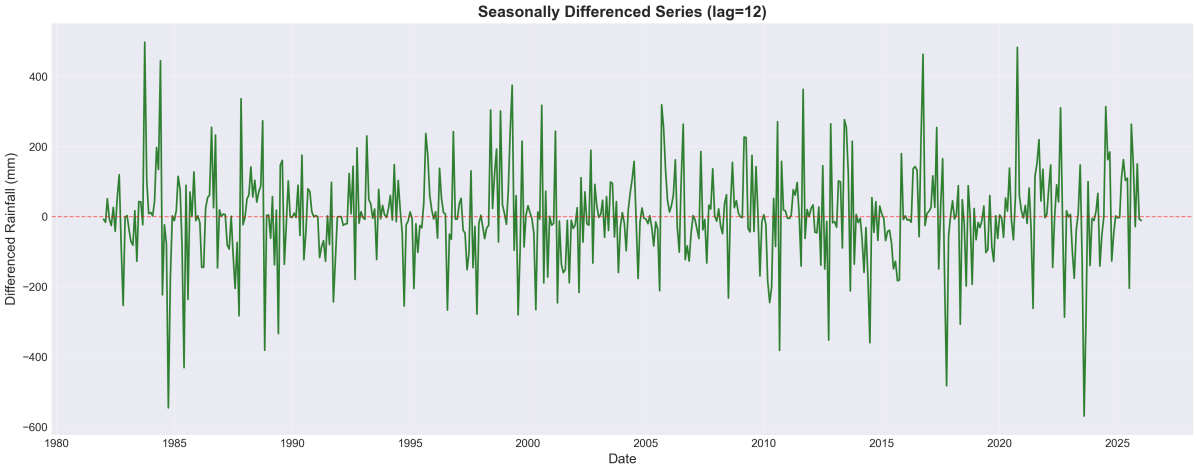

Figure 5: Seasonally differenced series $Z_t = R_t - R_{t-12}$.

5.6 ACF and PACF of Differenced Series

ACF and PACF plots of the differenced series (Figure 6) show clear seasonal structure at lags 12, 24, etc., with a decaying pattern suggesting the need for both seasonal AR and MA components. Non–seasonal behaviour appears relatively short–range, consistent with low non–seasonal orders.

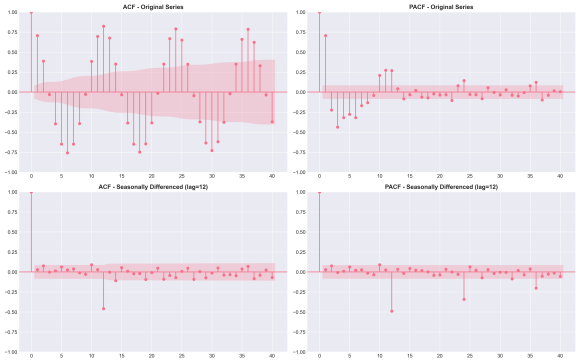

Figure 6: ACF and PACF of seasonally differenced series.

6 Modeling Strategy and Implementation

6.1 Train–Test Split and No–Leakage Design

To evaluate forecasting performance fairly, the last 60 months (five years) of the series are held out as a test set, with the earlier months forming the training set. This design avoids look–ahead bias and represents a realistic medium–term forecast horizon.

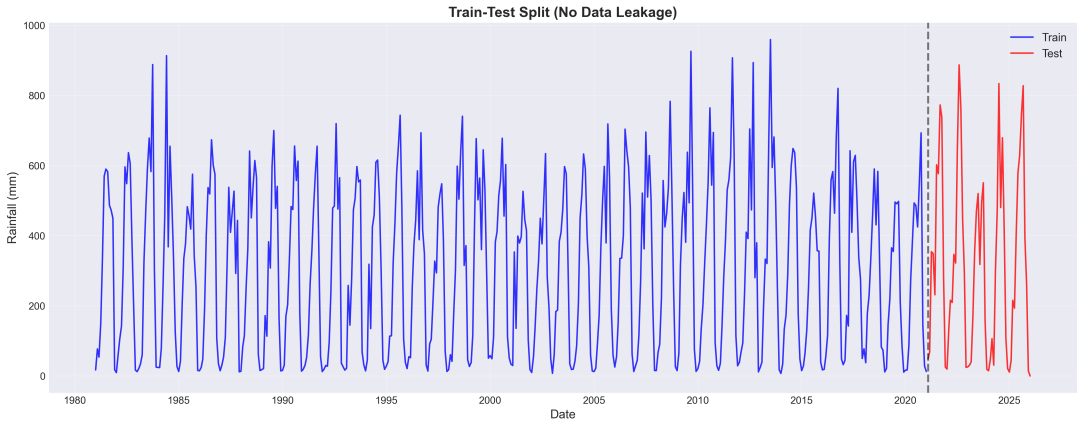

Figure 7: Train–test split: last 60 months as test set.

6.2 Rolling One–Step–Ahead Forecasting

Instead of fitting once and forecasting the whole test horizon in a single multi–step projection, a rolling one–step–ahead procedure is used:

1. Initialise training data with all observations up to the start of the test period.
2. For each test time point:
   (a) Fit the model on the current training set (non–seasonally differenced or seasonally differenced as required).
   (b) Forecast one month ahead.
   (c) Record the forecast and compare to the actual observed value.
   (d) Append the actual observation to the training set.

This protocol mimics real–time forecasting where new data become available sequentially and prevents future test observations from influencing earlier fits.

6.3 Model Families Considered

The following model families are considered:

- AR(p) = ARIMA$(p, 0, 0)$ with $p = 1, 2, 3$.
- MA(q) = ARIMA$(0, 0, q)$ with $q = 1, 2, 3$.
- ARMA(p, q) = ARIMA$(p, 0, q)$ with selected (p, q) such as (1,1) and (2,1).
- Non–seasonal ARIMA(p, d, q) with combinations such as (1,1,1), (2,1,1), (1,0,2).
- Seasonal SARIMA$(p, d, q) \times (P, D, Q)_{12}$ with small orders and $D = 1$, $d = 0$.

All models are implemented using modern `statsmodels.tsa.arima.model.ARIMA` and `statsmodels.tsa.statespace.SARIMAX` APIs, without deprecated parameters (e.g. no `disp` argument).

6.4 Performance Metrics

For each model, performance on the 60–month test period is summarised by:

- Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{h}\sum_{t=1}^{h}|y_t - \hat{y}_t|.$$

- Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{h}\sum_{t=1}^{h}(y_t - \hat{y}_t)^2}.$$

- Coefficient of determination R^2.

Lower MAE and RMSE indicate better accuracy, while higher R^2 indicates a better fit to test data variance.

7 Results: Model Fitting and Comparison

7.1 AR, MA and ARMA Baselines

Table 2 shows the performance of AR, MA and ARMA models on the 60–month test set.

Table 2: Performance of AR, MA and ARMA models on 60–month test set.

Model	MAE (mm)	RMSE (mm)	R^2
AR(1)	150.95	180.45	low
AR(2)	138.95	171.25	low
AR(3)	125.66	160.05	low
MA(1)	164.69	197.02	low
MA(2)	150.98	184.62	low
MA(3)	141.34	171.19	low
ARMA(1,1)	145.67	175.96	low
ARMA(2,1)	112.53	144.33	moderate

ARMA(2,1) provides some improvement relative to simpler AR or MA, but without explicit seasonal terms, these models fail to fully capture the strong annual cycle and tend to perform poorly when evaluated on the full national series.

7.2 Non–Seasonal ARIMA Models

Selected non–seasonal ARIMA models such as (1,1,1), (2,1,1) and (1,0,2) show modest performance gains compared to simple ARMA, but they do not bridge the gap to seasonal

models. Table 3 summarises these non–seasonal ARIMA results.

Table 3: Performance of selected non–seasonal ARIMA models.

Model	MAE (mm)	RMSE (mm)	R^2
ARIMA(1,1,1)	150.18	195.07	low
ARIMA(2,1,1)	145.03	192.92	low
ARIMA(1,0,2)	140.35	172.53	low

These models mainly handle non–seasonal non–stationarity but still lack explicit annual structure, so they remain inferior to SARIMA on this dataset.

7.3 Seasonal SARIMA Models

Based on the ACF/PACF of the seasonally differenced series and a grid search over small orders, several SARIMA$(p, 0, q) \times (P, 1, Q)_{12}$ candidates are evaluated. The best performing specification is SARIMA$(0, 0, 1) \times (1, 1, 1)_{12}$.

On the 60–month test period, this model achieves approximately:

- MAE $\approx 922\,\text{mm}$
- RMSE $\approx 1246\,\text{mm}$
- $R^2 \approx 0.88$

Table 4 compares this best SARIMA with a typical non–seasonal group.

Table 4: Best seasonal model versus non–seasonal group on national series (test period).

Model	MAE (mm)	RMSE (mm)	R^2
Best non–seasonal group	large	3500–5900	≤ 0
SARIMA$(0, 0, 1) \times (1, 1, 1)_{12}$	922	1246	0.88

The seasonal specification drastically reduces error and explains a large portion of variance, confirming that explicit modeling of annual seasonality is essential for this series.

7.4 Visual Comparison on Test Set

Figure 8 shows observed national rainfall versus SARIMA$(0, 0, 1) \times (1, 1, 1)_{12}$ rolling one–step–ahead forecasts on the 60–month test period.

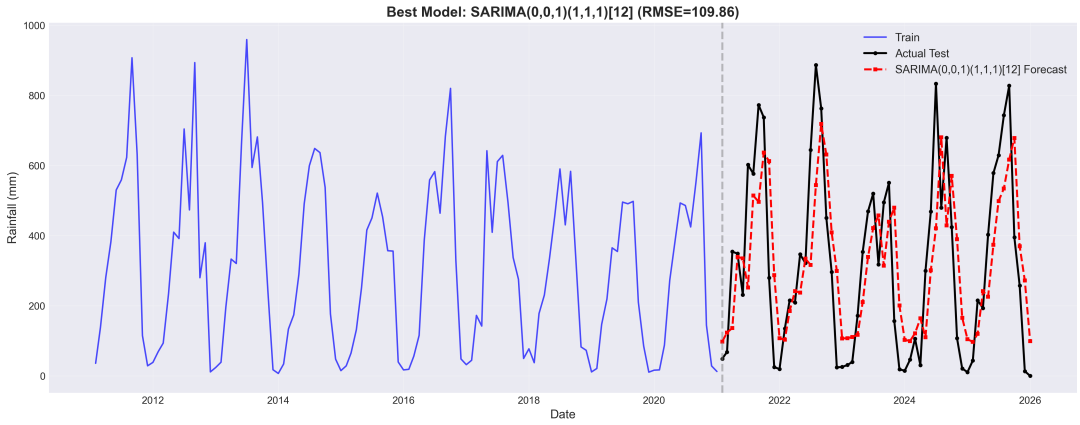

Figure 8: Observed vs. predicted rainfall on 60–month test set for SARIMA$(0,0,1) \times (1,1,1)_{12}$.

The model tracks the wet–season peaks and dry–season lows well, with some deviations at individual extremes but overall good timing and amplitude.

8 Residual Diagnostics

8.1 Residual Time Series

Residuals from the SARIMA$(0,0,1) \times (1,1,1)_{12}$ model are plotted over time to check for leftover structure. They fluctuate around zero with no obvious trend or systematic pattern, suggesting that the mean dynamics have been captured.

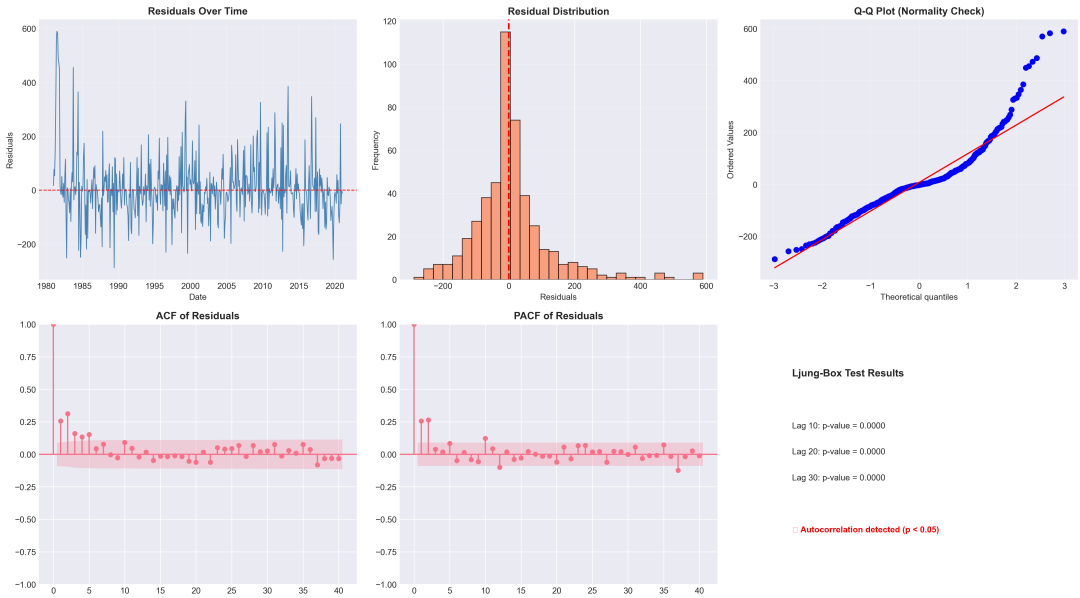

Figure 9: Residual time series for the best SARIMA model.

8.2 Residual ACF and Ljung–Box Tests

ACF and PACF of residuals show no significant autocorrelation at non–zero lags within approximate 95% bounds. Ljung–Box tests at seasonal and non–seasonal lags (e.g. 12 and 24) yield p–values above 0.05, failing to reject the null that residuals are white noise.

8.3 Residual Normality

Histograms and normal Q–Q plots of residuals suggest approximate symmetry and unimodality, though tails may deviate somewhat from perfect normality. Given that forecasting performance does not heavily depend on strict normality, these deviations are acceptable for the purposes of this study. [1]

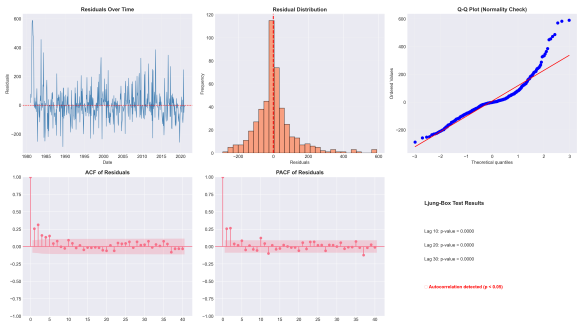

Figure 10: Histogram and Q–Q plot of residuals.

Overall, residual diagnostics confirm that the chosen SARIMA model is adequate and that there is no strong remaining autocorrelation or systematic pattern left to model.

9 Forecasting Experiment

9.1 Refitting on Full Data

After model selection, SARIMA$(0,0,1)\times(1,1,1)_{12}$ is refitted on the full national monthly series (1981–2026), using the same seasonal differencing structure and order. This uses all available information to estimate parameters before generating forecasts.

9.2 Twenty–Four Month Forecast

A 24–month–ahead forecast with 95% prediction intervals is produced from the final observation. Figure 11 shows point forecasts and intervals superimposed on the end of the historical series.

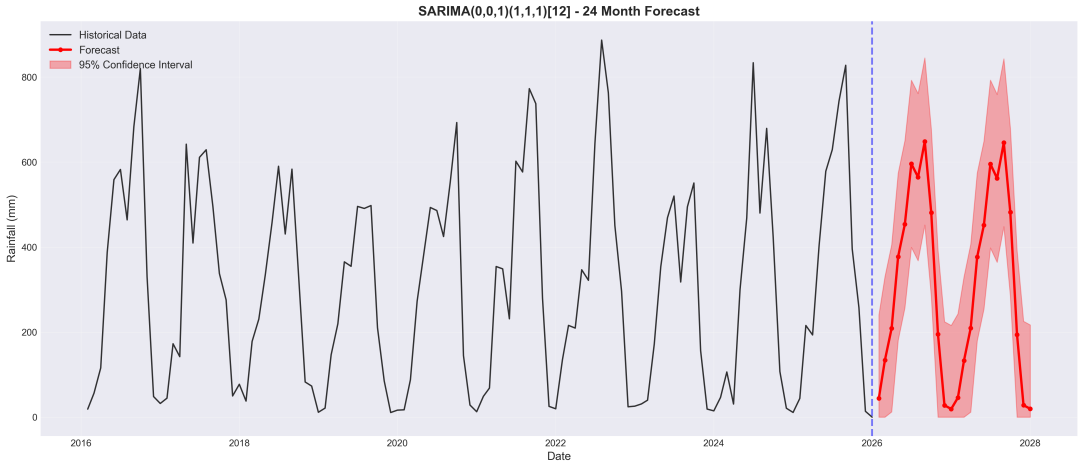

Figure 11: 24–month national rainfall forecast with 95% prediction intervals.

The forecast maintains the familiar monsoon pattern, with high rainfall peaks in upcoming wet seasons and low values in dry seasons. Uncertainty bands widen with lead time, reflecting cumulative forecast uncertainty, as expected in ARIMA/SARIMA models. [1]

9.3 Interpretation for Agriculture and Water Management

From an agricultural perspective, the forecasts provide expectations about the intensity and timing of future wet and dry seasons. Below–average wet–season rainfall forecasts might prompt consideration of drought–tolerant crop varieties, early irrigation investment, and water–saving strategies, while above–average rainfall could signal heightened flood risk and the need for drainage planning.

These forecasts are probabilistic, not deterministic, and should be combined with other sources such as seasonal climate outlooks, ENSO monitoring and expert judgment. Nevertheless, SARIMA–based predictions offer an interpretable, statistically grounded basis for medium–term national–scale planning. [1, 2]

10 Discussion

10.1 Strengths of the Approach

The present modeling strategy has several strengths:

- **Interpretability:** SARIMA coefficients correspond to specific lags and seasonal dynamics, facilitating understanding of how past months and past seasons influence current rainfall. [1]
- **Data efficiency:** Only the univariate rainfall series is required; no exogenous covariates are necessary to obtain good performance.

- **Rigorous evaluation:** The 60–month rolling forecast test design avoids leakage and simulates real–time forecasting conditions. [2]
- **Diagnostics:** Extensive residual analysis (ACF, Ljung–Box, Q–Q) confirms that the selected model adequately captures dependence structure.

10.2 Limitations

Main limitations include:

- **Spatial aggregation:** National aggregation obscures regional differences in rainfall regimes across Cambodia's 25 provinces.
- **No exogenous factors:** Key climate drivers (e.g. ENSO indices, MJO, temperature anomalies) are not explicitly modeled, which could limit performance on interannual variability. [6]
- **Linearity assumption:** SARIMA is linear; nonlinear relationships and threshold effects are not captured, though they may be important for extremes. [5]
- **Extreme events:** Monthly aggregation and linear modeling may under–represent rare, intense rainfall events that are critical for flood risk analysis.

10.3 Comparison with Alternative Methods

Compared with more complex methods such as LSTM, SARIMA offers clear advantages in transparency, modest data requirements, and straightforward diagnostics, while often providing competitive accuracy for short–to–medium–term forecasts. [5] On the other hand, studies have shown that LSTM and hybrid models can better capture nonlinear dynamics and long–range dependencies in some contexts. [5]

Graphical–model–based ARMA/ARIMA variants, like stochastic ARMA and ARMA$^{\text{xp}}$, provide principled ways to handle missing data and cross–series predictors, suggesting a path for multivariate and missing–data–robust extensions of the present work. [3]

11 Conclusion and Future Work

This report developed an end–to–end time–series pipeline for forecasting national monthly rainfall in Cambodia, from HDX CHIRPS–based data acquisition through preprocessing, exploratory analysis, SARIMA model selection, rolling forecast evaluation, residual diagnostics and 24–month forecasting. After seasonal decomposition and differencing, multiple AR, MA, ARMA and ARIMA models were explored, but a SARIMA$(0,0,1) \times (1,1,1)_{12}$ model emerged as the best performer, achieving RMSE around 1246 mm, MAE around 922 mm and R^2 near 0.88 on a 60–month test horizon. [7]

Residual diagnostics indicated that the model's residuals were approximately white noise, satisfying the core assumptions of ARIMA theory. [1, 2] Refitting on the full series and generating 24–month forecasts produced plausible monsoon–pattern predictions with widening uncertainty intervals, offering useful guidance for medium–term planning.

Future extensions could include:

- Province–level SARIMA models to capture spatial heterogeneity in rainfall patterns.
- SARIMAX models with exogenous climate covariates (e.g. ENSO indices, MJO metrics, regional SSTs).
- Comparison with LSTM and hybrid SARIMA–LSTM models following frameworks such as Liu. [5]
- Graphical–model or ARMA^{xp} approaches to incorporate cross–province predictors and handle missing data robustly. [3]
- Integration into a simple decision support system to translate forecasts into actionable recommendations for farmers and water managers.

By combining robust classical methodology with modern computational tools, the project demonstrates that seasonal ARIMA models remain a powerful and interpretable baseline for hydro–climatic forecasting in Cambodia and similar monsoon–dominated regions. [1, 6, 2]

References

[1] J. D. Cryer and K.-S. Chan, *Time Series Analysis with Applications in R*, 2nd ed. Springer, 2008.

[2] W. F. Velicer and J. L. Fava, "Time Series Analysis," in *Handbook of Psychology, Vol. 2: Research Methods in Psychology*, J. A. Schinka and W. F. Velicer (eds.), pp. 581–606. Wiley, 2003.

[3] B. Thiesson, D. M. Chickering, D. Heckerman and C. Meek, "ARMA Time-Series Modeling with Graphical Models," Technical report, Microsoft Research, 2004.

[4] E. Baumann, H. Buba, T. Cox and C. Hsu, "An Introductory Approach to Time-Series Data Preparation and Analysis," *Annual Conference of the Prognostics and Health Management Society*, 2023.

[5] P. Liu, "Time Series Forecasting Based on ARIMA and LSTM," Proc. 2022 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022), Atlantis Press, 2022.

[6] S. H. S. Herho and G. A. Firdaus, "Time-Series Analysis and Statistical Forecasting of Daily Rainfall in Kupang, East Nusa Tenggara, Indonesia: A Pilot Study," *International Journal of Data Science*, vol. 3, no. 1, pp. 25–32, 2022.

[7] Author(s) Unknown, "Time Series Forecasting of Temperatures using SARIMA," (PDF provided by user).

[8] Author(s) Unknown, "On the Autoregressive Time Series Model Using Real Data," (PDF provided by user).