# RAINFALLS IN NAVARCLES: VISUALISATION AND FORECASTING

Liubov Shubina[*]

[*] *IT Academy, Barcelona Activa,*
*Carrer de Roc Boronat, 117, 127, Sant Martí, 08018 Barcelona, Spain*

## Abstract

This study aims to forecast monthly rainfall using various time series models, including SARIMA, SARIMAX, and machine learning techniques, while also providing an interactive dashboard for public use. The primary objectives were to evaluate the effectiveness of forecasting models and to develop a user-friendly visualization tool for Navarcles. Historical rainfall data, including a unique privately collected dataset from 1995 to 2024, was analysed along with external climatic features. Forecasting methods such as Auto-ARIMA, Prophet, SARIMA, SARIMAX, ElasticNET, CatBoost, and LightGBM were tested, with Auto-ARIMA achieving the best performance (test MAE of 29.19). External regressors in SARIMAX did not significantly enhance accuracy. Complementing the forecasting, an interactive dashboard was created using Tableau Desktop (Public Edition), allowing citizens to explore 30 years of rainfall and snow data by year and month. The dashboard provides insights into annual and monthly rainfall trends, drought periods, and snow occurrences. This project not only advances rainfall forecasting in the region but also empowers Navarcles residents with accessible climate data.

## Introduction

Rainfall forecasting plays a crucial role in agriculture, water resource management, and urban planning. Traditional time series models like SARIMA (Seasonal Autoregressive Integrated Moving Average) have been extensively used for forecasting meteorological data due to their ability to capture seasonality and trends. Advancements in machine learning (ML) also offer new possibilities to enhance model accuracy, particularly when external features (exogenous variables) are considered.

This study focuses on precipitation patterns in Navarcles, a village in the Bages region, province of Barcelona, within the autonomous community of Catalonia, Spain. Located approximately 11 km from the regional capital, Manresa, Navarcles experiences rainfall patterns that vary significantly between years and months. Data used in this study is a private dataset: manually collected by a resident of Navarcles daily precipitation from 1995 until 2024. Navarcles lacks publicly accessible rainfall records before 2008, making the privately collected dataset particularly valuable. This unique dataset may interest local authorities, residents, and researchers concerned with long-term precipitation trends.

This study has two main goals: (1) to compare the performance of various time series and non-specific time series models in forecasting rainfall, and (2) to develop an interactive dashboard for visualising precipitation data.

ML models, such as SARIMA, SARIMAX (which includes external regressors), Prophet, LightGBM and CatBoost were explored to predict monthly rainfall. Performance of these models was compared, and the impact of including external variables, such as temperature and solar radiation, on forecast accuracy was investigated. The primary research question is whether more complex models, such as SARIMA/X or ML approaches, can outperform simpler constant model in forecasting rainfall.

## Materials and Methods

### *Data Collection and Preparation*

The study utilized a private rainfall dataset for Navarcles spanning 1995 to 2024. Rainfalls were measured manually by a resident of Navarcles using a pluviometer (unit: mm, equivalent to l/m²). Data was recorded in a notebook from 1995 until April 2020 and has been entered into an Excel spreadsheet. The dataset includes daily rainfall amounts (mm) and notes on snow events.

Daily data for Manresa was obtained from Catalan Daily Temperature and Precipitation dataset, which consist of continuous, homogeneous, and publicly accessible climate series at daily and monthly resolutions since 1950. Daily data for Manresa includes year, month, day, daily accumulated precipitation in mm, daily maximum/minimum temperature in ºC, insolation in h.

Missing values were addressed by filling the gaps using Manresa data, given the geographical proximity and the similarity of their rainfall patterns. The data was then resampled into monthly totals for both Navarcles and Manresa datasets.

The dataset was split into training and test sets (10%), with the training set used for model fitting and the test set for evaluating forecast accuracy.

### *EDA*

Stationarity of the data was assessed using visual inspection of the time series plot, its rolling mean and SD, and augmented Dickey-Fuller (ADF) statistical test.

To identify seasonality, trends, and potential irregular components, seasonal decomposition of the time series was performed. Additionally, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were analyzed to evaluate temporal dependencies and to guide the selection of appropriate SARIMA model parameters. These diagnostic tools helped in determining the orders of autoregressive (AR), differencing (I), moving average (MA), and seasonal components necessary for effective model fitting.

### *Modeling Approaches*

Data analysis and preprocessing, model evaluation and rainfall forecast were performed in Python3.10.13 with Jupyterlab 4.2.1, running on MacOS10.14.6.

Investigated models:

1. SARIMA (Statsmodels) and Auto-ARIMA (Pmdarima): Seasonal models were fitted, with Auto-ARIMA selecting the optimal parameters through iterative search.

2. SARIMAX (Statsmodels): External regressors were incorporated to assess their impact.

3. Prophet (Facebook): Implemented for its robustness in handling seasonality.

4. Non-specific time series models: ElasticNet were tested to implement regularization; LightGBM (Microsoft) and CatBoost (Yandex) were tested to capture potential non-linearities.

Time-series cross-validation (TSCV) was employed to evaluate model performance. The training set gradually increases with each fold, while the validation set remains the same size (matching the test period). Mean Absolute Error (MAE) was calculated for each fold, and the mean and standard deviation of the errors were recorded.

The best-performing model, identified through TSCV, was subsequently evaluated on the test set to verify its predictive accuracy. This validated model was then employed to generate future rainfall forecasts.

*Dashboard Development*

A user-friendly dashboard was created using Tableau Desktop (Public Edition) 2022.3.0.

## Results

Initial visualization of the data indicated significant variation in monthly rainfall, with occasional peaks exceeding 150 mm, especially in the spring and autumn months **(fig. 1, 2)**. Outlier detection was performed using the K-Nearest Neighbours method and revealed that most extreme rainfall events were natural occurrences and should be retained in the dataset for further analysis.

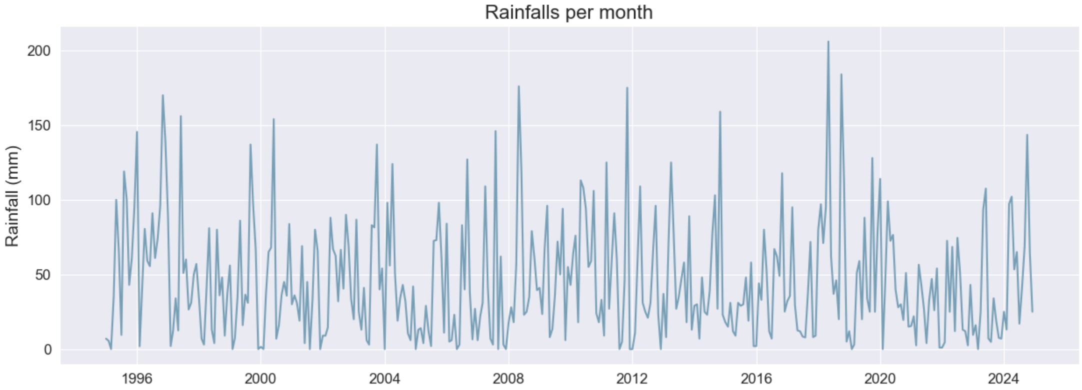

**Figure 1.** Daily rainfalls in Navarcles.

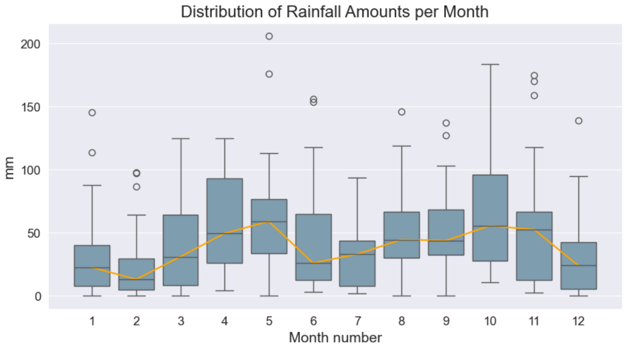

**Figure 2**. Distribution of rainfalls per month.

*Seasonality, ACF/PACF analysis*

Despite some indications of seasonality, no clear trend was identified in the data. Statistical ADF test (ADF Statistic: -15.63, p-value: 1.6966264846974422e-28), confirmed that the dataset was stationary, allowing for further analysis without the need for differencing **(Fig. A1).**

While there was some seasonal variation, particularly in spring and autumn, the overall seasonality was not strong enough to guarantee predictable patterns without sophisticated models. ACF analysis indicates short-term correlations at lag 1 and seasonal effects at lags 24 and 36, while PACF highlights significant seasonal autocorrelations at lag 12. Based on these findings, a SARIMA(1,0,1)(1,1,1,12) model, which captures both short-term and seasonal dependencies, was chosen as the starting point **(Fig. A2-3).**

*Forecasting Models*

For this project, several popular models for time series forecasting: SARIMA, SARIMAX, Prophet, ElasticNet, CatBoost, and LightGBM were evaluated. As a baseline, median model that predicts the median rainfall for each month based on past data was used. This simple model provided a reference point for more complex models.

The SARIMA (1,0,1)(1,1,1,12) model, with initial parameters determined through ACF and PACF analysis, showed suboptimal performance compared to the constant model, with an average MAE of 32.80. This suggests that simpler models may be more effective for the given dataset, given the high volatility in rainfall data.

Auto-ARIMA automatically selected the best parameters, producing a simpler model: SARIMA(0,0,1)(1,0,1)[12]. This model showed a slight improvement, with an average MAE of 29.35, but the performance was not dramatically better than the median model, indicating that the data may be challenging to model.

Prophet, designed for capturing seasonality and trends, achieved an MAE of 29.72, slightly worse than Auto-ARIMA, suggesting that while Prophet handles seasonality well, it did not provide a significant improvement for this dataset.

In an effort to enhance the performance of time series forecasting models, external regressors (exogenous features) were incorporated:

- **Month of the year**: Captures seasonal patterns.
- **Lagged values**: Including 1, 6, and 12-month lags of rainfall, rolling mean, and rolling standard deviation.
- **Rainfall-related statistics**: Number of rainy days, maximum daily rainfall, and the longest consecutive dry days.
- **Climatic data from the Manresa dataset**: Monthly median temperature (max/min) and insolation values.

Correlation and Multicollinearity Analysis were performed using Phik correlation coefficients. Rainfall does not show a significant correlation with any of the features. While high correlation between certain features could present multicollinearity risks for linear models, this could be mitigated by regularization techniques.

SARIMAX model were tested using a subset of the most correlated exogenous features: maximum rain, lag 6, and insolation. Despite the additional predictors, the average cross-validation MAE was 30.63, slightly worse than the performance of Auto-ARIMA and the median model. This suggests that while the external features captured some seasonal variation and short-term memory, they may not have added significant predictive value, or the model may have struggled with the complexity.

Next, non-specific time series models including ElasticNet, CatBoost, and LightGBM were tested. These models were trained using all available exogenous features and evaluated using cross-validation. For each model, we performed a Randomized Search CV to optimize hyperparameters. The best performing model was the LightGBM Regressor, achieving an average cross-validation MAE of 30.98. Despite using a more complex model and a wide array of features, its performance was worse than that of SARIMAX with external features and Auto-ARIMA **(Table 1, fig. A4).**

**Table 1.** Model comparison.

| Model | Average Cross-Validation MAE | Standard Deviation of MAE |
|---|---:|---:|
| Median Model | 30.68 | 4.16 |
| SARIMA(1,0,1)(1,1,1,12) | 32.80 | 7.88 |
| Auto-ARIMA(0,0,1)(1,0,1)[12] | 29.35 | 3.38 |
| Prophet | 29.72 | 3.44 |
| SARIMAX | 30.63 | 3.34 |
| LightGBM | 30.98 | 4.01 |

Auto-ARIMA, which automatically selects the best model parameters, has the smallest standard deviation (SD) and performs slightly better than the Constant model. Therefore, it was chosen for forecasting monthly rainfall. The Auto-ARIMA model was tested on the data, and the results showed **MAE=29.19** on the test set. The model displayed systematic inaccuracies, particularly struggling with the peaks and valleys of the time series, underestimating rainfall maxima and overestimating minima **(Fig. 3)**.

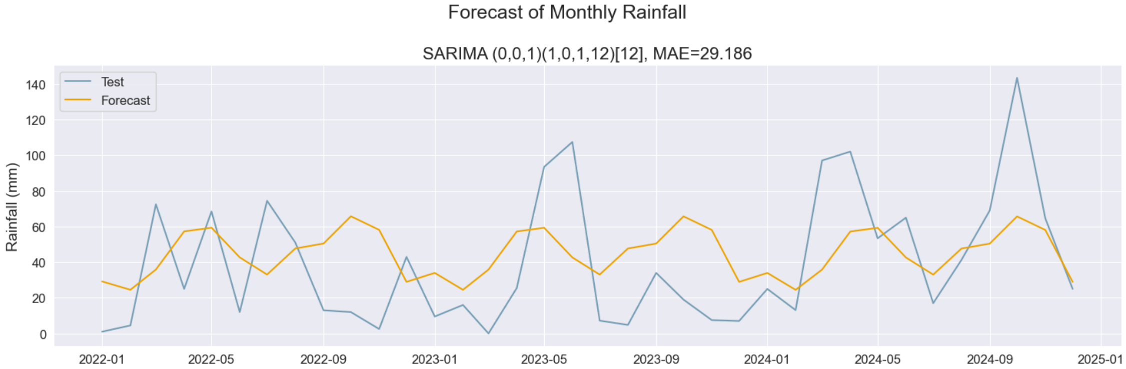

**Figure 3**. Forecast vs. original data.

After testing, Auto-ARIMA model was used to generate forecasts for the year 2025. Model was trained on the full dataset and forecasted monthly rainfall for the entire year of 2025. The forecast indicated not very variable rainfall across the year, with higher rainfall expected for May (61 mm) and October (65 mm). The forecasted values are as follows:

| **Month** | **Total Rainfall, mm** |
|---|---|
| 2025-01 | 31 |
| 2025-02 | 23 |
| 2025-03 | 38 |
| 2025-04 | 56 |
| 2025-05 | 61 |
| 2025-06 | 45 |
| 2025-07 | 33 |
| 2025-08 | 46 |
| 2025-09 | 50 |
| 2025-10 | 65 |
| 2025-11 | 54 |
| 2025-12 | 29 |

### *Dashboard Insights*

The Tableau dashboard provides multiple functionalities:

- Annual Analysis: Identify the rainiest and driest years, mean annual rainfall, and longest wet and dry periods.
- Monthly Analysis: Examine number of rainy days, total rainfall, and monthly extremes.

- Snowfall Data: Access historical snow occurrences.
- User Interaction: Citizens can select specific years or months to explore detailed precipitation patterns.

## Discussion

Due to the inherently unpredictable and random nature of weather patterns, accurately forecasting rainfall remains a challenging and complex task despite significant advances in meteorological research. The variability of rainfall patterns poses substantial obstacles to precise meteorological predictions.

While Auto-ARIMA provided the most accurate forecasts, the addition of external features in SARIMAX models did not yield notable improvements. This suggests that the intrinsic seasonality and autocorrelation in the rainfall data dominate predictive capability. Machine learning models did not outperform traditional approaches, indicating that for this dataset, simpler models are sufficient. The key concerns are:

- Rainfall is inherently difficult to predict with the provided features due to its high variability.
- Time series models like SARIMA may capture temporal dependencies more effectively than general ML models.
- ML models such as LightGBM and CatBoost did not provide significant improvements over SARIMA and Auto-ARIMA, possibly because they do not inherently capture the temporal structure of the data.

The results indicate that while simpler models like Auto-ARIMA and Prophet performed slightly better than the median model, the data's high variability and frequent outliers limited the effectiveness of models.

The Tableau dashboard represents a significant step toward community engagement with local climate data. By making 30 years of precipitation data accessible, residents can make informed decisions about water management and climate adaptation strategies.

## Conclusion

This study achieved two primary objectives: developing rainfall forecasts for Navarcles and creating a publicly accessible dashboard. Auto-ARIMA emerged as the best forecasting model, while the dashboard provides valuable insights into historical precipitation patterns. The unique manually recorded data from 1995 to 2007 adds significant value, filling a critical gap in public records. Future work could include expanding external features, refining machine learning models, and collaborating with local authorities for public dissemination.

## References

Amelia, R., Kustiawan, E., Sulistiana, I., & Dalimunthe, D.Y. (2022). Forecasting rainfall in Pangkalpinang city using seasonal autoregressive integrated moving average with exogenous (SARIMAX). BAREKENG: Jurnal Ilmu Matematika dan Terapan.

Analyzing Time Series Data. https://observablehq.com/blog/analyzing-time-series-data.

Dinesh Kumar, & Pankaj Richariya. (2024). Leveraging SARIMAX for accurate rainfall forecasting in India. International Journal of Innovations in Science, Engineering And Management, 3(2), 37-47.

Ibrahim, A., & Musa, A.O. (2023). On the performance of SARIMA and SARIMAX model in forecasting monthly average rainfall in Kogi State, Nigeria. FUDMA Journal of Sciences.

Mulla, S., Pande, C.B., & Singh, S.K. (2024). Times Series Forecasting of Monthly Rainfall using Seasonal Auto Regressive Integrated Moving Average with Exogenous Variables (SARIMAX) Model. Water Resour Manage, 38, 1825–1846.

Rodrigo, J.A., Ortiz J. E. ARIMA and SARIMAX models with Python. September, 2023 (last update November 2024). https://cienciadedatos.net/documentos/py51-arima-sarimax-models-python

Ryabenko, E. (2015). Applied Statistical Data Analysis. Time Series Analysis, Part I. http://www.machinelearning.ru/wiki/images/archive/e/ec/20150413075746!Psad_ts_ets.pdf.

Yandex Data Analysis School. (n.d.). Machine Learning Handbook. https://education.yandex.ru/handbook/ml.

# Appendix

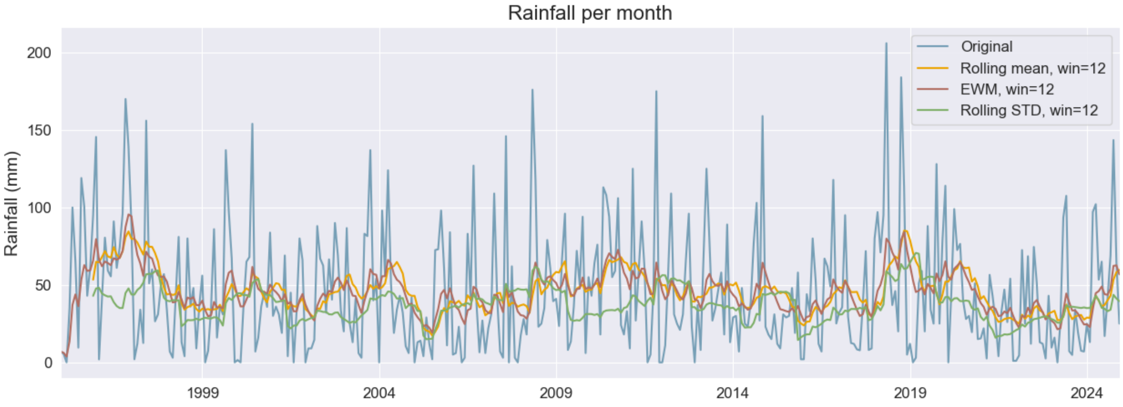

Figure 1. Original data and its rolling mean and SD with 12 months window.

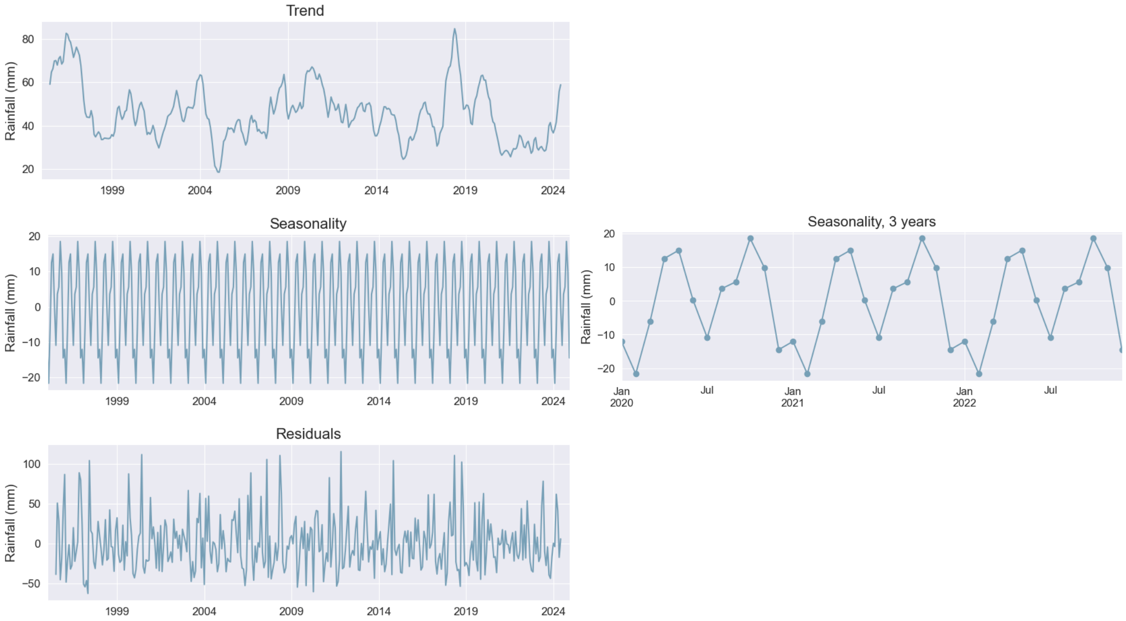

Figure 2. Decomposing Time Series.

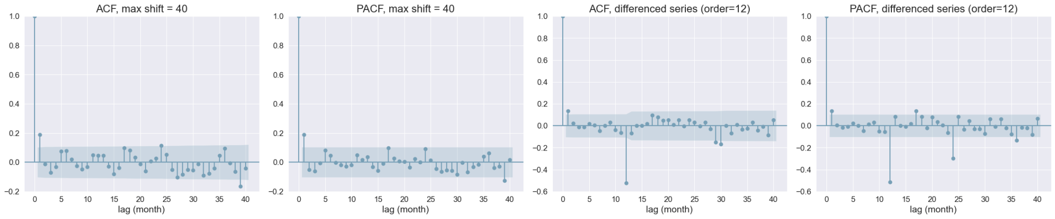

Figure 3. ACF and PCF plots for original (left) and differenced (right) Time Series.

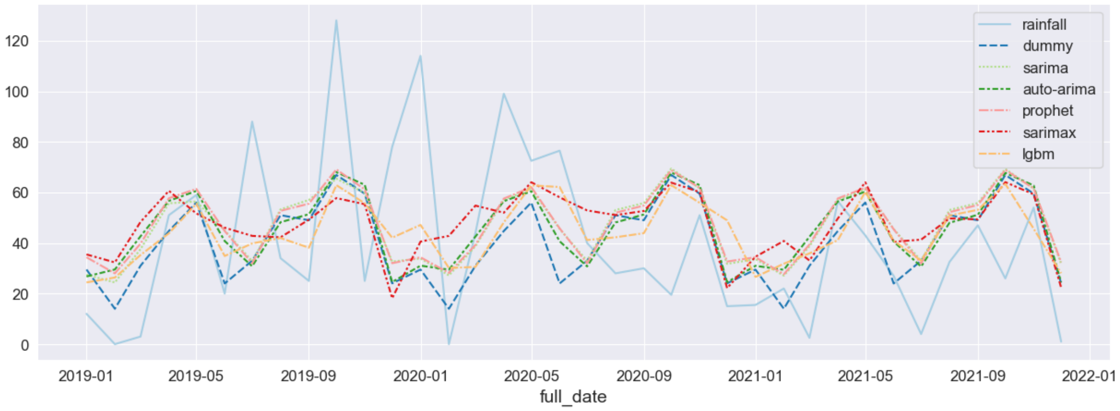

Figure 4. Model comparison.