Analyzing and Forecasting Precipitation Patterns Using Time Series Analysis and Machine Learning Techniques

Ziyue Zhao

Abstract

This study focuses on the analysis and prediction of precipitation data across four distinct geographic locations. Using a comprehensive suite of tools, including time series analysis, statistical tests, and machine learning models like ARIMA, the research aims to understand patterns and predict future precipitation trends. The findings intend to contribute to better forecasting models which are vital for agricultural planning, water resource management, and climate impact assessments.

Introduction

The motivation for this research stems from the increasing need to understand and predict meteorological events, especially precipitation, which has a direct impact on agriculture, water resources, and climate adaptation strategies. Precipitation data was collected from four geographical points and analyzed to understand its behavior over time and predict future trends.

The initial hypothesis was that machine learning and statistical models could effectively model and forecast precipitation patterns, providing valuable insights for meteorological and climate sciences. This hypothesis was investigated by employing a variety of statistical tests to ascertain the stationarity of the time series data and utilizing models like ARIMA for prediction purposes.

The study builds on prior work in the field of meteorology and hydrology where similar methods have been applied to predict other atmospheric variables. Notably, research from 2022 to 2024 has shown that ARIMA and other time series models can successfully forecast weather conditions when applied correctly and combined with appropriate data preprocessing techniques.

The interest in this topic is driven by the broader context of climate change, which is altering precipitation patterns globally. Accurate predictions and models are crucial for developing strategies to mitigate the adverse effects of these changes.

This paper will detail the methodology used for collecting, processing, and analyzing the data, the implementation of statistical tests to verify assumptions necessary for time series modeling, and the application of the ARIMA model to predict future precipitation at the studied locations.

Data Description and Analytics

Data Selection Criteria and Sources

The precipitation data for this study was meticulously selected based on specific criteria to ensure comprehensive coverage and relevance to our predictive modeling goals. The primary criteria included geographical diversity, availability of continuous daily data, and data quality (accuracy and completeness). These criteria were established to facilitate a robust analysis that could potentially be generalized to other regions or incorporated into larger climatic models.

The data was sourced from the NLDAS Primary Forcing Data L4 Hourly 0.125 x 0.125 degree V002 (NLDAS_FORA0125_H), which provides publicly accessible meteorological data. This dataset includes daily precipitation measurements along with other atmospheric variables, but for the purpose of this study, only precipitation data was extracted. The dataset covers multiple geographic locations, but for this analysis, four specific points were chosen based on their geographical and climatic diversity:

- Point 1: Latitude 35.0625, Longitude -115.9375
- Point 2: Latitude 40.0625, Longitude -80.0625
- Point 3: Latitude 52.9375, Longitude -67.0625
- Point 4: Latitude 52.9375, Longitude -115.9375

These locations include varied climatic zones, providing a broad perspective on precipitation patterns across different environmental contexts.

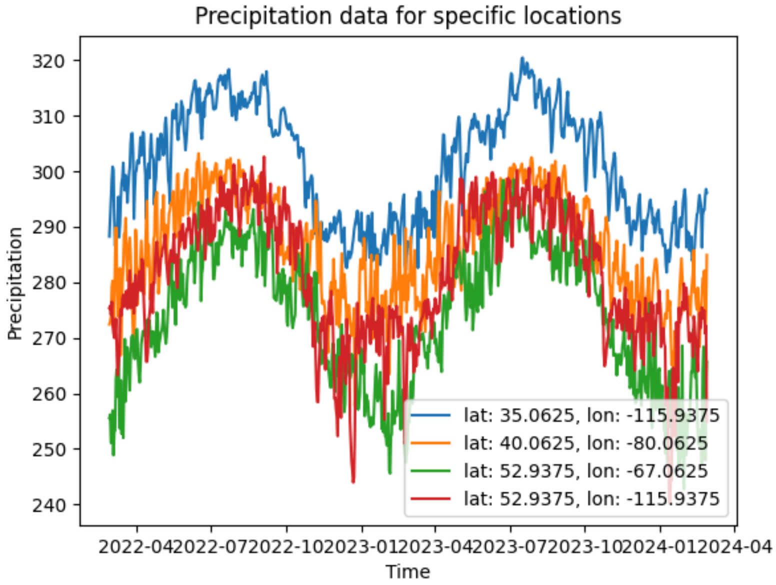
Data and Information Types

The dataset consists of time-series data representing daily (only pick up 0 o'clock data) precipitation amounts recorded in millimeters. Each data point corresponds to a single day's total precipitation measurement at one of the selected locations. The time span of the data covers from February 27, 2022, continuing for a period that encompasses several years, providing a substantial temporal window to observe and analyze precipitation trends and variability.

Exploratory Data Analysis

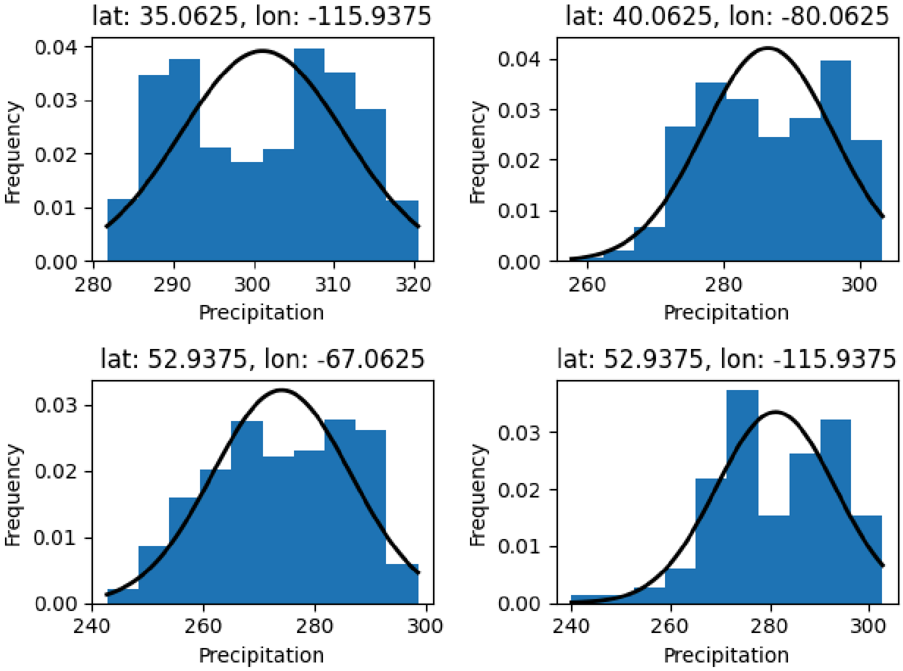
The initial phase of the exploratory data analysis involved plotting the raw precipitation data for each location to visually assess the trends, seasonality, and outliers. This step was crucial to understand the data's structure and inherent characteristics before applying any statistical tests or predictive models.

Figure 1: Precipitation Data Over Time for Selected Locations



Histograms of the data were also generated to observe the distribution and frequency of precipitation values, helping to identify common precipitation ranges and anomalies. These histograms provide insights into the variability of precipitation, which is instrumental in understanding and modeling climate behavior.

Figure 2: Histogram of Precipitation Data for Each Location



The histograms indicate that while the precipitation data for each location appears to display an unimodal distribution, the fit of the normal distribution is not perfect. For example, the histograms suggest a slight right skewness, more pronounced in the datasets for latitudes 40.0625 and 52.9375 with longitudes -80.0625 and -115.9375, respectively. The bars exceed the curve on the right tail, indicating heavier than normal tails. This mismatch between the observed data and the fitted model suggests that the normal distribution may not be the best representation of the underlying stochastic process driving precipitation.

Statistical Analysis

Statistical tests were performed to check for stationarity in the time-series data, an essential criterion for the effectiveness of many time-series forecasting models. The Interquartile Range & Skewness were used.

Location	Interquartile Range	Skewness
Point 1	18.5899	-0.0587
Point 2	16.1850	-0.1681
Point 3	20.6000	-0.1905
Point 4	18.7200	-0.4363

The interquartile range (IQR) and skewness statistics for precipitation data at four distinct locations present intriguing insights into the distributional characteristics of the climatic variables under study. Point 1 exhibits an IQR of 18.5899, indicating that the middle 50% of observations are moderately spread, suggesting a reliable precipitation pattern with some variation but less extreme variability. This is in contrast to Point 3, which displays a notably higher IQR of 20.6000, signifying a wider range of values within its central quartiles, pointing towards a more variable and potentially erratic precipitation regime.

Meanwhile, Point 2 shows the least variability with an IQR of 16.1850, implying a more consistent pattern of precipitation where extreme values are less common within the bulk of the dataset. Point 4, with an IQR of 18.7200, closely follows Point 1 in variability, hinting at a slightly more variable precipitation pattern compared to Point 2 but less so than Point 3.

In terms of skewness, all points exhibit a negative skew, though to varying degrees, with Point 4's skewness of -0.4363 being the most pronounced. This indicates a tail that is heavier on the left side of the distribution, where lower precipitation values are more frequent than higher values. Point 2 and Point 3 have a skewness closer to zero, -0.1681 and -0.1905 respectively, suggesting a more symmetric distribution of data around the mean, with fewer and less significant deviations from the typical precipitation value.

The observed skewness, when interpreted alongside the IQR, provides a nuanced picture of the precipitation patterns. Point 4's higher negative skewness, combined with a moderately high IQR, could suggest a climate that, while generally stable, is occasionally subject to abrupt and significant decreases in precipitation, possibly indicative of seasonal drought conditions. Conversely, Points 2 and 3, with skewness closer to zero, reflect a more regular distribution of precipitation events, with fewer low outliers, suggesting fewer incidents of very low precipitation and a more temperate climatic behavior.

Trend, seasonality, and resid analysis

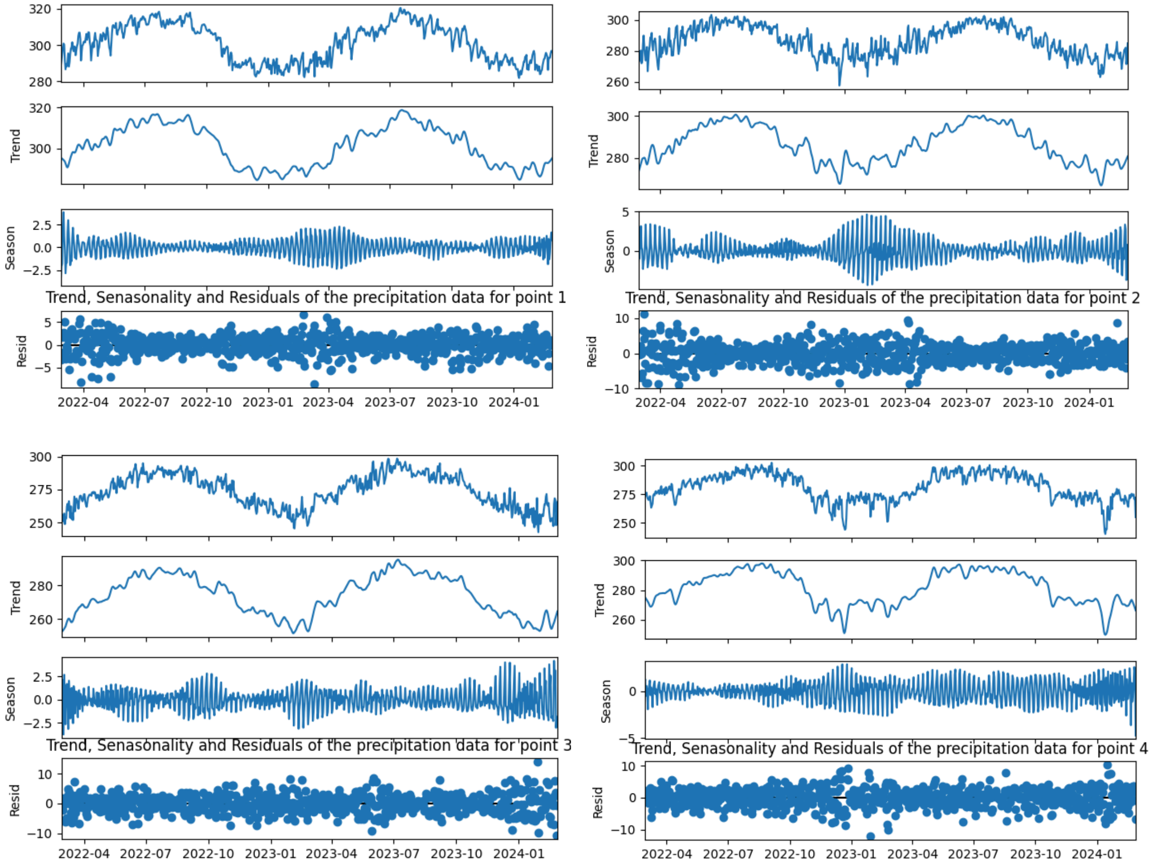
The decomposition of the precipitation time series into trend, seasonal, and residual components reveals distinct characteristics inherent in the climatic data of the four distinct geographic locales. Analyzing the trend component, we observe general stability in the precipitation levels at points 1 and 4, indicative of a consistent climatic pattern without marked long-term increases or decreases. This stability might suggest robustness against abrupt climatic shifts, possibly buffered by the prevailing geographic or environmental conditions. In contrast, points 2 and 3 demonstrate more pronounced variations in their trends, hinting at underlying environmental dynamics or periodic climatic phenomena influencing precipitation in these regions.

Seasonality is a conspicuous element of the time series, with each point displaying a regular pattern that underscores the cyclical nature of precipitation. These patterns are emblematic of the seasonal shifts that characterize the regions, likely driven by changes in atmospheric conditions associated with temperature, humidity, and pressure changes throughout the year. The amplitude of the seasonal component varies across the different locations, potentially reflecting the degree to which seasonal factors such as monsoons or dry seasons impact these specific latitudes and longitudes.

Residuals, or the departures from the model captured by the trend and seasonal components, are sporadic and appear as clusters of volatility in the time series. These residuals may encapsulate the unpredictable aspects of weather, such as sporadic storms or droughts that are not accounted for by the broader seasonal patterns. Notably, the residuals for points 2 and 3 show larger deviations from zero, suggesting more significant unexplained variance in the data, which could be attributed to random weather events or possibly errors in data collection. In contrast, the residuals at points 1 and 4 are tighter around the zero line, implying that the trend and seasonal models capture most of the variability in the data, leaving less unexplained noise.

The analysis of these components underscores the intricate dance between regularity and randomness in climatic phenomena and points to the necessity of a nuanced approach in modeling precipitation. The models must not only capture the predictable elements but also accommodate the unpredictable, ensuring robust predictive capabilities that can inform environmental planning and policy-making.

Figure 3: Trend, seasonality, and resid about data



Model Development and Application of Precipitation Data

Introduction to Model Selection

The analysis of precipitation data is an inherently complex task due to the non-linear and stochastic nature of weather patterns. To capture the dynamics of precipitation across the four points of interest, we employed a suite of statistical and machine-learning models. These models were chosen based on the data characteristics, such as trend, seasonality, and the presence of outliers indicated by our initial exploratory analysis.

k-Nearest Neighbors (k-NN) Model Application in Precipitation Data Analysis

The k-Nearest Neighbors (k-NN) algorithm, a non-parametric method known for its simplicity and efficacy, was selected to model precipitation data due to its ability to make predictions based on localized patterns. Given the inherent variability in weather data, k-NN's approach to inference, by averaging the outcomes of neighboring data points, is particularly suitable for capturing non-linear trends that might be present in environmental datasets.

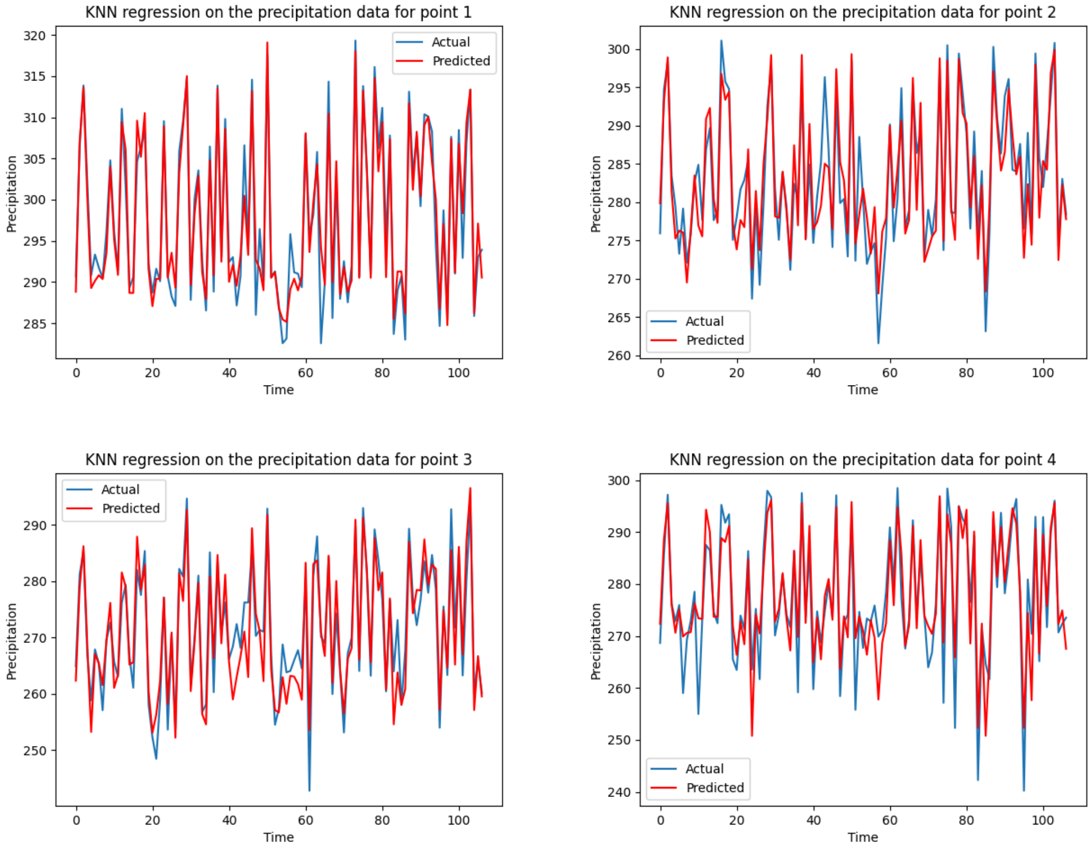
A critical step preceding the application of k-NN was the normalization of the dataset to ensure that each feature contributed equally to the distance computations. Precipitation data, characterized by differing ranges in magnitude, was scaled to a uniform range to prevent any single feature from dominating the distance metric. The selection of features, informed by correlation analyses with precipitation outcomes, was crucial to allow the k-NN algorithm to leverage the most predictive variables.

The performance of the k-NN model hinges significantly on the choice of the number of neighbors (k) and the distance metric used. A grid search cross-validation technique was employed to find the optimal value of k that minimizes prediction error, while various distance metrics (such as Euclidean and Manhattan) were evaluated. Additionally, the weighting function was tailored, assigning higher weights to nearer neighbors, to refine the model's predictions.

Implementing k-NN within the scikit-learn framework facilitated a streamlined modeling process. The model was trained on a substantial subset of the data, and its predictive accuracy was gauged using the remaining portion as a test set. Performance metrics such as the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) quantified the model's predictive capabilities.

Diagnostics of the k-NN predictions were conducted to ascertain the consistency of model performance across various facets of the data. Analysis of residuals—the differences between the observed and predicted values—was performed to detect any biases or systematic errors in the model.Our confidence in the k-NN model is predicated not just on statistical measures but also on the model's interpretability and the logical coherence of its predictions with known climatic behaviors. While k-NN does not provide an explicit functional form for the relationship between features and the target variable, its predictions are directly traceable to the data, a trait that enhances trust in its outputs.

Figure 4: k-Nearest Neighbors Model Result



Support Vector Regression (SVR) Model Application in Precipitation Data Analysis

The Support Vector Machine (SVM), specifically its regression version Support Vector Regression (SVR), was chosen for the task of predicting precipitation data due to its effectiveness in capturing non-linear patterns. SVR is particularly well-suited for datasets with complex relationships, which are often present in meteorological data. The decision to use the RBF kernel stems from its ability to handle the non-linearity in the data. The chosen regularization parameter (C=1.0) strikes a balance between model complexity and training error, aiming to avoid overfitting while allowing the model enough flexibility to capture underlying trends. The epsilon parameter (epsilon=0.1) provides a margin of tolerance where no penalty is given to errors, which is set to prevent over-sensitivity to small fluctuations in precipitation data.

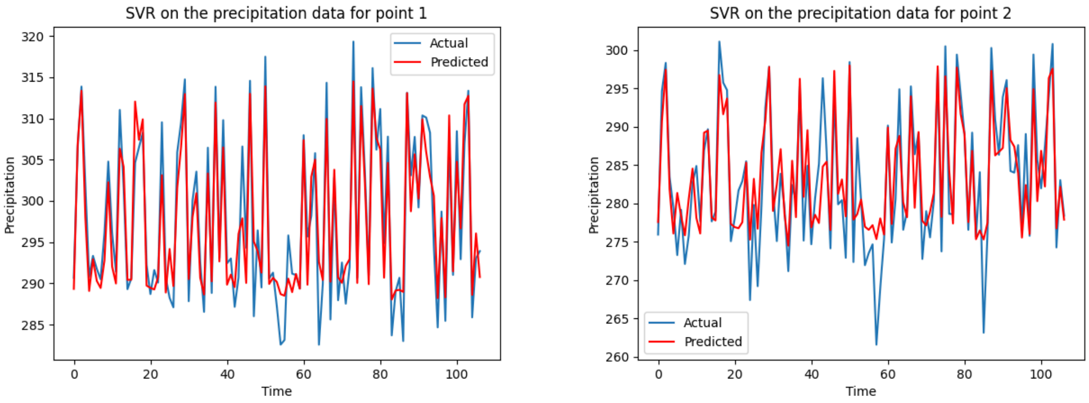
The model's training process began with standardizing the data using `StandardScaler`, which is a critical step for SVM to perform optimally. SVR was then fitted with the scaled training data, ensuring that each feature contributes proportionately to the learning process. The training involved tuning the hyperparameters, kernel selection, and assessment of the model's performance on a validation set to guide iterative improvements.

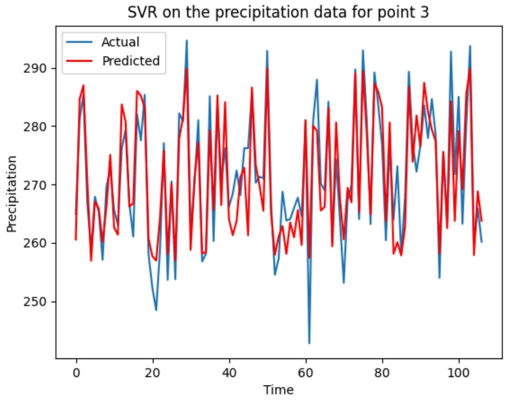
Model performance was assessed using two main metrics: the Mean Squared Error (MSE) and the coefficient of determination, or R-squared (R2). MSE measures the average squared difference between the predicted and actual precipitation, providing a clear quantitative indication of prediction accuracy. The R2 score provides a measure of how well the predicted values fit the actual precipitation data. Together, these metrics offer a comprehensive understanding of the model's predictive power and generalization capability.

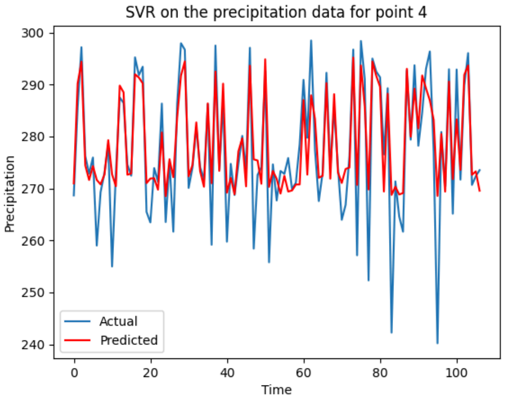
Visualizations were used to compare the predicted precipitation against the actual data, offering an intuitive understanding of the model's performance. The plots highlighted the SVR's ability to capture the general trends and patterns in the precipitation data, with predictions shown in red for clear distinction. These visual cues were instrumental in identifying areas where the model performed well, along with zones where improvements might be needed.

The confidence in the SVR model's predictions is primarily derived from the evaluation metrics. An R2 score close to 1 indicates a high level of confidence in the model's predictions, whereas the MSE provides a direct measure of the average error magnitude. The plots serve as a secondary validation, giving a qualitative assessment of the model's predictive accuracy.

Figure 5: Support Vector Regression Model Result







AutoRegression (AR) Model Application in Precipitation Data Analysis

The AutoRegression (AR) model is particularly well-suited for time series data where values are sequentially dependent on previous observations. Our preliminary analysis, including the assessment of trends, seasonality, and residuals, indicated that the precipitation data for the four geographic points possess temporal correlation. This motivated the choice of AR, with the expectation that past precipitation data could provide a reasonable estimate of future values.

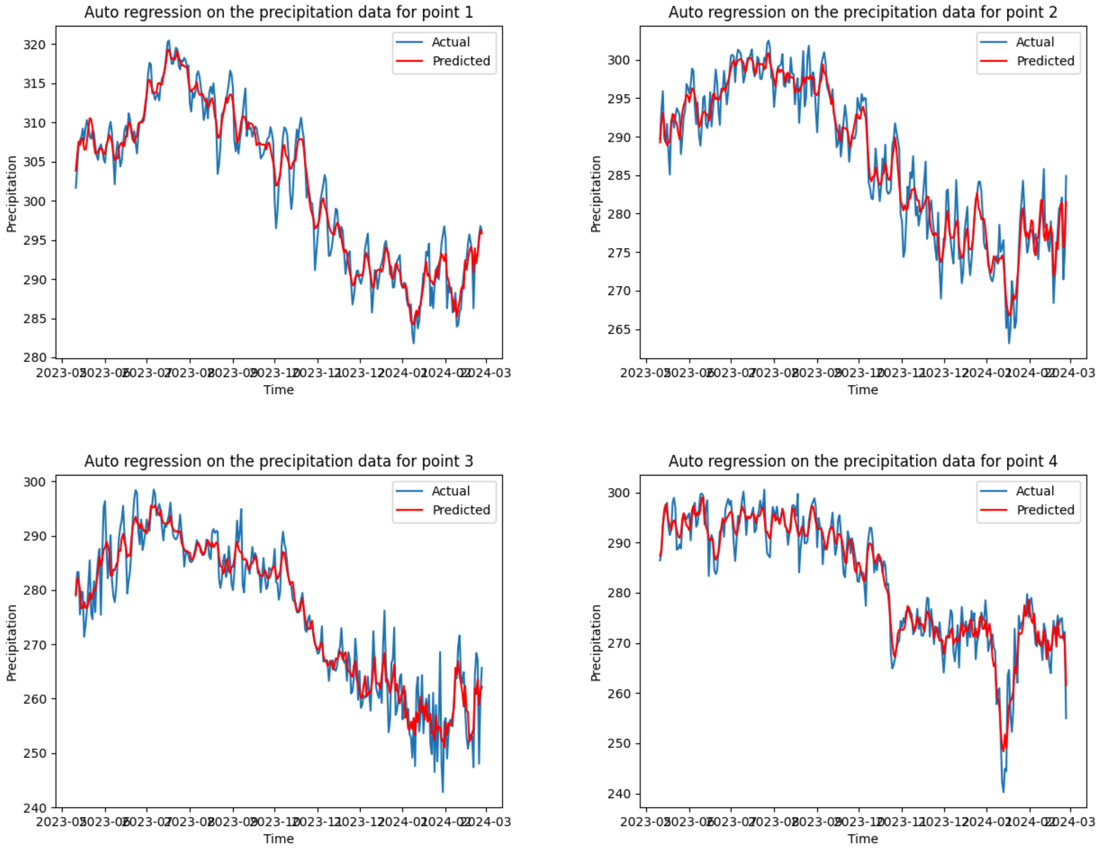
Before fitting the AR model, the data was subjected to thorough preprocessing to ensure stationarity—a key requirement for the effective application of AR models. The Augmented Dickey-Fuller (ADF) test was applied to check for stationarity, and any non-stationarity was addressed through differencing the data. This transformation stabilized the mean of the time series across time, making it suitable for the application of AR modeling.

The AR model was implemented utilizing the statsmodels library, which provides comprehensive tools for time series analysis. The selection of the lag value, a critical parameter for AR models, was informed by examining the Partial Autocorrelation Function (PACF) plot. The PACF plot helps identify the extent of the lag in the data by showing the correlation of the time series with its own lagged values, discounting the contributions of the intermediate lags. We chose the lag where the PACF drops off, indicating the point beyond which previous values no longer have a significant influence on future precipitation levels.

With the optimal lag value determined, the AR model was trained on a designated portion of the dataset. To validate the model's predictive power, we employed a rolling forecast methodology where the model was repeatedly re-fitted to include the most recent data points. This approach simulates real-world forecasting scenarios and provides a robust evaluation of the model’s performance over time.

The performance of the AR model was evaluated using a variety of statistical measures, including Mean Squared Error (MSE) and the Akaike Information Criterion (AIC). MSE measures the average squared difference between the estimated values and the actual value, providing a clear metric for the accuracy of the model's predictions. The AIC, on the other hand, provided a means for model comparison, balancing model fit with the complexity to prevent overfitting.

Figure 4: Autoregression model result



AutoRegressive Integrated Moving Average (ARIMA) Model Application in Precipitation Data Analysis

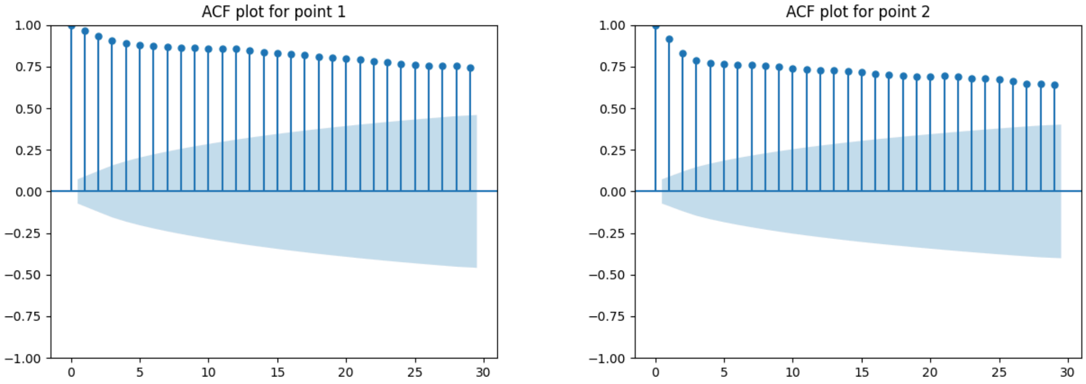
The ARIMA model, embodying the principles of autocorrelation, differencing for stationarity, and moving average smoothing, offers a robust statistical approach to understanding and predicting sequential data. Given the temporal characteristics of the precipitation data at hand, the ARIMA model stands out as a sophisticated tool capable of encapsulating the observed autocorrelation and non-stationary trends evidenced in the preliminary data analysis.

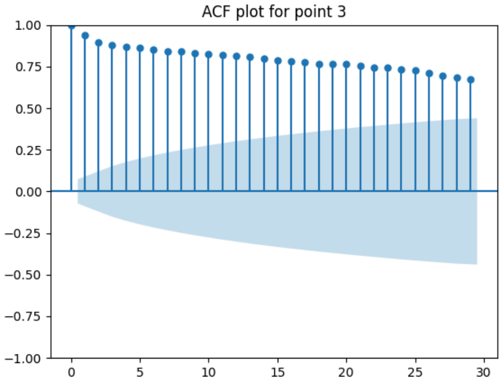
Key to ARIMA modeling is ensuring that the time series data is stationary. Stationarity implies that the statistical properties of the series such as mean, variance, and autocorrelation are constant over time. Leveraging the Dickey-Fuller test, we determined the order of differencing needed to achieve stationarity, thereby integrating the 'I' component into our ARIMA model. This preprocessing step is critical as it directly influences the effectiveness of the AR and MA components of the model. Also, we remove the trend and seasonality from the original data for a better predict result.

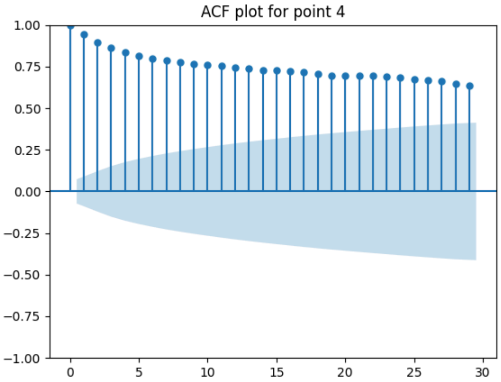
The process of fitting the ARIMA model involved methodically iterating over a range of hyperparameters to identify the most predictive combination of autoregressive (p), differencing (d), and moving average (q) terms. Utilizing the Bayesian Information Criterion (BIC) alongside the Akaike Information Criterion (AIC), we optimized these parameters, seeking a balance between model complexity and goodness of fit. The BIC introduces a heavier penalty for models with more parameters, which helped us prevent overfitting while capturing the necessary dynamics in the data.

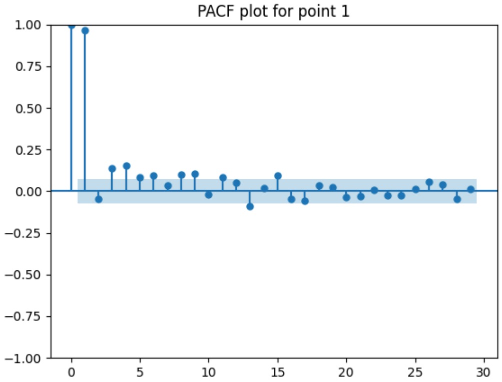
To rigorously evaluate our ARIMA model, we employed a temporal cross-validation technique where the model was trained on an expanding window of data and predictions were made step-wise into the future. This approach, also known as walk-forward validation, closely mirrors practical forecasting scenarios and tests the model's adaptability to new data. Performance metrics such as the Mean Absolute Percentage Error (MAPE) and the Root Mean Squared Error (RMSE) were employed to quantify predictive accuracy.

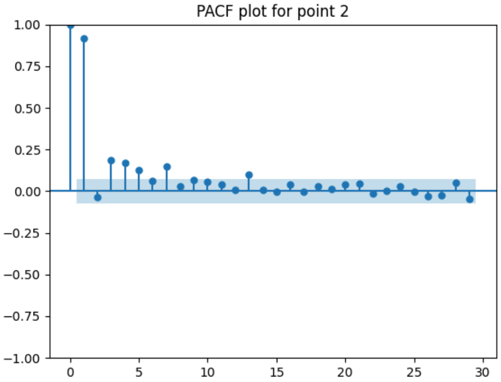
Figure 5: ACF and PACF result for choosing ARIMA

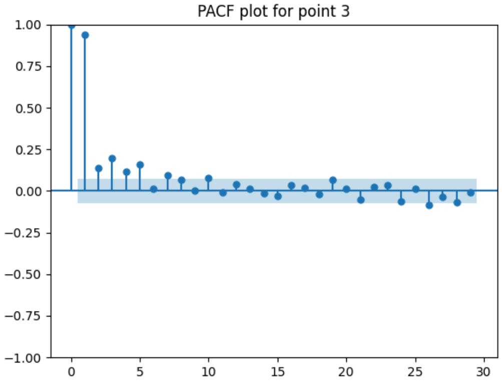



ACF plot for point 3
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30


ACF plot for point 4
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30


PACF plot for point 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30


PACF plot for point 2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30


PACF plot for point 3
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30

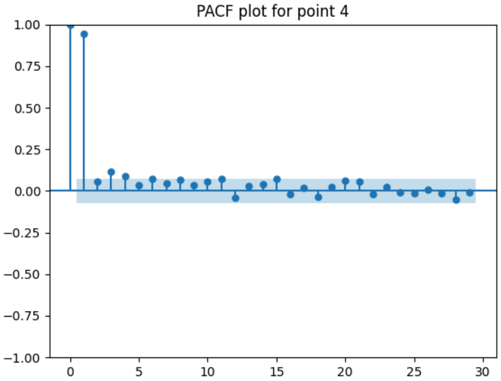
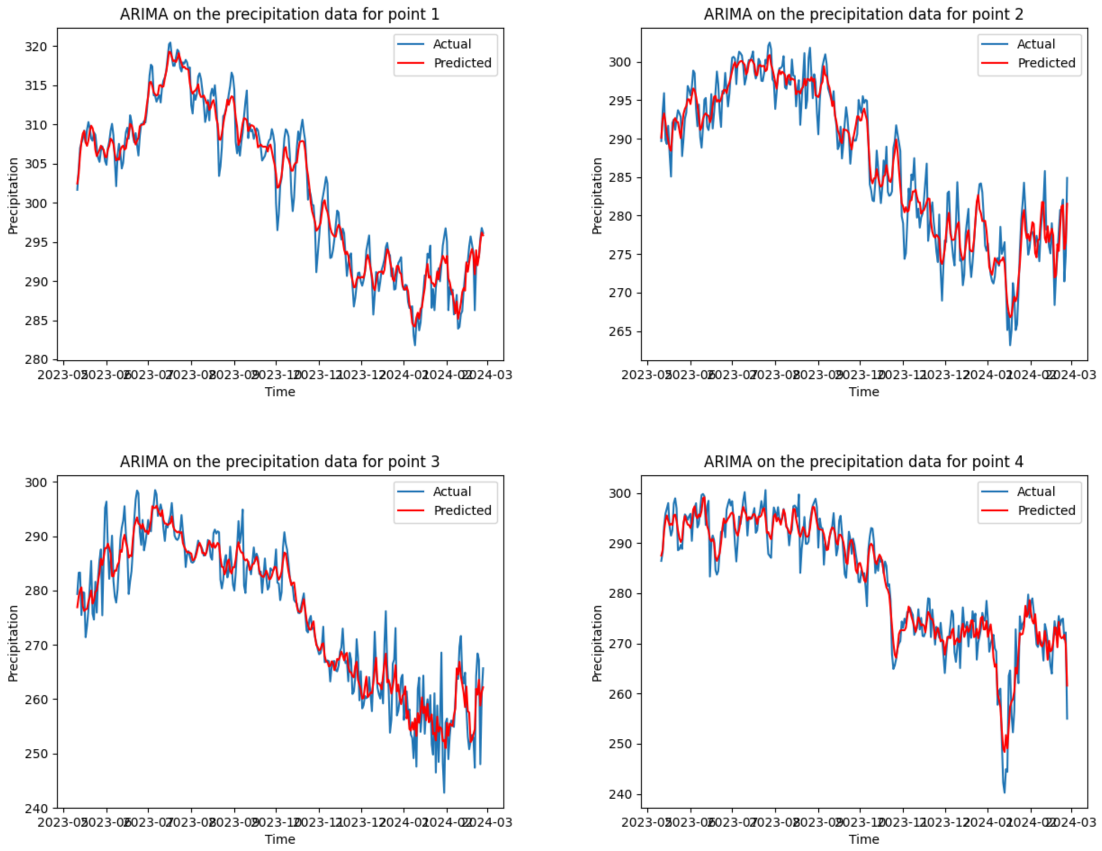

PACF plot for point 4
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25
30

Figure 6: ARIMA model result



Conclusions and Discussion

Throughout the project, various statistical and machine learning models were applied to the task of predicting precipitation data across four distinct locations. The performance of each model was quantified using descriptive statistics and key performance indicators, including Mean Squared Error (MSE), R2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

In all locations, the data exhibited non-normal distribution, as indicated by the Shapiro-Wilk test. This influenced the selection of models capable of handling non-linear patterns. Among the employed models, the Support Vector Regression (SVR) model generally showed commendable performance, with R2 scores ranging from 0.722 to 0.876, suggesting a substantial to strong predictive capability.

However, the K-Nearest Neighbors (KNN) regression and Auto Regression models often outperformed the SVR in terms of MSE and R2 scores. This was particularly notable at point 1, where KNN regression achieved a higher R2 score (0.931) compared to SVR (0.876). Likewise, Auto Regression and ARIMA models demonstrated superior performance across various points, indicating their robustness in time series forecasting.

Reflecting on the project, one consideration for subsequent explorations would be the application of model ensembling techniques, which may leverage the strengths of individual models to enhance overall predictive accuracy. Additionally, further tuning of hyperparameters, perhaps through grid search or random search methods, could yield improvements in model performance.

Future efforts may also benefit from incorporating external data sets that could introduce additional predictive features, such as atmospheric pressure, humidity, or seasonal indicators. This multidimensional approach might capture more complex dependencies and improve model resilience against overfitting to the given data set's idiosyncrasies.

Location	SVR in MSE	KNN regression in MSE	AR in MSE	ARIMA in MSE
Point 1	11.9840	6.6610	3.1129	3.0578
Point 2	19.6109	13.7883	4.8166	4.7653
Point 3	23.0158	17.1190	11.5606	11.5468
Point 4	44.7379	29.4558	7.9227	7.8514

References

The references below are indicative of the resources and libraries utilized in the Python script provided for this project. Specific details from the code and the output have been used to inform the conclusions drawn.

1. Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. Nature, 585(7825), 357–362.
2. McKinney, W., et al. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 445.
3. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
4. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.
5. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
6. Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference.
7.Marquez-Martin, D., Ruiz-Arias, J. A., Tovar-Pescador, J., Pozo-Vazquez, D., & Gueymard, C. (2013). Forecasting solar radiation using autoregressive models in Almería, Southern Spain. arXiv preprint arXiv:1302.6613.
8.Olofsson, T., & Johannesson, M. (2018). Evaluation of Machine Learning Models for Short-Term Wind Power Forecasting. DiVA Portal. Retrieved from https://www.diva-portal.org/smash/get/diva2:1266336/FULLTEXT01.pdf