# Comparative Analysis of ARIMA and STARIMA Models for Temperature Time Series Forecasting

March 2025

## Abstract

This study investigates the effectiveness of incorporating spatial dependence into time series forecasting of monthly temperature data using a Space–Time Autoregressive Integrated Moving Average (STARIMA) framework compared to an Autoregressive Integrated Moving Average (ARIMA) model. Monthly average temperature observations from 274 weather stations across the Mid-Atlantic United States were analyzed over the period 2000–2015. Exploratory spatio-temporal analysis revealed strong seasonal patterns, temporal persistence, and clear spatial structure associated with latitude, elevation, and regional geography. Forecast performance was evaluated using ARIMA models applied independently to each station, a seasonal naïve baseline, and a STARIMA model incorporating spatial relationships defined via a k-nearest neighbors weight matrix. Results show that ARIMA models struggled to outperform a simple seasonal benchmark, highlighting the limitations of non-spatial approaches in this context. In contrast, STARIMA achieved consistently lower forecast errors across multiple evaluation metrics, demonstrating that modelling spatial dependence provides measurable improvements beyond seasonality alone. Residual diagnostics indicate that some seasonal dependence remains unmodelled, suggesting opportunities for further refinement. Overall, the findings support the use of spatio-temporal statistical models for regional climate forecasting and illustrate the practical benefits of integrating spatial structure into time series analysis.

## 1. Introduction

Understanding regional climate patterns and their impacts on ecosystems, infrastructure, and human activities is essential for applications such as environmental planning, energy demand forecasting, and climate risk assessment. This project focuses on analyzing temperature data from the Mid-Atlantic United States – Delaware, Maryland, New Jersey, Pennsylvania, Virginia, and West Virginia – over the period January 2000 to January 2015. The objective is to employ and compare two statistical forecasting models: ARIMA (Autoregressive Integrated Moving Average) and STARIMA (Space-Time Autoregressive Integrated Moving Average), assessing their effectiveness in modeling and predicting temperature variations.

The methods employed in this study, ARIMA and STARIMA, are widely used for time series forecasting. ARIMA models are widely used for univariate time series forecasting and are particularly effective for data exhibiting strong seasonal structure. They model temporal dependence, allowing future values to be forecast based on past observations [1]. It can be used to predict future points by leveraging past data points. ARIMA is broken into three components: AR($p$) – the autoregressive component which is the number of lag observations included in the model (lag order), I($d$) – the integration component which is the number of times that the raw observations are differenced (degree of differencing), and MA($q$) – the size of the moving average window (order of moving average)[1].

STARIMA extends the ARIMA framework by explicitly incorporating spatial dependencies, making it particularly suitable for geographic data analyses[2]. Recent studies, such as those by Rathod et al.[3] and Lin et al.[4], have demonstrated the effectiveness of these models in climatological and traffic flow studies, highlighting their robustness in capturing seasonal patterns and spatial heterogeneity in different types of data. This study tests the hypothesis that incorporating spatial dependence through a STARIMA framework improves medium-term temperature forecasting accuracy compared to independent ARIMA models applied at individual weather stations.

The dataset used in this study is sourced from the National Oceanic and Atmospheric Administration (NOAA), known for its comprehensive climate data records. The "Global Summary of the Month" dataset[5] provides monthly average of meteorological variables, gathered from various weather stations, updated weekly and spanning from 1763 to present. The data include details such as unique station identifiers, elevation, latitude/longitude, and monthly average temperature readings.

To prepare this dataset for analysis, extensive preprocessing was necessary. This involved selecting the relevant states, filtering out files without an average temperature column or files with no data inside the 2000-2015 range, and consolidating each weather station file into one large file grouped by state. To address missing temperature values in the data, a method of imputing missing temperature data was used[6], averaging adjacent temperature records (the two temperature records behind and in front of the missing cell) to maintain data integrity.

## 2. Exploratory Spatio-Temporal Data Analysis

### *2.1. Non-spatio-temporal analysis*

Figure 2.1 has a bimodal distribution – there are two peaks. This suggests that there are two dominant groups of temperature values in the dataset, potentially corresponding to warmer months and cooler months. The distribution appears slightly skewed to the right, indicating there are more instances of higher temperatures compared to lower temperatures beyond the mean.

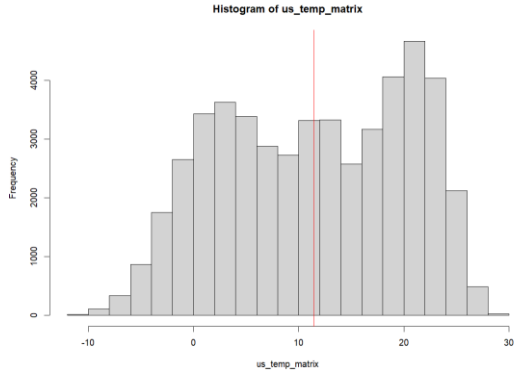


Figure 2.1: Histogram of temperature data across entire dataset

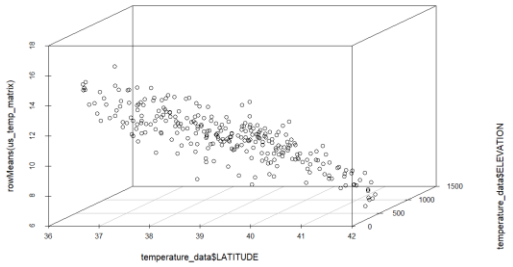


Figure 2.2: 3D scatterplot of temperature data, latitude, and elevation

In Figure 2.2, there is a negative correlation between temperature and latitude: as latitude increases, temperature generally seems to decrease. This makes sense as higher latitudes are further from the equator, leading to lower temperatures. It also appears that higher elevations are associated with lower temperatures.

### *2.2. Temporal analysis*

Figure 2.3 shows a clear and consistent seasonal cycle. This is expected, as the climate of the Mid-Atlantic U.S. has distinct seasonal changes. The range of temperature fluctuations is consistent, and the frequency of the cycles is annual, confirming seasonality in the data.

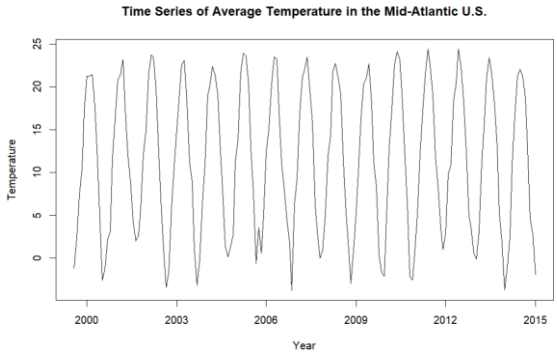


Figure 2.3: Average temperature from 2000-2015

Figure 2.4 again confirms the seasonality of the data, seeing that there is consistent and uniform cycling between warmer and colder temperatures as time progresses. There is also uniformity across the stations, suggesting that regional climate conditions affect the stations similarly throughout the year.

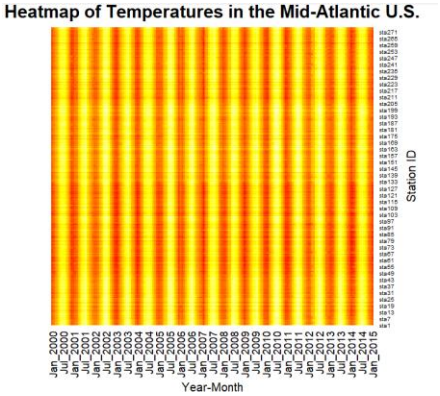


Figure 2.4: Heatmap of temperatures in the Mid-Atlantic

### *2.3. Spatial analysis*

Since this dataset contains monthly average temperature, Figure 2.5 shows the monthly temperatures for January, April, July, and October across the area of study in 2014. The larger circles indicate higher elevation. There is a geographical pattern in the temperature distribution. The coastal areas, along the Atlantic Ocean, tend to show different temperature patterns compared to inland regions.

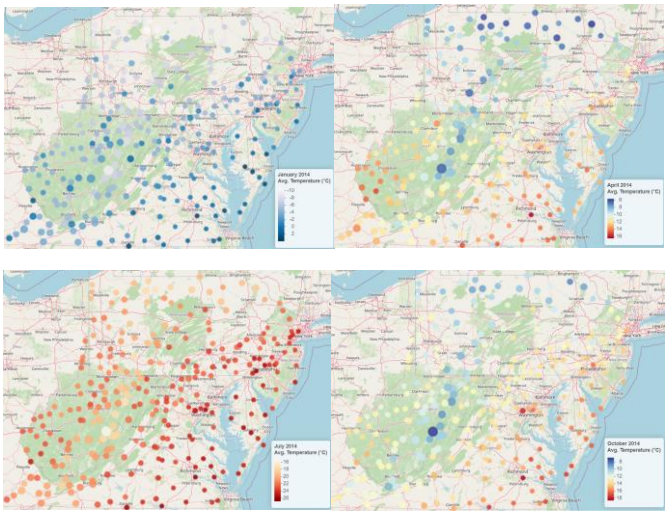

Figure 2.5: Temperatures in the Mid-Atlantic U.S. in 2014

Urban Heat Island Effect can be seen in cities such as Washington D.C., Philadelphia, and New York City, especially in July, where there are clusters of high temperatures in those cities. In the western portion of these maps, the weather stations in the Appalachian Mountains consistently recorded lower temperatures in all four months due to their higher elevation.

### *2.4. Autocorrelation analysis*

#### *2.4.1. Temporal autocorrelation analysis*

The PMCC (0.842) in Figure 2.6 indicates a strong positive autocorrelation in the data, suggesting that temperatures from one month are closely related to temperatures from an adjacent month. Temperature is relatively stable from month to month in the Mid-Atlantic, as reflected in the strong positive autocorrelation. The clustering of points along the regression line may also indicate seasonality.

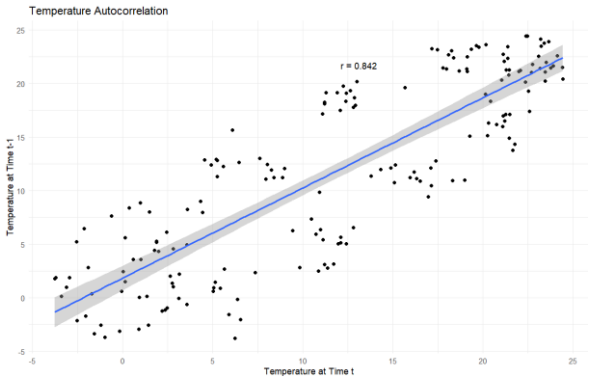


Figure 2.6: Temperature autocorrelation

Figure 2.7 indicates seasonality in the temperature data, seeing that there are periodic spikes.

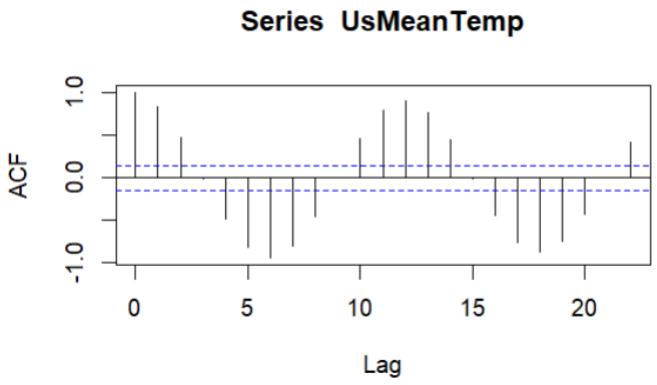


Figure 2.7: ACF of mean temperature

#### *2.4.2. Spatial autocorrelation analysis*

As the temperature observations are recorded at fixed weather station locations, the data can be treated as spatial point data; therefore, a semi-variogram was used to examine spatial autocorrelation. Figure 2.8 indicates that as distance increases, semi-variance increases, suggesting that data points that are farther apart are less similar than those that are closer together. This is also indicative of positive spatial correlation, where closer locations tend to have more similar temperatures.

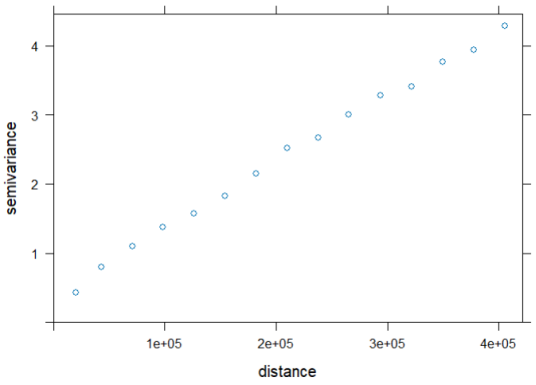


Figure 2.8: Semi-variogram of temperature data

In Figure 2.9, where semi-variance is examined in different directions, there is more semi-variance in the horizontal direction (0°) than in the vertical direction (90°). This could be due to the landscape changing from coastal plains to mountain ranges travelling from East to West.

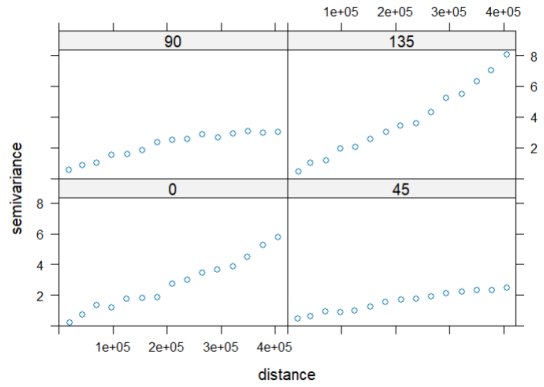


Figure 2.9: Directional semi-variance in temperature data

#### *2.4.3. Space-time autocorrelation analysis*

As the spacing of the data points is irregular, k-nearest neighbors (KNN) was used to define spatial relationships among the data points (Figure 2.10).

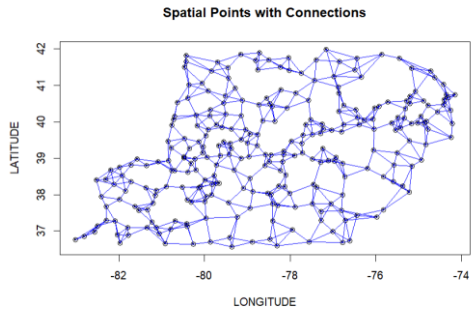


Figure 2.10: KNN connectivity plot

The number of neighbors is set to 4, which is the smallest k-value that connects all points. This allowed for the creation of a spatial weight matrix[7], since each point can now be analyzed in context with a fixed number of nearby points.

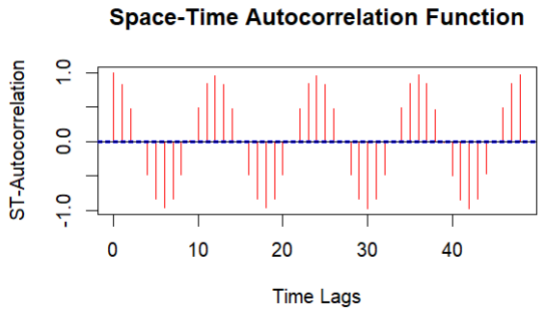


Figure 2.11: STACF of temperature data

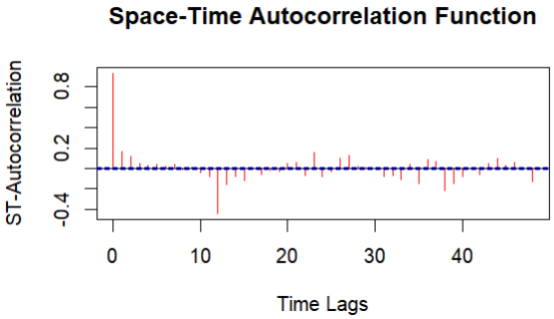


Figure 2.12: STACF seasonally differenced

After removing seasonal patterns (Figure 2.12) through differencing (lag of 12 months), the overall level of autocorrelation is reduced across all lags compared to the original STACF (Figure 2.11), indicating the seasonal component has been mitigated. There are significant spikes around lags 1 and 12 in Figure 2.12.

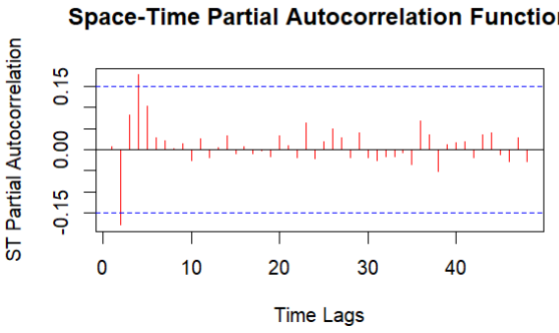


Figure 2.13: STPACF of temperature data

Figure 2.13 shows significant partial autocorrelations around the first and third lags. These lags (one and three months apart) have a strong influence on the series, likely indicative of short-term persistence in temperature changes, possibly driven by seasonal shifts captured within those lag intervals.

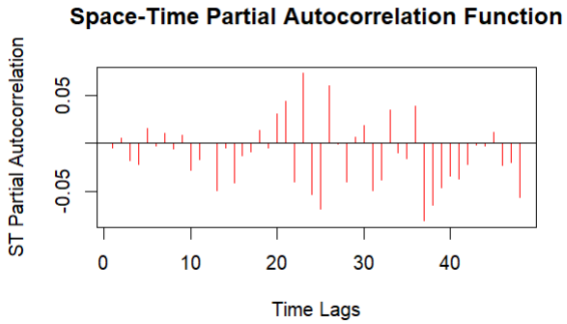


Figure 2.14: STPACF seasonally differenced

In Figure 2.14, all the partial autocorrelations fall within the lines of confidence, meaning they are not statistically significant. Here, the differencing process has made the data more stationary by removing seasonal influence, as there is no significant autocorrelation at these lag intervals. Building on the exploratory analysis and identified spatio-temporal dependencies, the following section evaluates the performance of ARIMA, STARIMA, and a seasonal naïve baseline model.

## 3. Methodology and Results

### *3.1. ARIMA*

#### *3.1.1. Experimental setup*

The data used in this project is multivariate, there are data from 274 different weather stations used. ARIMA does not work well with multivariate time series, and it would take too much time to tailor each station's ARIMA model parameters to find the best fit. To understand model behavior across a diverse set of locations, a loop was created to fit an ARIMA model to each station using auto.arima. First, indices were defined to split the data into training and testing sets. An 80/20 training/test split was chosen. The training set was the first 12 years of data (01/2000-01/2012). The final 3 years (02/2012-01/2015) were used to compare the model prediction vs. the actual temperatures. A for loop was used to iterate over each station, extracting the training data from the first 12 years, then extracting the remaining testing data. Inside the loop, auto.arima automatically fits an ARIMA model, selecting the best model based on AIC (Akaike Information Criterion) and BIC (Bayesian

Information Criterion)[8]. Seasonal = TRUE was set to consider the seasonal components in the data. The forecast function was then used to generate a forecast from the fitted ARIMA model.

### *3.1.2. Results*

Figure 3.1 shows two examples of the effectiveness of ARIMA forecasting on two stations. For Station 2, the auto-assigned parameters appear to capture the dominant seasonal structure of the series, resulting in forecasts that closely track observed temperatures. For Station 29, the model exhibits reduced accuracy during periods of extreme temperature variations, failing to capture extreme highs and lows. Forecast performance was consistently varied across stations. The average NRMSE, MAE, and sMAPE were calculated to gauge the performance of this methodology.

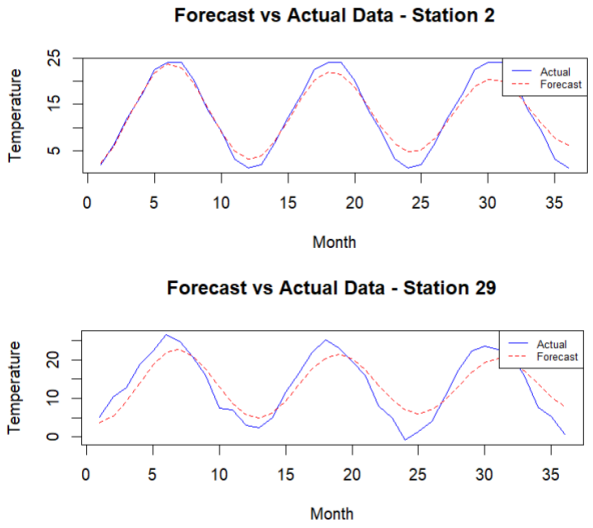


Figure 3.1: Examples of ARIMA forecasting

### *3.1.3. Performance assessment*

Forecast performance was evaluated using Normalized Root Mean Squared Error (NRMSE), Mean Absolute Error (MAE), and Symmetric Mean Absolute Percentage Error (sMAPE). NRMSE was used to allow comparison of forecast accuracy across stations with differing temperature ranges[9], while MAE provides an interpretable measure of average absolute error in degrees Celsius[10]. sMAPE was selected in place of MAPE to ensure symmetric treatment of over- and under-predictions and to avoid instability arising from zero or near-zero values in the data[11]. The temperature values in this dataset range from -5°C to 25°C. An NRMSE of about 32% (Table 1) means that the average size of the forecast errors is around 32% of the average value of the actual temperatures observed in the dataset.

| Avg. NRMSE | Avg. MAE | Avg. sMAPE |
|---|---|---|
| 0.321156 | 2.899523 | 50.08913 |

Table 1: Performance metrics of ARIMA

This level of prediction accuracy is moderate and indicates substantial scope for improvement. The same can be said for MAE, as the predictions are, on average, off by about 2.89°C from the true values. Given the observed temperature range, an average error of approximately 3°C represents a relatively large deviation from observed values. sMAPE suggests that, on average, the forecasts deviate from actual values by about 50.08%, which indicates limited forecast accuracy.

## *3.2. STARIMA*

### *3.2.1. Experimental setup*

For the STARIMA model, KNN was used to define spatial relationships (Figure 2.10) as the point data were previously unrelated, and a spatial weight matrix was created from that to be able to model spatial dependencies. The parameter selection was guided by autocorrelation analyses: the STACF for the seasonally differenced data (Figure 2.12) shows a significant spike at lag 1, justifying a moving average parameter (q=1). The STPACF plot for the same data (Figure 2.14) displayed no significant lags above the confidence lines, indicating that no autoregressive terms (p=0) were necessary. Seasonal differencing with period 12 (D=1, s=12) was applied to remove annual seasonality.

### *3.2.2. Results*

The STARIMA model (Figure 3.2) demonstrates improved forecast performance over the three-year prediction period. Forecast accuracy is highest during transitional seasons, particularly spring and autumn, where temperatures gradually increased and decreased. It did not perform as well in forecasting the colder temperatures, though it seems there may have been some unusually warm temperatures in those winter months. The model appears less effective at capturing abrupt temperature

fluctuations. Forecast accuracy was highest during warmer months, although some overestimation of peak temperatures was observed. It did particularly well in the last few months of the forecasting, with very close agreement between predicted and observed values toward the end of the forecast horizon, which may indicate a degree of overfitting.

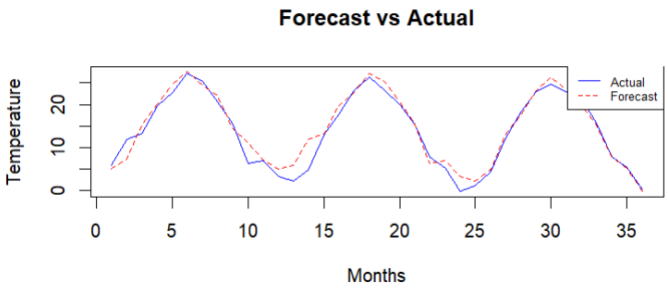


Figure 3.2: STARIMA forecasting – station 1

### *3.2.3. Performance assessment*

An NRMSE value of approximately 0.195 (Table 2), which indicates that the forecast error is about 19.5% of the average observed temperature values in the dataset, suggests a relatively efficient model performance in capturing temperature variation in the data. In the context of this data, a value below 0.2 indicates relatively strong predictive performance, implying the model does a good job of forecasting temperatures. An MAE of 1.42°C suggests that the average absolute error in the temperature forecasts by the STARIMA model is about 1.42°C.

| NRMSE | MAE | sMAPE |
|---|---|---|
| 0.1951761 | 1.422565 | 32.50108 |

Table 2: Performance metrics of STARIMA

This is a decent result, given the range of temperatures in the dataset, although this error magnitude remains non-negligible in the context of monthly temperature variability. A sMAPE of 32.5% suggests moderate accuracy of the model, and there is room for improvement, potentially in reducing systematic biases or model errors. The Box–Ljung test returned a p-value of 0.00089, indicating statistically significant residual autocorrelation remains in the model, suggesting the STARIMA specification does not fully capture the temporal dependence structure.

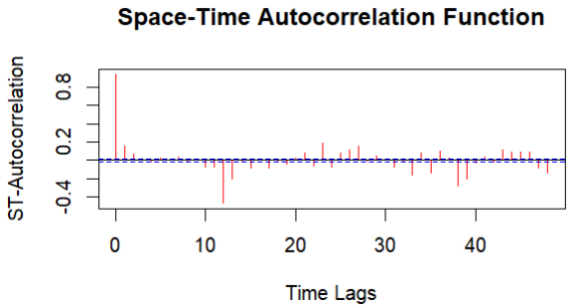


Figure 3.3: STACF of residuals


Figure 3.4: Histogram of residuals – station 1

It appears there is still a significant spike at lag 12 (Figure 3.3), suggesting that some seasonal dependence remains unmodelled. However, there is a significant drop after time lag 1. Figure 3.4 also shows about 50% of the residuals are within 2.5 and almost all the residuals within 5.

### *3.3. Baseline model: seasonal naïve forecast*

To provide a meaningful baseline for evaluating forecast performance, a seasonal naïve model was implemented. This approach predicts each monthly temperature value using the observed temperature from the same month in the previous year, exploiting the strong annual seasonality evident in the dataset while requiring no parameter estimation.

| Avg. NRMSE | Avg. MAE | Avg. sMAPE |
|---|---|---|
| 0.1956788 | 1.421893 | 32.45935 |

Table 3: Performance metrics of seasonal naïve forecast

The seasonal naïve model achieved an average NRMSE of 0.196, MAE of 1.42 °C, and sMAPE of 32.46% (Table 3) indicating that simple year-to-year persistence explains a substantial proportion of temperature variability in the Mid-Atlantic region. Compared to the ARIMA model, which produced notably higher error values, the seasonal naïve baseline outperformed, suggesting that ARIMA struggled to outperform a simple seasonal

benchmark when spatial information was ignored. In contrast, the STARIMA model achieved marginally lower errors than the seasonal naïve model across all metrics, demonstrating that incorporating spatial dependence provides measurable predictive improvement beyond what can be achieved through seasonality alone. These results confirm that the improved performance of STARIMA reflects genuine modelling gains rather than the exploitation of strong seasonal structure in the data.

## 4. Discussion

### *4.1. Model comparison*

The STARIMA model performed significantly better than the ARIMA model in terms of forecasting accuracy. STARIMA outperformed ARIMA across all three evaluation metrics (NRMSE, MAE, and sMAPE). This improvement is likely attributable to STARIMA's ability to model both spatial and temporal dependence structures. ARIMA produced smoother, more general forecasts, which limited its ability to capture sharp increases or decreases in temperature. This suggests that incorporating spatial relationships provides a significant advantage in geographical data analysis, especially in regions where temperature variations are influenced by spatial factors such as elevation and proximity to water bodies. ARIMA was found to be easier to implement, being well-suited for quick analyses, especially with the auto.arima function. However, its inability to account for spatial dependencies limits its utility in geographical studies. STARIMA has higher computational demands and is a longer overall process as it is more complex. Despite this, it provided improved forecasting accuracy by modeling the spatial correlations among data points, something crucial for the accurate representation of temperature data. These findings support the stated hypothesis that incorporating spatial dependence improves medium-term temperature forecasting accuracy.

### *4.2. Limitations and model improvements*

Several limitations of this study also point toward opportunities for future model improvement. The primary limitation was most likely the handling of the missing data and the potential bias introduced by the imputation method used for this project. More robust data collection would benefit this project. Future studies could also improve this by employing more sophisticated imputation techniques or by integrating additional climatic factors that influence temperature. Since some of the temperature values were engineered, the models have been influenced, potentially to be worse off. Auto.arima is useful for completing a large stack of ARIMA models, but the parameters it assigns are not always the best available for the series. Working though each series and tailoring the parameters to each series would most likely have resulted in better ARIMA performance. Experimenting with different spatial weight matrices could potentially improve STARIMA's performance. Finally, more training data would be useful for improving both the ARIMA and STARIMA models. Due to computational and time constraints, 15 total years of data were used. The models likely would have improved performance with more training data.  Overall, this study demonstrates the practical value of spatio-temporal statistical modelling for regional climate analysis while highlighting the importance of aligning model complexity with data structure and forecasting objectives.

## 5. References


1. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time Series Analysis: Forecasting and Control. Hoboken, NJ: John Wiley & Sons, Inc; 2016.

2. Pfeifer PE, Deutsch SJ. A three-stage iterative procedure for space-time modeling. Technometrics. 1980 Feb;22(1):35. doi:10.2307/1268381

3. Rathod S, Gurung B, Singh KN, Ray M. An improved Space-Time Autoregressive Moving Average (STARMA) model for Modelling and Forecasting of Spatio-Temporal time-series data. Journal of the Indian Society of Agricultural Statistics. 2018 Oct 20;72:239–53.

4. Lin S-L, Huang H-Q, Zhu D-Q, Wang T-Z. The application of space-time Arima model on traffic flow forecasting. 2009 International Conference on Machine Learning and Cybernetics. 2009 Jul;3408–12. doi:10.1109/icmlc.2009.5212785

5. Lawrimore JH, Ray R, Applequist S, Korzeniewski B, Menne MJ. Global summary of the month (GSOM), version 1 [Internet]. 2016 [cited 2025 Apr 4]. Available from: https://doi.org/10.7289/V5QV3JJ5

6. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology. 2006 Oct;59(10):1087–91. doi:10.1016/j.jclinepi.2006.01.014

7. Cheng T, Wang J, Haworth J, Heydecker B, Chow A. A dynamic spatial weight matrix and localized space–time autoregressive integrated moving average for network modeling. Geographical Analysis. 2014 Jan;46(1):75–97. doi:10.1111/gean.12026

8. Singh A. Build High Performance Time Series Models using Auto Arima in Python and R [Internet]. 2024 [cited 2025 Apr 4]. Available from: https://www.analyticsvidhya.com/blog/2018/08/auto-arima-time-series-modeling-python-r/

9. Bi L, Feleke A->Genetu, Guan C. A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration. Biomedical Signal Processing and Control. 2019 May;51:113–27. doi:10.1016/j.bspc.2019.02.011

10. What is mean absolute error (MAE) in time series forecasting? [Internet]. [cited 2025 Apr 4]. Available from: https://milvus.io/ai-quick-reference/what-is-mean-absolute-error-mae-in-time-series-forecasting

11. Smape - symmetric mean absolute percentage error [Internet]. [cited 2025 Apr 4]. Available from: https://permetrics.readthedocs.io/en/latest/pages/regression/SMAPE.html

12. Cheng T, Haworth J. (2024) Spatio-temporal Analytics in R – Chapter 2: Spatio-temporal Dependence and Autocorrelation, Chapter 3: Statistical modelling of time series and spatio-temporal series.

https://moodle.ucl.ac.uk/pluginfile.php/8470339/mod_resource/content/24/_book/index.html