1 Introduction

Toronto experiences distinct seasonal variations due to its temperate climate, making accurate temperature forecasting critical for various sectors within the city. Reliable temperature forecasts provide valuable insights for applications in energy management, agriculture, and public health. For example, Toronto's cold weather response plan is designed to mitigate adverse health impacts caused by extreme winter conditions (City of Toronto (2021)). Accurate forecasts enable better preparedness, enhancing the plan's effectiveness and ensuring more residents receive timely assistance during cold weather events. Furthermore, precise temperature predictions contribute significantly to managing city infrastructure and operations (City of Toronto (2012)). For instance, accurately forecasting extreme temperatures allows for proactive resource allocation, such as increasing public transit availability during adverse weather, efficiently managing energy demands, and preventing outages.

The primary goal of this study is to develop a time series forecasting model for Toronto's monthly average temperatures using historical data from January 1999 to December 2019. By performing a detailed time series analysis of this data, this study aim to have a deep understanding of the seasonal patterns in the data and identify the key components required for a Seasonal Autoregressive Integrated Moving Average (SARIMA) model, such as the non-seasonal orders (p, d, q) and seasonal orders (P, D, Q). Using the Box-Jenkins methodology, the SARIMA model will be constructed, and its predictive performance will be evaluated by comparing forecasted temperatures to actual observations. The modeling process involves data preprocessing, exploratory data analysis, systematic model selection, diagnostics checking, and forecasting. The report will detail each step, present the results obtained, and discuss any limitations identified during the analysis.

2 Modeling

The dataset used in this study was obtained from Kaggle and includes temperature data for 1000 cities from 1980 to 2020, originally sourced from the Copernicus Climate Service. It contains daily average temperatures (in degrees Celsius) for 100 cities worldwide. For this analysis, data for Toronto from January 1999 to December 2019 was extracted, and daily averages were converted into monthly averages. This study specifically focuses on analyzing and forecasting monthly average temperatures for Toronto.

2.1 Data Preprocessing

The dataset contains 252 observations with two variables: Date and Avg. Temperature. The dataset has no missing values or duplicated dates. Visual checks confirmed that there are no extreme outliers, and temperature values range between -12.01°C and 23.96°C. The dataset was then split into a training set (1999–2017, 228 observations) and a test set (2018–2019, 24 observations), after which both sets were converted into time series formats.

2.2 Exploratory Data Analysis (EDA)

The time series plot for the training dataset (Figure 1) shows a strong seasonal pattern, with peak temperatures occurring around mid-year (summer) and lower temperatures in winter. The series oscillates between approximately -10°C and 25°C, indicating an annual temperature cycle in Toronto. There is no clear upward or downward trend over the studied period.

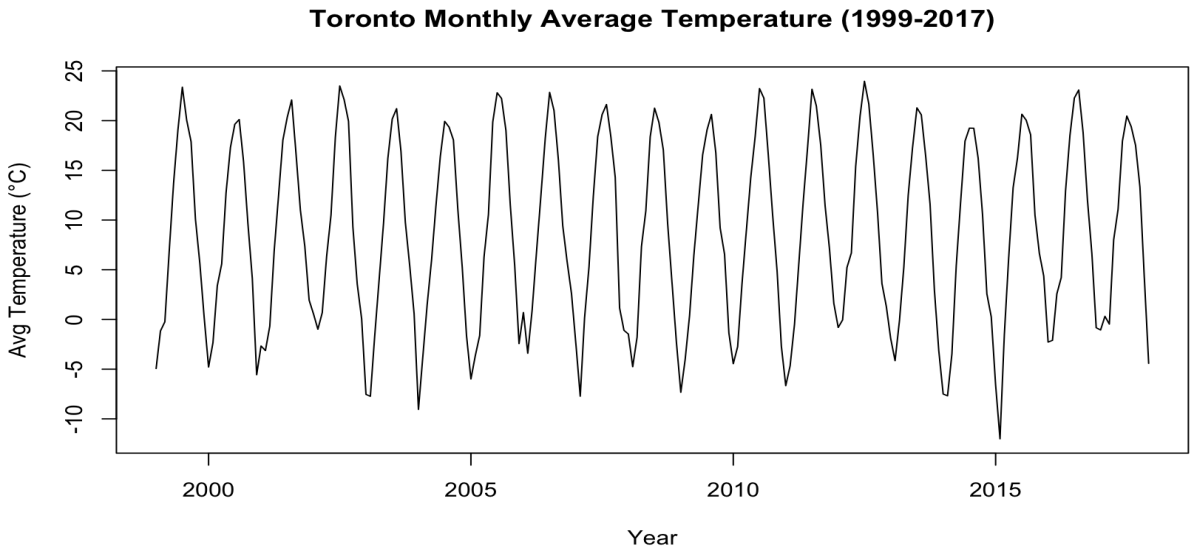

Figure 1: Training Data Time Series

To further explore the data, the training series was decomposed using the STL (Seasonal-Trend Decomposition using Loess) function, separating it into trend, seasonal, and remainder components. The remainder component may include cyclical effects. The seasonal component of the decomposition plot (Figure 2) confirms that seasonality has the strongest influence on the data. The trend component relatively flat, indicating no significant long-term increase or decrease, although there are short-term changes. The spikes seen in the remainder component around 2015 might represent random weather events, as the dataset does not include extreme outliers.

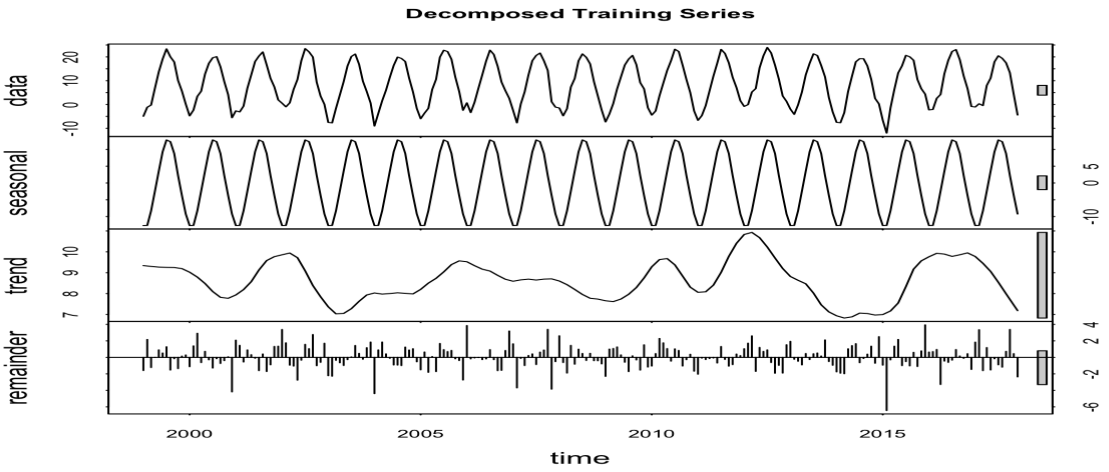

Figure 2: Decomposition of Training Series

2.3 Model Specification

The Augmented Dickey-Fuller (ADF) test, with a p-value of 0.01, suggests that the series is stationary. Therefore, no non-seasonal differencing is needed (d = 0).

The ACF plots (Figure 3) provide clear evidence of strong seasonal patterns in the data, as shown by the significant spikes in the ACF at lags 12, 24, and 36, corresponding to the annual temperature cycle. The PACF plot (Figure 4) indicates that partial autocorrelations drop close to zero after lag 6; however, fitting an AR(6) model would be overly complex. There is a large positive partial autocorrelation spike at lag 1, suggesting an AR(1) component as a suitable starting point. Additionally, the PACF shows seasonal spikes at lags 12 and 24 that slightly exceed the significance bounds. These observations imply that seasonal influences need to be addressed separately using seasonal differencing. Removing seasonal effects could also help more accurately identify any underlying moving average components that may otherwise be masked by dominant seasonal patterns.

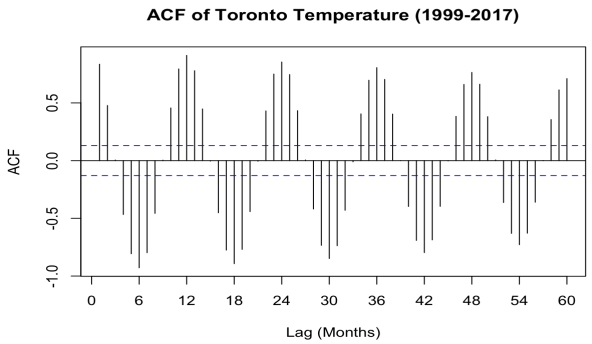

Figure 3: ACF of Training Series

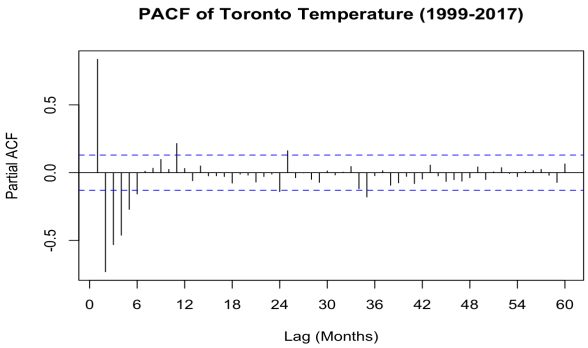

Figure 4: PACF of Training Series

The series remains stationary after seasonal differencing, as indicated by the ADF test p-value still being less than the significance level of 0.05. Seasonal differencing effectively removes the dominant annual pattern. The ACF plot after seasonal differencing (Figure 5) shows a large negative spike at lag 12, suggesting a seasonal MA component with order $Q = 1$. Additionally, the sharp decline in the ACF after lag 1 suggests an MA(1) component.

The PACF plot after seasonal differencing (Figure 6) closely resembles the original PACF. Based on the earlier analysis of the original PACF plot, starting with an AR(1) model and seasonal AR with order $P = 1$ seems appropriate.

After carefully analyzing the ACF and PACF plots both before and after seasonal differencing, the tentative SARIMA model orders were determined to be $(p = 1, d = 0, q = 1) \times (P = 1, D = 1, Q = 1)$.

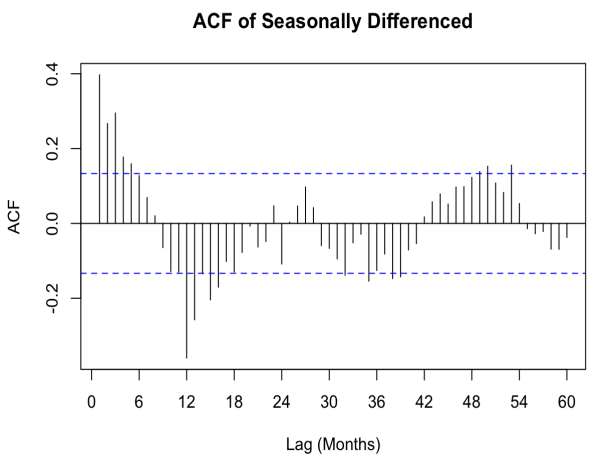

Figure 5: ACF after Seasonal Differencing

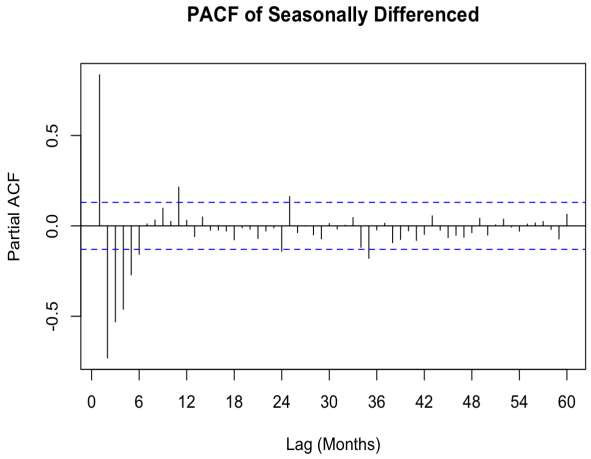

Figure 6: PACF after Seasonal Differencing

2.4 Model Fitting and Diagnostics

Starting with the tentative model SARIMA$(1, 0, 1)(0, 1, 1)[12]$, the residual ACF in the diagnostic plot (Figure 7) shows significant autocorrelation at lag 24. This suggests there is some remaining autocorrelation at this lag, meaning the model has not fully captured a seasonal effect occurring every 24 months. This finding indicates that adjusting the seasonal part of the model might improve its fit.

Additionally, the residual plot reveals two outliers exceeding three standard deviations from zero. To see if handling these outliers could improve the model, I imputed their values with the average for that particular month across all years. Then, I compared the residual normality using the Shapiro-Wilk test. The original model had a W-statistic of 0.984 (p-value = 0.012), and the model after imputation had a W-statistic of 0.978 (p-value = 0.001). The results show that imputation did

not improve the model fit in terms of normality of residuals . Since the residuals of the original model appear centered around zero without any clear patterns or trends, and the dataset itself does not contain extreme outliers, I decided to proceed without imputing these residual outliers.

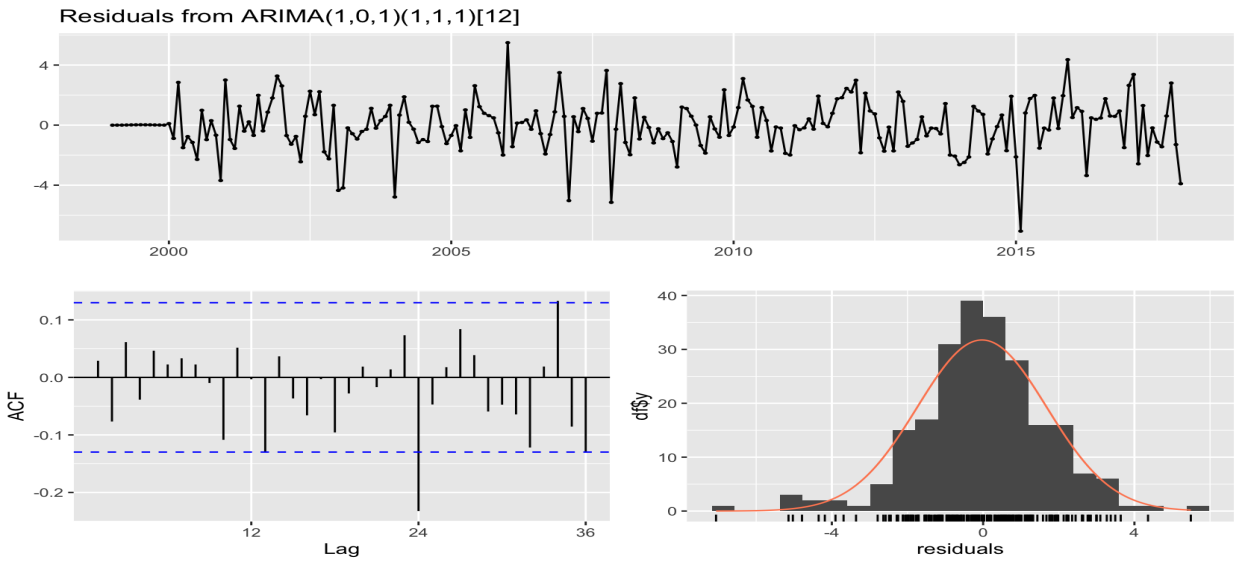

Figure 7: Diagnostic Plot for Model 1

Next, the model was adjusted to SARIMA$(1,0,1)(2,1,1)[12]$ by increasing the seasonal P from 1 to 2. The diagnostic plot (Figure 8) shows that this model successfully removed the seasonal spike at lag 24. However, the residual ACF slightly exceeds the significance bounds (-0.12) at lags 32 and 36. Overall, although the main seasonal effect has been addressed, the residual autocorrelation at lags 32 and 36 indicates some remaining structure that could be further investigated.

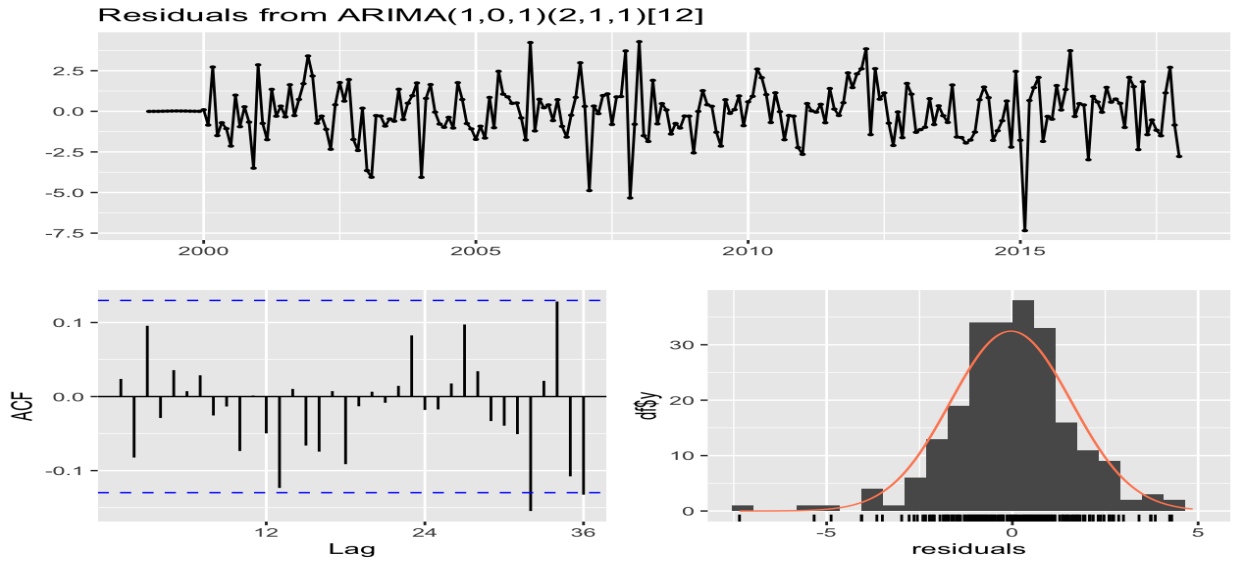

Figure 8: Diagnostic Plot for Model 2

Continuing with the model adjustments, I first overfitted SARIMA$(2,0,1)(2,1,1)[12]$ (model_3). However, this did not result in any noticeable improvement compared to model_2. Then, I fitted SARIMA$(2,0,1)(2,1,2)[12]$ (model_4) by increasing the seasonal Q from 1 to 2. The residual ACF for model_4 only has one lag (lag 32) that slightly exceeds the significance bounds, indicating a minor improvement in model fit.

In practical scenarios, small deviations in the residual ACF are common and usually do not significantly reduce the model's forecasting accuracy. Therefore, I decided to proceed with both model_2 and model_4 for forecasting.

Model	**Order**	**Period**	**Ljung-Box p-value**	**AIC**	**BIC**
Model_1	$(1,0,1) \times (1,1,1)$	12	0.05404	900.6	917.48
Model_2	$(1,0,1) \times (2,1,1)$	12	0.5994	889.01	909.27
Model_3	$(2,0,1) \times (2,1,1)$	12	0.5045	890.72	914.35
Model_4	$(2,0,1) \times (2,1,2)$	12	0.621	887.3	914.31

Table 1: Comparison of Model Fitting Results

3 Results

3.1 Forecast Results

Figures 9 and 10 illustrate the actual temperatures (black line with points) against their forecasted 95% CIs (light blue shading for Model_2, light pink shading for Model_4) over the test period.

Based on these result graphs, both SARIMA models show good performance in capturing Toronto's monthly average temperature pattern for the forecasting period of 2018-2019. The actual temperature values closely follow the forecasts produced by both models, falling mostly within the 95% confidence intervals.

Model_2 (SARIMA$(1,0,1)(2,1,1)[12]$) generally captures the seasonal temperature trends effectively, with slightly broader confidence intervals at the peaks.

Model_4 (SARIMA$(2,0,1)(2,1,2)[12]$) shows slightly narrower confidence intervals and provides slightly more precise predictions, shown by narrower confidence intervals, particularly at the highest summer temperatures.

Overall, both models are reliable for forecasting seasonal temperature variations, with model_4 showing a slight advantage in precision.

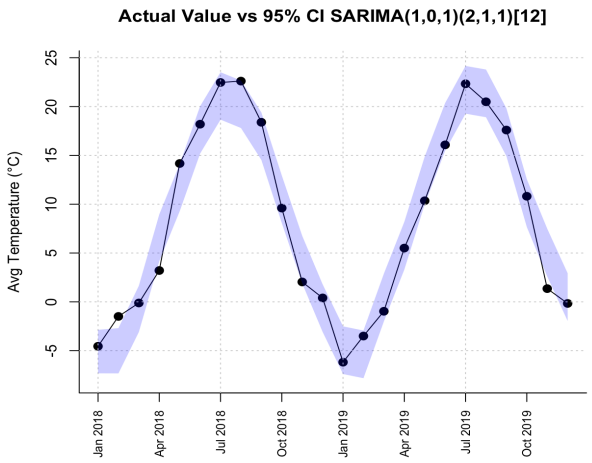

Figure 9: 95% Confidence Interval of Model_2

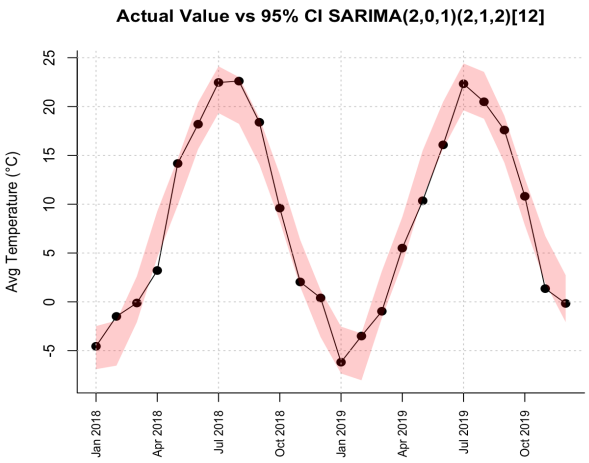

Figure 10: 95% Confidence Interval of Model_4

Figure 11 shows that the forecasted temperatures (in blue) closely follow the historical seasonal pattern. This suggests the selected models effectively captured the regular cycles in Toronto's climate. The smooth continuation from past data to forecasted data supports the accuracy and reliability of the chosen SARIMA models.

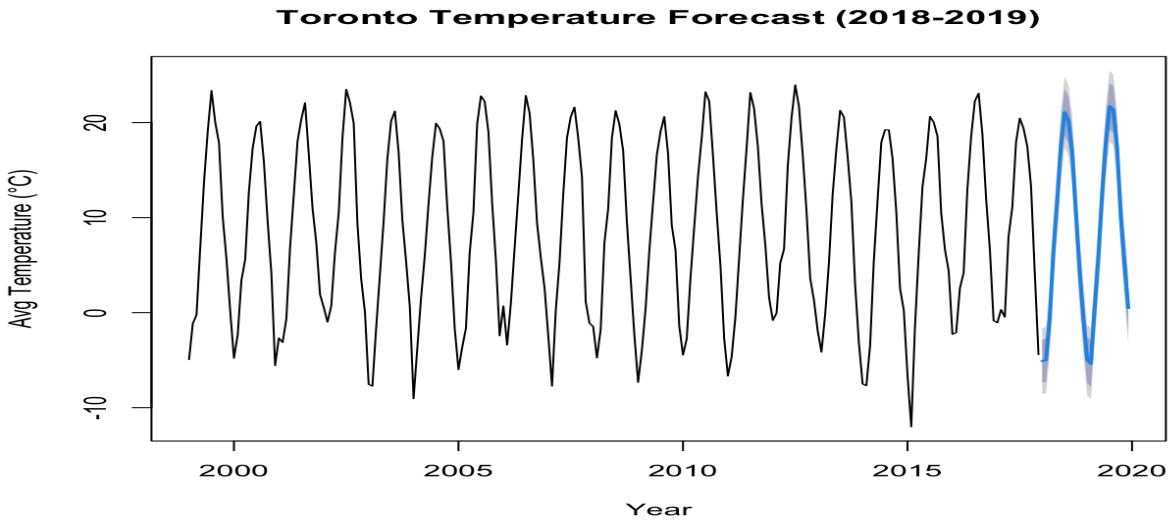

Figure 11: Toronto Monthly Average Temperature Forecast (2018-2019)

3.2 Performance Metrics

When comparing the performance of both models using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE), model_2 has an MAE of 1.495°C. This means, on average, its forecasts differ from actual temperatures by about 1.5°C.

Model_4 performs slightly better with a lower MAE of 1.452°C. Both models have MAE values below 3°C, indicating good prediction accuracy.

The MSE values for the two models are also similar. Model_2 has an MSE of 3.245°C², while model_4 has a slightly lower MSE of 3.031°C². However, this difference (0.214°C²) is quite small, suggesting that both models handle larger errors similarly.

Looking at MAPE, model_2 has a value of 93.22%, while model_4 has a lower value of 84.345%. Both values are relatively high, meaning both models face challenges with relative accuracy. Despite small differences in MAE and MSE, the improvement seen in the MAPE for model_4 is more noticeable, suggesting model_4 provides a better overall fit.

Model	**Order**	**Period**	**MAE**	**MSE**	**MAPE**
Model_2	$(1,0,1) \times (2,1,1)$	12	1.495	3.245	93.22%
Model_4	$(2,0,1) \times (2,1,2)$	12	1.452	3.031	84.345%

Table 2: Performance Comparison

Furthermore, the coefficient significance tests show that model_2 has one non-significant coefficient (sar1). In contrast, all coefficients in model_4 are significant, indicating that model_4 is better specified and more reliable.

The model equation is:

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

$$\begin{aligned}(1-0.8192B+0.0505B^2)\big(1-0.6605B^{12}+0.3154B^{24}\big)(Y_t-Y_{t-12}) \\ = (1+0.5162B)\big(1+1.6294B^{12}-0.7020B^{24}\big)e_t\end{aligned}$$

4 Conclusion

This study developed a SARIMA model to forecast Toronto's monthly average temperatures for 2018–2019, achieving good results in capturing seasonal patterns. However, several limitations remain, including a high MAPE indicating poor relative accuracy, especially during colder months, and the model's inability to capture extreme temperature events effectively, as shown by two residual outliers. Additionally, residuals are not normally distributed, and the relatively short 19-year dataset might miss longer climate cycles. Future work could address these limitations by using longer datasets, incorporating intervention analysis for extreme events, exploring alternative models like exponential smoothing to enhance predictive accuracy and reliability.