International Journal of Data Science and Analysis
2022; 8(3): 82-93
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20220803.12
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Time Series Analysis in Forecasting Monthly Average Rainfall and Temperature (Case Study, Minot ND, USA)

Upul Rupassara[*], Dion Udokop, Favour Ozordi

Department of Mathematics and Computer Science, Minot State University, Minot, the United States

Email address:
upul.rupassara@minotstateu.edu (U. Rupassara)
[*]Corresponding author

To cite this article:
Upul Rupassara, Dion Udokop, Favour Ozordi. Time Series Analysis in Forecasting Monthly Average Rainfall and Temperature (Case Study, Minot ND, USA). *International Journal of Data Science and Analysis*. Vol. 8, No. 3, 2022, pp. 82-93. doi: 10.11648/j.ijdsa.20220803.12

Received: May 12, 2022; **Accepted**: May 25, 2022; **Published**: May 31, 2022

Abstract: This project analyzes the monthly average rainfall and temperature from 2005 January to 2021 December in Minot, ND, USA. Since both rainfall and temperature time series represent seasonal components, Seasonal Auto Regressive Integrated Moving Average (SARIMA) models were used to forecast the average rainfall and temperature. The main objective was to identify the SARIMA models based on Akaike's Information Criteria (AIC). The graphical and diagnostic analysis techniques validated the models having the smallest AIC values. Among the competitive tentative models, the SARIMA (2, 0, 0) (2, 0, 1, 12) and SARIMA (1, 0, 1) (2, 0, 1, 12) were found to be the best time series forecasting models that capture the existing pattern of the rainfall and temperature data, respectively. Nevertheless, these models satisfy the model diagnostics test assumptions on the residuals such as randomness, independency, normality, and heteroscedasticity. Therefore, SARIMA (2, 0, 0) (2, 0, 1, 12) and SARIMA (1, 0, 1) (2, 0, 1, 12) models were used to forecast the mean rainfall and temperature, respectively, from the 2022 January to 2023 December.

Keywords: Rainfall, Temperature, Average, SARIMA, Forecasting, Models

1. Introduction

Time series analysis is a statistical technique that deals with data collected at different times. Usually, data are collected at the adjacent period, and there is a potential for correlation between the observations. The intervals on which data are collected are called time series frequency. Forecasting is obviously a difficult activity, but many advanced forecasting techniques have been developed with new software packages to overcome this challenge. Good forecasts capture the genuine patterns and relationships in the historical data. The primary purpose of weather forecasting is to provide knowledge that people and organizations could utilize to decrease weather-related losses and improve societal advantages, such as life and property protection, public health and safety, and economic prosperity and quality of life [12].

Forecasting should be an integral part of the decision-making activities in various sectors of society. Depending on the specific application, modern organizations require short-term, medium-term, and long-term forecasts. Forecasting average temperature and rainfall are essential for planning and formulating agricultural strategies. It helps farmers to manage risks, especially; short-range forecasts are mainly used in agriculture, water management, and many other purposes. Crop growth, or crop yield, requires appropriate amounts of moisture, light, and temperature. Detailed and accurate forecast weather information can help farmers better understand and track the growth stages to make decisions. Having access to this information can guide farmers in making significant and potentially costly decisions, such as when and how much to irrigate. These facts emphasize that accurate forecasting models are essential to a critical geographic region like Minot. For most of the months of the year, the temperature drops too low, and outdoor farming is not possible during this time. Therefore, accurate weather forecasting models help farmers get maximum benefit during the cultivation period in the summer months.

Since Minot gets snow for approximately five months, the snow water equivalent data is used as rainfall data because of the unavailability of the rainfall data in the major climate centers. Minot poses a challenge when collecting rainfall data for these months as snow is the main form of precipitation during that period. From northern to southeastern North Dakota, mean annual precipitation varies from 14 to 22 inches. Around 75 percent of the yearly precipitation occurs within the crop-growing season, which runs from April to September, and 50 to 60 percent between April and July. The coldest months, November to March, get only approximately 0.50 inches of precipitation each month, largely in the form of snow [20].

This research aims to identify the pattern of the historical data for average monthly rainfall and temperature and to find the best SARIMA model that can be used to forecast the temperature and rainfall for the geographic region of Minot, North Dakota. Observations from 2005 January to 2019 December were used as the training data, and 2020 January to 2021 December were used as the testing data to validate the models. The data collected from the Minot Airforce base and Minot North Hill reading locations were used, and these data were retrieved from the National Operational Hydrological Remote Sensing Center website [11]. Since Minot is experiencing snow from November to March, the rainfall data is not available during this period. But instead, snow water equivalent data was used to replace the missing rainfall data. In some rare situations, when snow water equivalent data is not available, the rainfall height is approximated by the snow height adopting the measurement that ten inches of snow are equivalent to one inch of rain. A simple and cost-effective procedure for estimating solid and liquid precipitation can be found in [8].

The graphical analysis demonstrated that both series have seasonal components, and hence, the SARIMA model was chosen as the most appropriate model for analyzing the series. The Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) were used to determine the order of the Auto Regressive Moving Average (ARIMA) model. The Akaike's Information Criteria (AIC) was used as the primary model selection criterion, and residual analysis techniques were used to validate the selected models. Among the competitive tentative models, the models with the smallest Root Mean Squared Error (RMSE) and largest R-squared value were further discussed as those criteria used for different purposes in our analysis. Models with the lowest AIC work best for forecasting as they fit well with the testing data compared to other models. The models with the highest R-squared and lowest RMSE values determine the goodness of fit against the training data. Since our primary purpose is forecasting, models with the smallest AIC were selected to predict mean rainfall and temperature from January 2022 to December 2023.

Several studies have been performed to analyze the climate data using different statistical models in the literature. Many researchers who work on forecasting problems use ARIMA and SARIMA models to do the initial analysis and identify the series's pattern. Also, in addition to Moving Average (MA) smoothing techniques, exponential smoothing techniques can transform the non-stationary series into a stationary series. A comprehensive description of the exponential smoothing techniques was given [9]. However, there should be an interrelation between temperature and rainfall. Because of that, it is difficult to do an accurate analysis of the joint distribution of the rainfall and temperature. The possible interdependence of temperature and precipitation was studied in [14]. A similar case study can be found in [1, 17, 18] in the recent literature.

2. Different Types of Covariance and Correlation Functions

2.1. The Autocovariance Function

Covariance is a measure of linear dependence between two random variables. The Autocovariance function is defined as the covariance between two observations of a series, x_s and x_t for all s and t.

$$\operatorname{cov}(x_s, x_t) = \mathrm{E}[(x_s - \mu_s)(x_t - \mu_t)] \tag{1}$$

where μ_s and μ_t are the mean of each time series respectively. Autocovariance measures the linear dependence between two points on the same time series observed at different times.

2.2. Cross-Covariance Function

Cross-covariance function is used when the time series x_t is used to measure the predictability of another time series y_t assuming both have finite variances.

$$\operatorname{cov}(x_s, y_t) = \mathrm{E}[(x_s - \mu_{xs})(y - \mu_{yt})] \tag{2}$$

whare, μ_{xs} and μ_{yt} are the mean of the respective time series.

2.3. Autocorrelation Function (ACF)

Autocorrelation function measures the linear predictability of the time series at a time t, say x_t, using the value x_s at time s. The ACF is defined as

$$\rho(s,t) = \frac{\operatorname{cov}(x_s, x_t)}{\operatorname{cov}(x_s, x_s)\operatorname{cov}(x_t, x_t)} \tag{3}$$

ACF is used to estimate how the current value (x_t) of the time series depends on past values, directly and indirectly, depending on the time lag. For example, if time lag (l) is chosen to be 3 then ACF uses all the past values x_{t-1}, x_{t-2}, x_{t-3} and use their linear dependencies to calculate the correlation between current and past time series values. Significant spikes of the ACF plot can be used to determine the order of the (Moving Average) MA model.

Similarly, the Cross-Correlation Function (CCF) is given by

$$\rho_{xy}(s,t) = \frac{\operatorname{cov}(x_s, y_t)}{\operatorname{cov}(x_s, x_s)\operatorname{cov}(y_t, y_t)} \tag{4}$$

2.4. Partial Autocorrelation Function (PACF)

The PACF provides more information about the order of the linear dependency that MA models may not be captured.

ACF provides a considerable amount of information about the order of the dependency when the process is a moving average process. At the same time, PACF removes the intermediate linear dependencies of the autocorrelation for different lag values. It can be used to determine the number of lags that need to explain the linear dependency of the time series. Significant spikes of the PACF plot can be used to determine the order of the AR model.

3. Model Selection Criterion and Tests for Stationarity

3.1. Model Selection Criterion

The Akaike's Information Criterion (AIC), Corrected Akaike's Information Criterion (AICC), and Schwarz Bayesian Information Criterion (SBIC) is used to compare the relative goodness of the fit for generalized linear regression modeling.
AIC is given by the formula,

$$\mathrm{AIC} = 2k - \ln L \tag{5}$$

where k is the number of parameters in the regression model and L is the log-likelihood function. The AICC (for small samples) is given by

$$\mathrm{AICC} = \frac{2kn}{n-k-1} - 2\ln L \tag{6}$$

SBIC is computed as

$$S\mathrm{BIC} = k\ln n - 2\ln L \tag{7}$$

For each criterion, the model with the smallest value gives the best-fitted model. AIC and SBIC explain how well the model will fit the new data, and hence lower values improve the validity of the predictions. More details of the derivation and recent developments of these criteria can be found in [2, 3, 10, 15, 16].

3.2. Auto Regressive Integrated Moving Average Model

One of the key applications in the time series analysis is that the current value depends on past observations. This dependence is one of the major advantages of using time series models for forecasting than the classical regression models. The property that the current value depends on its past value described by the Auto-Regressive Moving Average models. The correlation that may be generated through the lagged relations of the variable leads to propose the Autoregressive (AR) and Autoregressive Moving Average (ARMA) models. Adding nonstationary models leads to an Auto-Regressive Integrated Moving Average model (ARIMA). A significant contribution to the development of ARIMA models was made by Box, G. E., and David A. Pierce [5]. ARIMA model is a combination of the following major components.

I. AR(p) denotes an autoregressive part of the ARIMA model. Autoregression is a regression where the target variable depends on its time-lagged values. It is very natural to predict future values based on past or current values. Therefore, the order of the AR model can be used to determine how many lags are significantly contributed to defining the linear dependency of the time series. An autoregressive model of order p is a model of the form.

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + w_t, \tag{8}$$

where w_t is Gaussian white noise, and $\alpha_1, \alpha_2, \dots, \alpha_p$ are constants with $\alpha_p \neq 0$.

AR(p) models can be written more consciously using the backshift operator B.

$$\phi(B)x_t = w_t \tag{9}$$

where, $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$,

$$B^k x_t = x_{t-k}. \tag{10}$$

see ([19], p. 56, 57) for more details.

II. I (d) define the degree of differencing involved. Differencing is a method of removing non-stationarity by calculating the change between each observation. In many situations, time series can be thought of as being composed of two components, stationary, and nonstationary trend components. Differencing such a process eliminates the nonstationary component and this process will lead to a stationary process. If the trend components $\gamma_0 + \gamma_1 t$ and stationary components y_t composed to give the time series x_t, where

$$x_t = \gamma_0 + \gamma_1 t + y_t, \tag{11}$$

γ_0 and γ_1 are constants. then the first differencing $(d = 1)$ yields,

$$x_t - x_{t-1} = \gamma_1 + y_t - y_{t-1} \tag{12}$$

The degree of the differencing determines the value of d.

III. MA(q) indicates the order of the moving average part of the model, which is given as a moving average of the error series. It can be described as a regression against the past error values of the series.

The moving average model of order (q) is defined to be

$$x_t = w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2} + \cdots + \beta_q w_{t-q}, \tag{13}$$

where w_t is Gaussian white noise and $\beta_1, \beta_2, \dots, \beta_q$ $(\beta_q \neq 0)$ are parameters. This also can be represented more concisely using the backshift operator B.

$$x_t = \beta(B)w_t, \tag{14}$$

where,

$$\beta(B) = 1 + \beta_1 B + \beta_2 B^2 + \cdots + \beta_q B^q \tag{15}$$

The stationary property of the moving average component does not depend on the parameters $\beta_1, \beta_2, \dots, \beta_q$. ACF plot can be used to determine the order of the MA model.

3.3. Seasonal AutoRegression Integrated Moving Average Model

Seasonality is a prevalent feature in most time series that appear in weather forecasting. Suppose any predictable fluctuation or pattern is available. In that case, the appropriate time series model should be used that captures the seasonal patterns because there may be a significant contribution from the different seasonal factors. The time-series data consists of two parts: the non-seasonal and the seasonal models. The non-seasonal part incorporates long-term changes, while the seasonal part looks at seasonal cycles. Therefore, ARIMA(p, d, q)(P, D, Q, S) is the most appropriate model to analyze time series with seasonal patterns. P, D, and Q represent the parameters for the seasonal model, while S represents the period of repeating seasonal patterns. The SARIMAX is the updated version of the SARIMA model that captures the seasonal changes and Xogenous factors.

3.4. Probabilistic Behavior and Regularity

The regularity of the time series was introduced by the concept called stationarity. Strict stationarity is tangible property, and in general, it isn't easy to find a time series that satisfy this condition. The changes of the stationary time series do not change over time. Therefore, stationary series are more accessible to analyze than non-stationary series.

If the time series is strictly stationary, its probabilistic behavior is identical to every collection of values at different times.

That is for a time series x_t,

$$P\left(x_{t_1} \le c_1, \dots, x_{t_k} \le c_k\right) = P(x_{t_1+l} \le c_1, \dots, x_{t_k+l} \le c_k) \quad (16)$$

for all $k = 1,2,\dots$, all the time shifts $l = 0, \pm 1, \pm 2, \dots$, all-time points $t_1, t_2, \dots, t_k$ and all numbers $c_1, \dots, c_k$. Since this version is too strong for most applications, a milder version called weak stationary that imposes conditions only on the first two moments is used to fulfill this requirement.

In our discussion, the term stationary means weak stationary, and we will mention it if strong stationery is required.

3.5. Dickey-Fuller Unit Root Test

One of the robust formal tests that can be used to determine whether the time series is the AR(1) model. Let the observations $Y_1, Y_2, \dots, Y_n$ generated by the first-order linear difference equation.

$$Y_t = aY_{t-1} + \epsilon_t,\ a \text{ is a real number}, \epsilon_t \sim N(0, \sigma^2) \quad (17)$$

The time series converges to a stationary series for $|a| < 1$ when $t \to \infty$, and it is not stationary if $|a| = 1$. If $|a| > 1$ the time series is not stationary as the variance grows exponentially [7]. The augmented Dickey-Fuller (ADF) Test facilitates checking the stationarity of higher-order AR models.

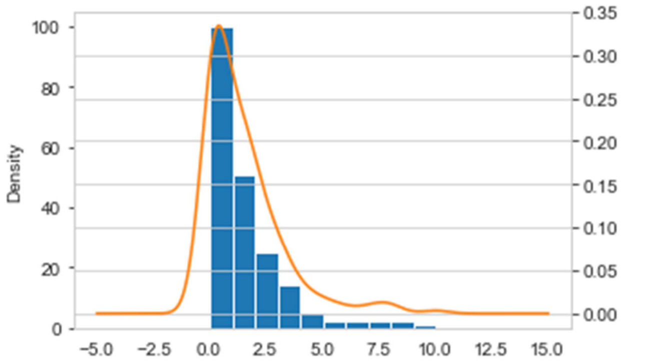

Figure 1. *KDE plot and histogram for average rainfall.*

4. Computations and Results

The following computations, analyses, and discussions are based on the data collected from Minot, North Dakota, USA, from January 2005 to December 2021.

4.1. Descriptive Statistics of the Rainfall and Temperature Data

In any statistical test, it is a common practice to compute the summary of the statistics that provide the researchers with the complete overall descriptive coefficients of the information in the data set. Usually, mean, standard deviation, and five-number summary are used to identify the nature of central tendency and spread of the data. Table 1 shows the descriptive statistics for the monthly average rainfall and temperature data from 2005 January to 2021 December for 204 observations.

Table 1. *Descriptive statistics for monthly average rainfall and temperature.*

Variable	Count	Min	Max	Mean	Std
Rainfall	204	0	10.05	1.48	1.72
Temperature	204	-6.15	74.39	41.14	21.42

Minot experiences heavy snow during the winter months. In January and February, the maximum temperature is recorded as negative values. According to the data, the minimum temperature is -6.15 Fahrenheit, while the maximum is 74.39 Fahrenheit. Figure 1 shows the Kernel Density Estimation (KDE) and histograms for the average rainfall, and figure 2 demonstrates those plots for the monthly average temperature. The temperature distribution is approximately symmetric with a mean of 41.14 Fahrenheit, while the average rainfall variation is skewed right with a mean of 1.48 inches. The minimum recorded rainfall for this period is 0 inches, and the maximum is 10.05 inches.

4.2. Time Series Decomposition and Model Identification

Monthly average rainfall and temperature data from 2005-January to 2019-December are used as training data, and the observations from 2020-January to 2021-December are used as testing data. Both rainfall and temperature distribution are stationary, and Dickey-Fuller test values are 0.04 and 0.05. One lag differencing further decreases the Dickey-Fuller test values for both series, and d=0 and d=1 are the most appropriate values for the integrated order of the ARIMA model. Although there is no clear trend, both series demonstrate a seasonal pattern.

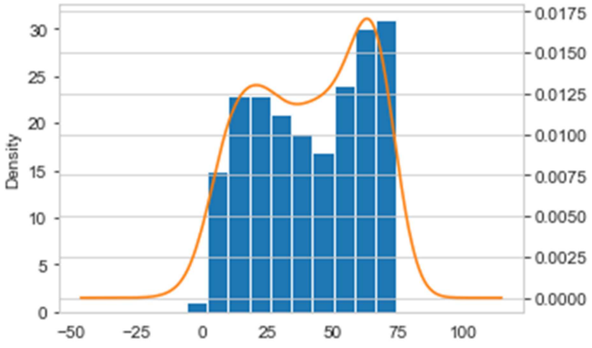

Figure 2. *KDE plot and histogram for average temperature.*

4.2.1. Time Series Decomposition

Decomposition is a powerful statistical technique to deconstruct the time series into seasonal (S_t), cyclic (C_t), trend (T_r) and noise (ϵ_t) components. Considering the variation around the trend, this can be done in two different ways, namely additive and multiplicative decompositions. This can be given by the equations:

i. Additive model

$$x_t = S_t + C_t + T_r + \epsilon_t \tag{18}$$

ii. Multiplicative model

$$x_t = S_t \times C_t \times T_r \times \epsilon_t \tag{19}$$

In general, the amplitude of the multiplicative model changes more drastically than the amplitude of the additive model. Figures 3 and 4 show the graphical representation of the individual components of the decomposed time series, namely, trend and seasonality. In both average temperature, and rainfall data, a strong 12-month seasonal pattern can be observed, and the order of the seasonal parameter obviously should be 12. A comprehensive overview of techniques and methods in time series modeling and decomposition can be found in [4, 13].

Table 2. *Possible values for p, d, and q.*

Model	Rainfall	Temperature
AR (p)	0, 1	0, 1, 2
MA (q)	0, 1, 2	0, 1, 2
I (d)	0, 1	0, 1

4.2.2. Identification of the Order of the AR and MA Models

ACF and PACF provide strong support in determining the order of the AR and MA components of the time series. Figures 5 and 6 represent the ACF and PACF plots for average rainfall and temperature data. According to the plots, the possible combinations of the orders for the AR and MA components, along with integrated orders, are given in Table 2.

4.3. SARIMA Model Selection

An iterative process is used to calculate the AIC values for the different combinations of the ARIMA parameters $p, d,$ and q and the SARIMA parameters P, D, and Q. The table shows ten different SARIMA models with their AIC along with log-likelihood values, R- squared values, and Root Mean Square Error (RMSE). The AIC value is used as the primary model selection criterion. SARIMA model with the lowest AIC is chosen as the best among the other competitive models.

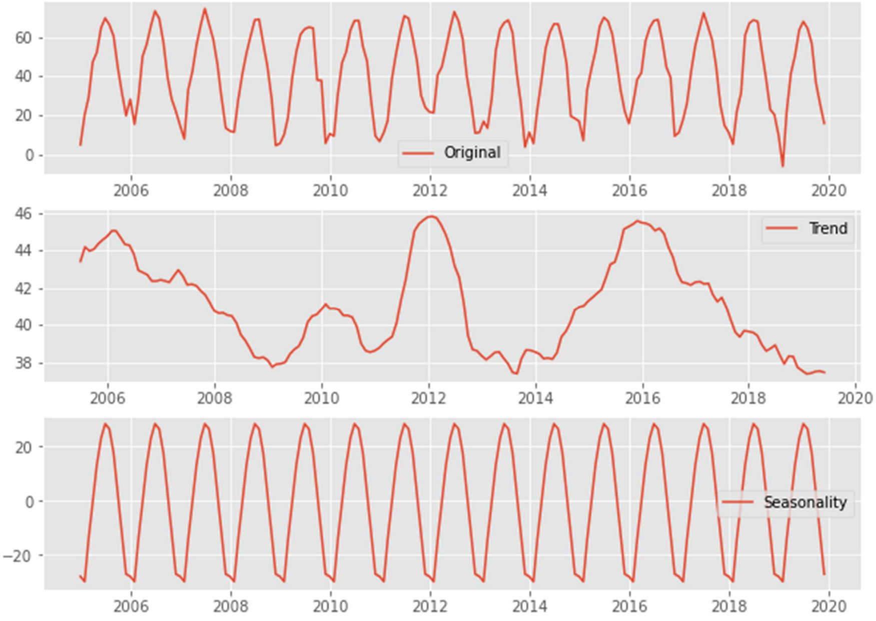

Figure 3. *Additive seasonal decomposition of monthly average rainfall.*

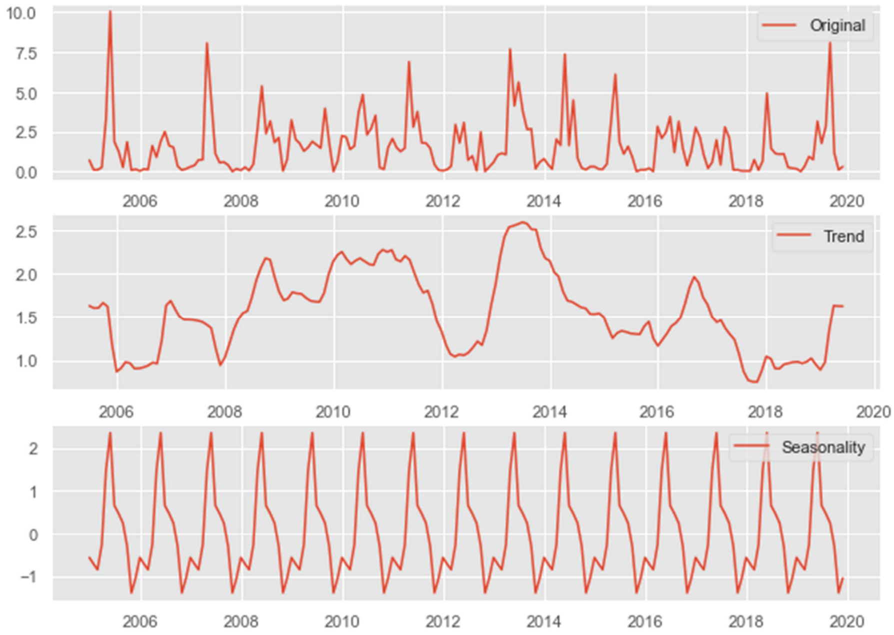

Figure 4. *Additive seasonal decomposition of monthly average temperature.*

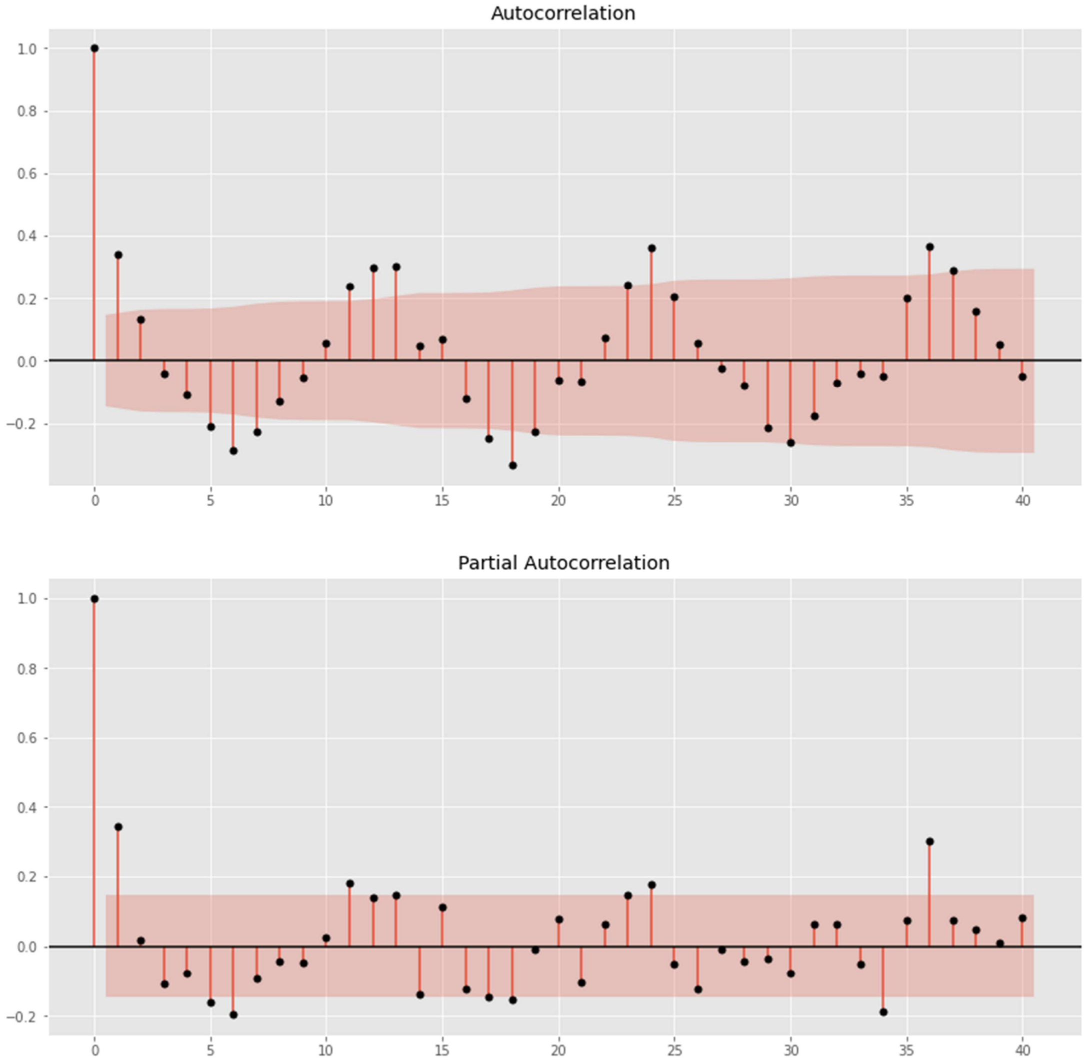

Figure 5. *ACF and PACF plots for monthly average rainfall.*

4.4. Model Results for Rainfall Data

The AIC, Log-Likelihood value, R-squared values, and RMSE for the top SARIMA models for average rainfall data are shown in Table 3. Among the top ten competitive models, the SARIMA (2, 0, 0) (2, 0, 1, 12) has the smallest AIC, while the model (2, 0, 0) (3, 0, 1, 12) qualified for the smallest RMSE and largest R-squared value.

Table 3. *SARIMA (2, 0, 0) (2, 0, 1, 12) model results for monthly average rainfall data.*

SARIMA Model	AIC	Log Likelihood	R - squared	RMSE
(2, 0, 0)(2, 0, 1, 12)	664.332	-326.166	0.0848	0.9679
(1, 0, 0)(2, 0, 1, 12)	666.474	-328.237	0.0604	0.9807
(2, 0, 0)(1, 0, 2, 12)	664.407	-326.203	0.1069	0.9561
(1, 0, 0)(1, 0, 2, 12)	666.651	-328.325	0.0807	0.9700
(1, 0, 1)(2, 0, 1, 12)	665.324	-326.662	0.0677	0.9769
(3, 0, 0)(2, 0, 1, 12)	666.221	-326.111	0.0792	0.9708
(2, 0, 1)(2, 0, 1, 12)	666.240	-326.120	0.0799	0.9705
(1, 0, 0)(2, 0, 2, 12)	668.470	-328.235	0.0660	0.9778
(2, 0, 0)(3, 0, 1, 12)	666.267	-326.134	0.1169	0.9507
(2, 0, 0)(2, 0, 2, 12)	666.308	-326.154	0.0970	0.9614

4.5. Model Results for Temperature Data

The AIC, Log-Likelihood value, R-squared value, and RMSE for the SARIMA models with the lowest AIC value for the average temperature data are shown in Table 4. Among the top ten competitive models, the SARIMA (1, 0, 1) (2, 0, 1, 12) has the smallest AIC, while the model (0, 0, 1) (1, 0, 2, 12) qualified for the smallest RMSE and largest R-squared value.

Table 4. *SARIMA (1, 0, 1) (2, 0, 1, 12) model results for monthly average temperature data.*

SARIMA Model	AIC	Log Likelihood	R - squared	RMSE
(1, 0, 1)(1, 0, 2, 12)	1155.658	-571.829	0.9441	4.8392
(1, 0, 1)(1, 0, 1, 12)	1153.879	-571.940	0.9448	4.8098
(1, 0, 1)(2, 0, 1, 12)	1155.424	-571.712	0.9449	4.8021
(0, 0, 1)(1, 0, 1, 12)	1176.319	-584.159	0.9491	4.6170
(0, 0, 0)(1, 0, 1, 12)	1188.346	-591.173	0.9501	4.5727
(2, 0, 0)(1, 0, 2, 12)	1164.430	-576.215	0.9498	4.5878
(0, 0, 1)(1, 0, 2, 12)	1177.267	-583.634	0.9502	4.5669
(0, 0, 0)(1, 0, 2, 12)	1192.269	-592.134	0.9472	4.7011
(0, 0, 2)(1, 0, 2, 12)	1175.792	-581.896	0.9499	4.5810
(0, 0, 1)(2, 0, 0, 12)	1227.418	-609.709	0.8746	7.2484

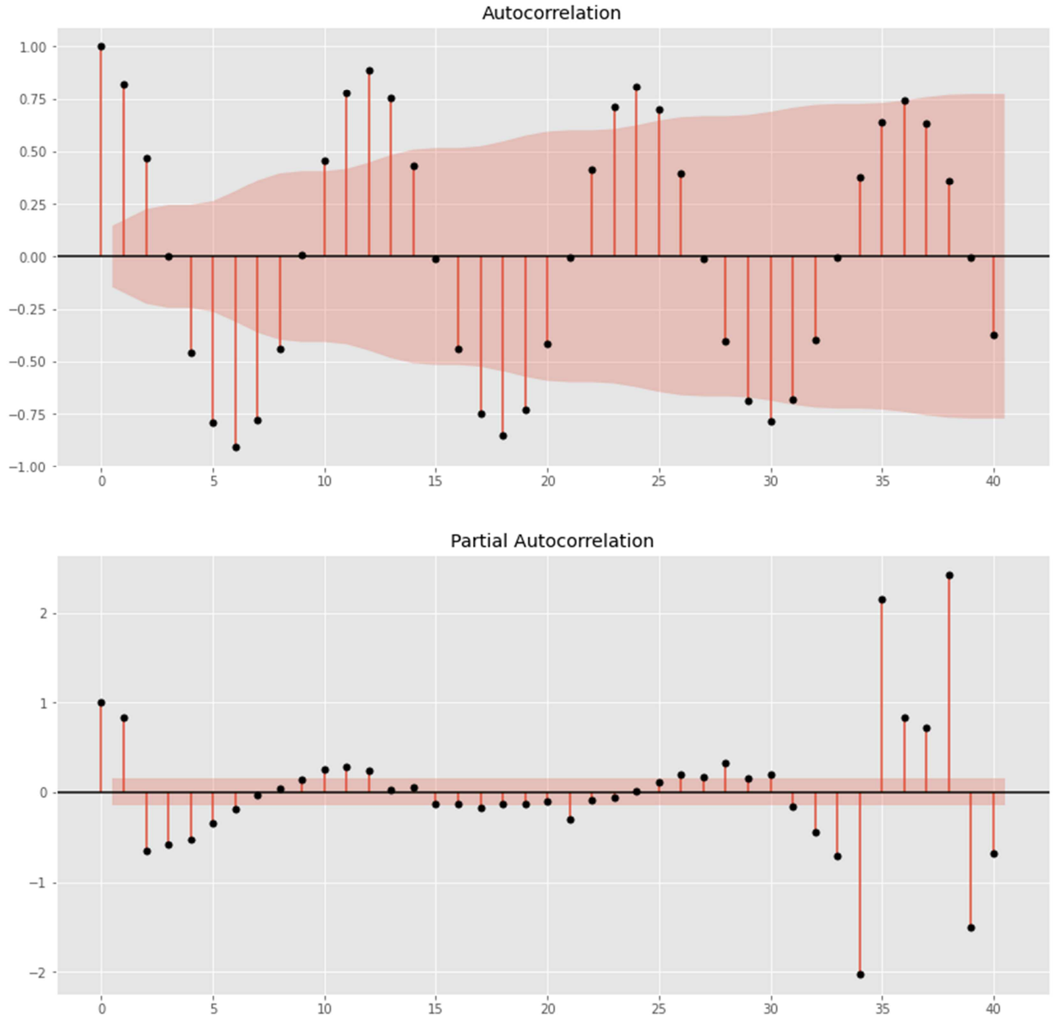

Figure 6. *ACF and PACF plots for monthly average temperature.*

5. Model Validation

Model validation is the process of confirming the introduced model achieves the intended purpose, and different types of diagnostics tests are available to justify the validity of the fitted model. Without validating the model, it is not right to rely on the predictions. A widely used diagnostic technique is residual analysis. Residual is the difference between actual observations and fitted values. Residuals of a well-fitted model are uncorrelated and follow a Gaussian distribution with a mean zero and satisfy the homoscedasticity property [9].

5.1. Parameter Estimation for Rainfall Data

The coefficients of the (Auto Regression) AR, (Seasonal Auto Regression) SAR, (Seasonal Moving Average) SMA components, along with their standard error values and p-values for the SARIMA (2, 0, 0) (2, 0, 1, 12) are given in Table 5.

Even though the model fitted well with the observed data, predicting rainfall using the model is challenging because Minot experiences snow for more than four months over the year.

Table 5. *Parameter estimation (rainfall).*

Component	Coefficient	Std. Error	p-value
AR(1)	0.1932	0.084	0.022
AR(2)	0.1522	0.086	0.075
SAR(1)	0.8111	0.088	0.000
SAR(2)	0.1785	0.083	0.033
SMA(1)	-0.8368	0.088	0.000

5.2. Parameter Estimation for Temperature Data

The coefficients of the (Auto Regression) AR, (Seasonal Auto Regression) SAR, (Seasonal Moving Average) SMA components, along with their standard error values and p-values for the SARIMA (1, 0, 1) (2, 0, 1, 12) are given in Table 6.

Table 6. *Parameter estimation (temperature).*

Component	Coefficient	Std. Error	p-value
AR(1)	0.9127	0.051	0.000
MA(1)	-0.6709	0.075	0.000
SAR(1)	0.9907	0.016	0.000
SAR(2)	0.0091	0.016	0.572
SMA(1)	-0.9095	0.116	0.000

6. Model Diagnostics Tests

Various diagnostic tests are available to determine the fitted model's validity. Each diagnostic tests focus on a different dependence structure. Figures 7 and 8 show the residual plot, normal Q-Q plot, Histogram plus KDE plot, and correlogram for the average rainfall and temperature data.

6.1. Residual Plot

The residual plot visually demonstrates how the fitted model captures the actual data. A good forecasting method will yield uncorrelated residuals. If there are correlations between residuals, there is information left in the residuals that should be used in computing forecasts. The residuals have zero mean. The estimates are biased if the residuals have a mean other than zero [9]. If the appropriate model is chosen, there will be zero autocorrelation in error [5].

6.2. Histogram and Estimated Density

Histograms of the residual of the fitted models for the rainfall and temperature have approximately bell shape, and this can be further confirmed by comparing the standard normal curve and Kernel Density Estimator (KDE).

6.3. Normal Q-Q Plot

The normal Q-Q plot is a visual representation that can compare the normality of theoretical and observed models. Further, it can be used to confirm the validity of the normality assumptions. Normal Q-Q plots compare the quantile residuals against the theoretical quantile. If most of the data points lie on a straight line, then residuals are normally distributed.

6.4. Correlogram Plot

Correlogram plots for the residual show the significance of the autocorrelation of the residuals at each lag value.

7. Model Diagnostics Tests Results

According to the rainfall diagnostics, by the Q-Q plot in Figure 7, it is hard to confirm that the residuals are perfectly normally distributed because a significant number of the observations deviate from the straight line. The KDE plot is skinnier than the standard normal density curve. The correlogram shows no considerable correlation between the residuals at each lag value.

Figure 8 shows the diagnostic plots for average temperature data. Unlike rainfall data, the temperature data residuals closely follow the standard normal distribution, and the Normal Q-Q plot reflects this. Moreover, the KDE density curve for the temperature data approximately follows the standard normal density curve, and hence the residuals are approximately normally distributed. The correlogram plot also suggests no significantly visualized correlation between residuals for each lag value. However, according to the overall diagnostic analysis plots, the fitted model for the average temperature is more accurate compared to the rainfall model.

7.1. Ljung-Box Test

Ljung-Box test can be used to check if autocorrelation exists in the time series. The Ljung-Box test statistics Q is given by,

$$Q = n(n+2)\sum_{k=1}^{h}\frac{r_k^2}{(n-k)} \tag{20}$$

where n is the number of observations, r_k is the autocorrelation for lag k, and h is the number of lags tested. This test is a hypothesis test that the null hypothesis is that the residuals are independently distributed, and the alternative hypothesis is that the residuals are not independently distributed and exhibit a serial correlation.

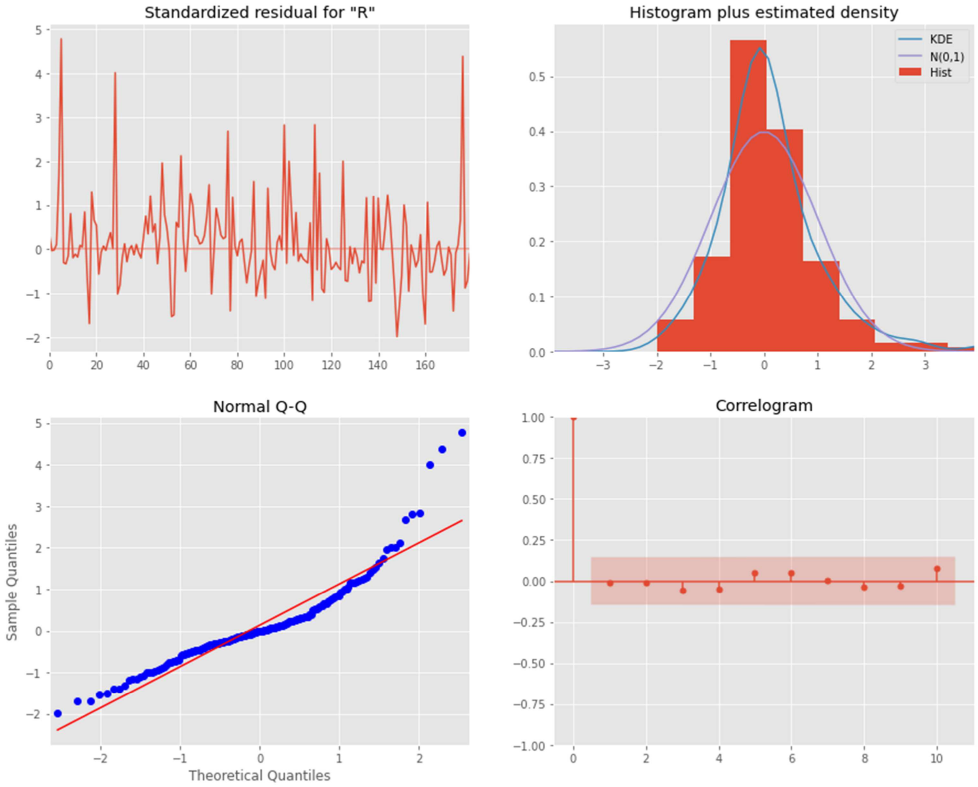

Figure 7. *Model diagnostic plots for the SARIMA (2, 0, 0) (2, 0, 1, 12), monthly average rainfall.*

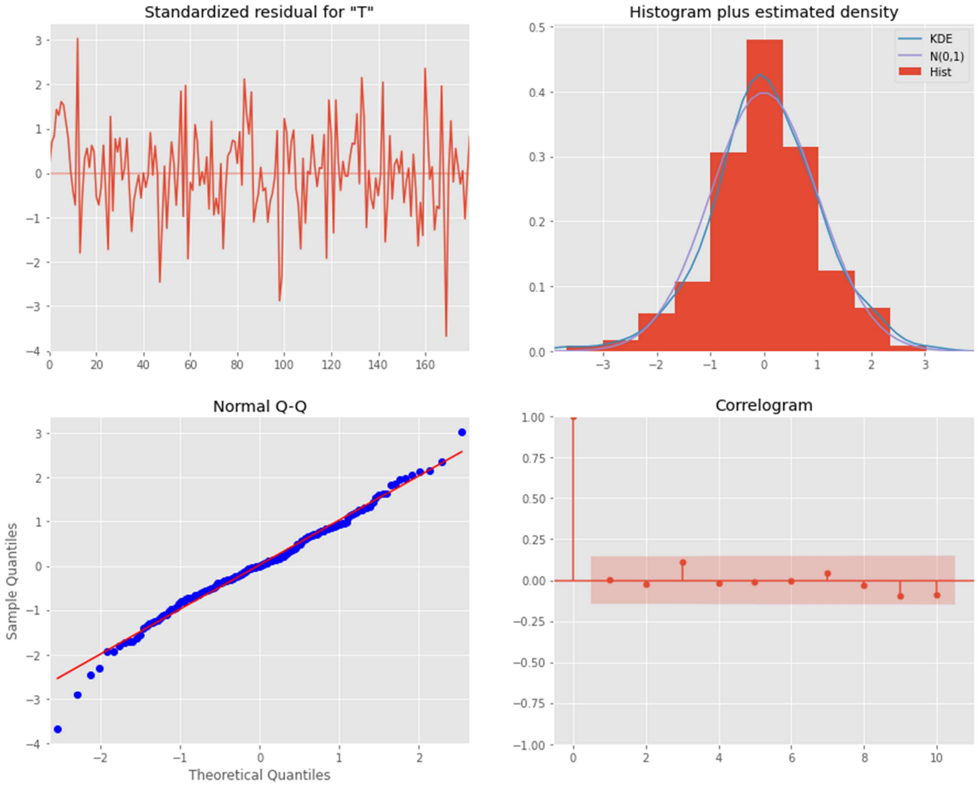

Figure 8. *Model diagnostic plots for SARIMA (1, 0, 1) (2, 0, 1, 12), monthly average temperature.*

Table 7. *Ljung-Box test results.*

Test values	Average Rainfall	Average Temperature
Ljung-Box (L1) (Q)	0.03	0.01
Prob (Q)	0.86	0.94

The Ljung-Box test summary is given in Table 7. The Ljung-Box p-value for rainfall data is 0.86, while it is 0.94 for temperature data. So in both cases, the null hypothesis fails to reject at the significance level $\alpha = 0.05$. So the test is not significant and sufficient evidence does not exist to conclude that residuals are not uncorrelated.

7.2. Heteroskedasticity

Heteroskedasticity (or heteroscedasticity) can be used as a reference to check the uniformity of the variance of the residual over time. Table 8 shows the heteroscedasticity and corresponding p-values. In both situations, we fail to reject the null hypothesis and do not have enough evidence to suggest that residuals do not have equal variance at the significance level of $\alpha = 0.05$.

Table 8. *Test results for heteroskedasticity.*

Test Values	Average Rainfall	Average Temperature
Heteroskedasticity (H)	0.79	1.03
Prob (H) (two sided)	0.37	0.90

If the residual has a constant variance, it is known as homoscedasticity. More details on efficient tests for normality and homoscedasticity can be found in [6].

7.3. Jarque-Bera Test

Jarque-Bera test is another diagnostic test that can be used as a test for normality. This test is based on the sample skewness and sample kurtosis. It is a goodness of fit test that can be used to determine whether the skewness and kurtosis follow the normal distribution. Jarque-Bera (JB) test statistic Q is given by

$$Q = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right) \tag{21}$$

where $n, k,$ and s represent the sample size, kurtosis, and skewness respectively.

The null hypothesis for the tests is data are normally distributed, while the alternative hypothesis is data are not normally distributed. Table 9 shows the JB test statistics and p-values for average rainfall and temperature data.

Table 9. *Test results for the Jarque-Bera test.*

Test values	Average Rainfall	Average Temperature
Jarque-Bera (JB)	252.15	8.10
Prob (Q)	0.00	0.02

8. Prediction with Fitted Models

The average monthly rainfall and temperature data from 2005-January to 2019-December were used to fit the data with SARIMA models, and observations from 2020-January to 2021-December were used as testing data to validate the model. Figure 9 shows the graphical representation of the observed data and predictions for the SARIMA (2, 0, 0) (2, 0, 1, 12) model for average rainfall data with a 95% confidence band. Figure 10 demonstrates the observed and predicted values for SARIMA (1, 0, 1) (2, 0, 1, 12) for the average temperature data with a 95% confidence band. Moreover, each graph shows the average rainfall and temperature forecasting pattern from January 2022 to December 2023 for the average temperature and rainfall.

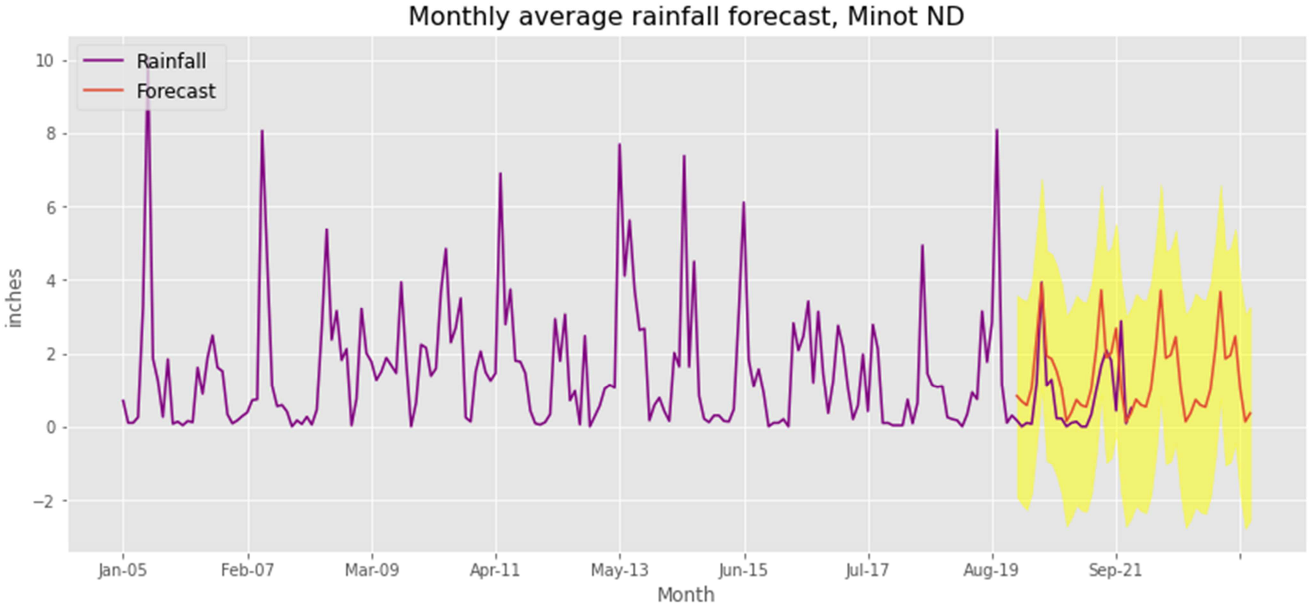

Figure 9. *Observed and SARIMA (2, 0, 0) (2, 0, 1, 12) predictions.*

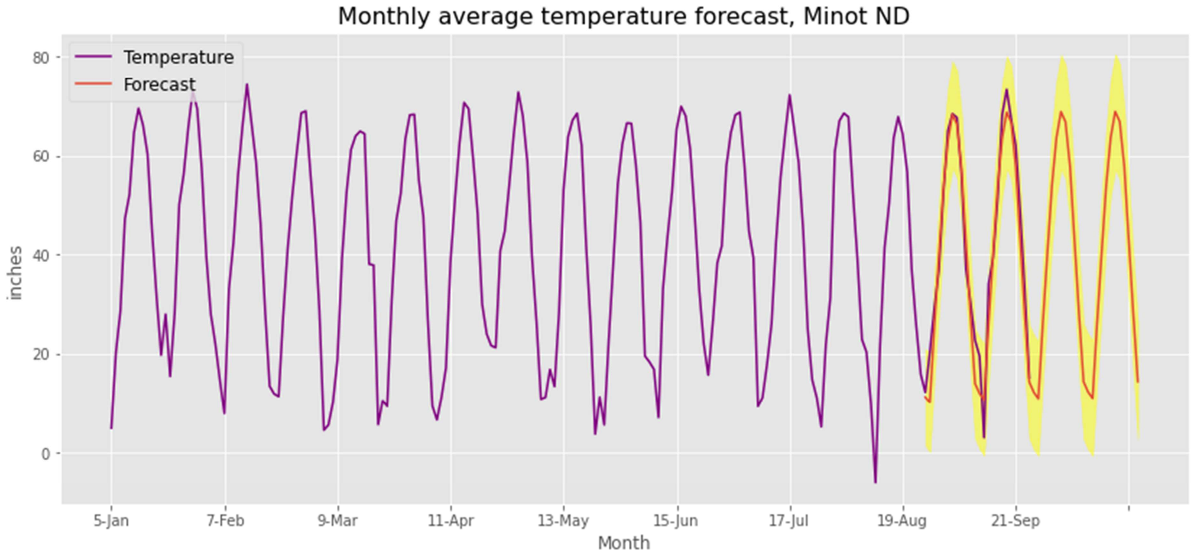

Figure 10. *Observed and SARIMA (1, 0, 1) (2, 0, 1, 12) predictions.*

9. Conclusion and Recommendation

9.1 Conclusion

The best model is selected based on the AIC criterion and the SARIMA (2, 0, 0) (2, 0, 1, 12) has the smallest AIC for the rainfall data while the SARIMA (1, 0, 1) (2, 0, 1, 12) has the smallest AIC for the temperature data. Although the SARIMA (2, 0, 0) (2, 0, 1, 12) has the lowest AIC value the SARIMA model (2, 0, 0) (3, 0, 1, 12) has the smallest RMSE and largest R-squared value compared to the previous model. But the difference between those values is very small (Table 3). For the average temperature, SARIMA (0, 0, 1) (1, 0, 2, 12) has the lowest RMSE and largest R-squared value among the other comitative models (Table 4). The model we selected as the largest R-squared value for the rainfall yielded an R-squared value of 11.69%, and this is obviously not a good value for a forecasting model. The reasonable justification for the low R-squared value of rainfall data is the irregularity of rainfall patterns due to the heavy snowfall during the winter and the practical difficulties of measuring the snow water equivalent amount to replace the missing rainfall data. The unavailability of accurate rainfall data is one of the major challenges in fitting a good time series model for the average rainfall. For the temperature data, the SARIMA (0, 0, 1) (1, 0, 2, 12) has the largest R-squared value of 95.02% among the other competitive models. Although the temperature is very low for some months there are no technical difficulties to measure the temperature compared to the rainfall measurements in the winter. Therefore the fitted SARIMA model captures most of the temperature variations including cyclic patterns throughout the period considered in the analysis.

9.2. Recommendation

The SARIMA model for the monthly average rainfall could be improved by considering the eXogenous factors and using advanced techniques to find out the snow water equivalent amount to replace the missing rainfall data for the months where Minot experiences heavy snowfall.

Funding

The authors would like to acknowledge the financial support from the ND EPSCoR STEM grants program for this research.

References

[1] Ademola, A., Emmanuel, C. O., and Aderemi, K. A. (2018). Statistical Modeling of Monthly Rainfall in Selected Stations in Forest and Savannah Eco-climatic Regions of Nigeria. *Journal of Climatology & Weather Forecasting*. 6: S1. DOI: 10.4172/2332-2594.1000226.

[2] Akaike, H. Fitting autoregressive models for prediction. *Ann Inst Stat Math* 21, 243–247 (1969). https://doi.org/10.1007/BF02532251.

[3] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions of Automatic Control, 19 (6): 716-723.

[4] Bee Dagum, E. (2010). Time series modeling and decomposition. *Statistica*, *70* (4), 433–457. https://doi.org/10.6092/issn.1973-2201/3597.

[5] Box, G. E. P., and David A. Pierce. "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models." *Journal of the American Statistical Association* 65, no. 332 (1970): 1509–26. https://doi.org/10.2307/2284333.

[6] Carlos M. Jarque, Anil K. Bera, Efficient tests for normality, homoscedasticity and serial independence of regression residuals, Economics Letters, Volume 6, Issue 3, 1980, Pages 255-259, ISSN 0165-1765, https://doi.org/10.1016/0165-1765(80)90024-5.

[7] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, *74* (366), 427–431. https://doi.org/10.2307/2286348.

[8] E. Mair, G. Leitinger, S. Della Chiesa, G. Niedrist, U. Tappeiner & G. Bertoldi (2016) A simple method to combine snow height and meteorological observations to estimate winter precipitation at sub-daily resolution, Hydrological Sciences Journal, 61: 11, 2050-2060, DOI: 10.1080/02626667.2015.1081203.

[9] Hyndman, R. J., & Athanasopoulos, G. (2018) Forecasting: Principles and Practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

[10] Jan de Leeuw. (2011). Information Theory and an Extension of the Maximum Likelihood Principle by Hirotogu Akaike. UCLA: Department of Statistics, UCLA. Retrieved from https://escholarship.org/uc/item/0fd986xb.

[11] National Weather Service, National Operational Hydrological Remote Sensing Center. Retrieved from https://www.nohrsc.noaa.gov/

[12] National Research Council. 2010. When Weather Matters: Science and Services to Meet Critical Societal Needs. Washington, DC: The National Academies Press. https://doi.org/10.17226/12888.

[13] Prema. V, Rao, U. (2015), Time series decomposition model for accurate wind speed forecast, Renewables: Wind, Water, and Solar V-2, DOI 10.1186/s40807-015-0018-9.

[14] Rong-Gang Cong, Mark Brady, “The independence between Rainfall and Temperature: Copula Analysis”, The Scientific Journal, vol. 2012, Article ID 405675, 11 pages, 2012. https://doi.org/10.1100/2012/405675.

[15] Schwarz, G. (1978). “Estimating the dimension of a model”. Annals of Statistics, 6 (2): 461-464.

[16] Sugiura, N. (1978). “Further analysis of the data by Akaike’s information criterion and the finite correlations”. Communications of Statistics-Theory and Methods, A7: 13-26.

[17] Tektaş, Mehmet. (2010). Weather forecasting using ANFIS and ARIMA models. A case study for Istanbul. Environmental Research, Engineering and Management 51 (1): 5–10.

[18] Teshome Hailemeskel Abebe. Time Series Analysis of Monthly Average Temperature and Rainfall Using Seasonal ARIMA Model (in Case of Ambo Area, Ethiopia). *International Journal of Theoretical and Applied Mathematics*. Vol. 6, No. 5, 2020, pp. 76-87. doi: 10.11648/j.ijtam.20200605.13.

[19] Shumway. R. H, Stoffer, D. S., Time Series Analysis and Its Applications, 4ed, ISSN 1431-875X, Springer Text in Statistics, ISBN 978-3-319-52451-1, DOI 10.1007/978-3-319-52452-8.

[20] U.S. Climate Normals. National Centers for Environmental Information (NCEI). Retrieved April 23, 2022, from https://www. ncei. noaa. gov/products/land-based-station/us-climate-normals.