Regression and Time Series Analysis for Climate Trends in Kazakhstan

Madina Sarsembayeva

May 4, 2025

`Capstone Thesis, Spring 2025`

Department of Mathematics, School of Sciences and Humanities
Nazarbayev University
`Supervisor:` Dr. Piotr Skrzypacz
`Second reader:` Dr. Erum Rehman

Abstract

Climate change is an important issue for Kazakhstan due to its diverse geography, which results in varying temperature and precipitation patterns in different regions. This study examines climate trends in Kazakhstan using B-spline smoothing, functional regression, and time series forecasting models, SARIMA and ETS. The analysis focuses mainly on temperature and precipitation data from 2000 to 2024. To begin, B-spline smoothing was applied to refine the data, while functional regression helped to explore deeply the relationship between temperature and precipitation. A correlation matrix revealed positive, negative, and no linear relationships between seven climate factors. The SARIMA and ETS models were then used to predict temperature trends for the next 24 months. The SARIMA model effectively captured seasonal variations, while the ETS model delivered more precise forecasts, reflecting significant seasonal fluctuations. These results provide valuable information on Kazakhstan's climate, helping to shape strategies for agriculture and sustainable development going forward.

1 Introduction

This study aims to examine the trends of climate in Kazakhstan between 2000 and 2024, specifically using data from the capital Astana. Kazakhstan, situated in Central Asia, has experienced varying climate changes in recent decades, particularly in temperature and precipitation patterns. Although some researchers

argue that these changes align with global climate trends, others suggest that they may be part of natural regional variability. Recent studies have shown a marked increase in temperature extremes, especially during the summer months. This warming trend aligns with global patterns, as observed by Salnikov et al. (2023), who documented a significant increase in heatwaves and temperature extremes in Kazakhstan in recent decades [1]. On the other hand, precipitation patterns have shown more variability, with fewer dramatic changes compared to increased temperatures, as Fallah et al. (2024) pointed out [2]
. The World Bank (2021) also notes that the effects of warming in Kazakhstan are not uniform across the country. The northern and central regions have seen more pronounced temperature increases, while the southern areas have experienced more stable temperature changes [3]. These regional differences highlight the complex nature of climate change in Kazakhstan and emphasize the need for more detailed studies to better understand long-term climate trends.

In this study, we analyzed historical Kazhydromet climate data and used advanced modeling techniques to investigate the relationship between temperature and precipitation. First, we performed a correlation analysis by plotting a correlation matrix between seven climate variables such as temperature, precipitation, relative humidity, evapotranspiration, cloud cover, duration of sunshine, and wind speed to examine how temperature is related to other climate variables and their linear relationships between each other. We also applied B-spline smoothing to generate smooth, continuous representations of the data. The functional regression was then used to model the interaction between temperature and precipitation.

For forecasting, we used two well-established time series models, Seasonal Autoregressive Integrated Moving Average (SARIMA) and Exponential Smoothing State Space (ETS). The first model, SARIMA, effectively captured seasonal patterns with long-term dependencies in the data, while the second model, ETS, which accounts for both trends and seasonality, provided more accurate forecasts. We compared the results of both models, presenting the temperature forecast for 2025 and 2026 along with the 95% confidence intervals. The study concludes with a discussion on the precision of both models and their usefulness in understanding future climate trends in Kazakhstan.

2 Theoretical Part

2.1 Correlation Set and Matrix

The concept of correlation is fundamental in statistics. In the context of time series analysis and climate data, correlations are used to quantify the strength and direction of the relationship between different variables, such as temperature, precipitation, evapotranspiration, wind speed and other.
The Pearson correlation coefficient measures the linear relationship between two variables [4]. It is calculated as:

$$\rho(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}$$

- $\text{Cov}(X_i, X_j)$ is the covariance between variables X_i and X_j, which measures how the variables change together.
- σ_{X_i} and σ_{X_j} are the standard deviations of X_i and X_j, respectively, which measure the variability of each variable.

The correlation coefficient $\rho(X_i, X_j)$ ranges from -1 to 1:

- $\rho = 1$ means a perfect positive linear relationship,
- $\rho = -1$ means a perfect negative linear relationship,
- $\rho = 0$ means no linear relationship between the variables.

A correlation matrix is a square matrix that contains the pairwise correlation coefficients between all pairs of variables. For the n variables $X_1, X_2, \ldots, X_n$, the matrix is of size $n \times n$, and each element $\rho(X_i, X_j)$ represents the correlation between the corresponding pair of variables X_i and X_j. The correlation matrix R is represented as:

$$R = \begin{bmatrix} \rho(X_1, X_1) & \rho(X_1, X_2) & \ldots & \rho(X_1, X_n) \\ \rho(X_2, X_1) & \rho(X_2, X_2) & \ldots & \rho(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(X_n, X_1) & \rho(X_n, X_2) & \ldots & \rho(X_n, X_n) \end{bmatrix}$$

Since the correlation of a variable with itself is always 1, the diagonal elements of the correlation matrix are all equal to 1:

$$\rho(X_i, X_i) = 1 \quad \text{for all} \quad i = 1, 2, \ldots, n.$$

2.2 B-Spline Smoothing and Functional Regression Model

The analysis of climate data begins with smoothing of the raw temperature and precipitation data. B-splines, which are piecewise polynomial functions, are used to create a smooth representation of the data [5]. The B-spline function for temperature $Y(t)$ at time t is given by:

$$S(t) = \sum_{i=1}^{n} c_i B_i(t)$$

- $S(t)$ is the smoothed temperature or precipitation at time t,
- c_i are the coefficients of the B-spline functions,
- $B_i(t)$ are the B-spline basis functions,
- t is the time, represented by the dates from January 1, 2000, to December 31, 2024.

As an example, after applying the smoothing function, the following values for January 2, 2000, are obtained: smoothed temperature: -21.03°C, smoothed precipitation: 2.03 mm.
These smoothed values are then used in the regression analysis. Once the data are smoothed, we proceed with functional regression to explore the relationship between temperature $Y(t)$ and precipitation $X(t)$. The general regression model is as follows:

$$Y(t) = \alpha(t) + \beta(t) \cdot X(t) + \epsilon(t)$$

- $Y(t)$ is the smoothed temperature,
- $X(t)$ is the smoothed precipitation,
- $\alpha(t)$ is the intercept term (constant over time),
- $\beta(t)$ is the regression coefficient representing the influence of precipitation on temperature,
- $\epsilon(t)$ represents the residuals (errors).

The coefficients $\beta(t)$ and $\alpha(t)$ are estimated using the least squares method to minimize the sum of squared residuals. The formula for calculating the regression coefficient β is:

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

- X_i and Y_i are the individual data points for precipitation and temperature,
- $\bar{X}$ and $\bar{Y}$ are the means of precipitation and temperature, respectively.

For our dataset, for January 2, 2000, we can calculate the regression coefficient:

$$\begin{aligned}\frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2} &= \frac{(2.03 - 0.446)(-21.03 - (-21.30))}{(2.03 - 0.446)^2} \\ &= \frac{(1.584)(0.27)}{(1.584)^2} \\ &= \frac{0.42768}{2.50906} \\ &\approx 0.17\end{aligned}$$

This calculation will be repeated for other days to estimate $\hat{\beta}$.
Finally, the residuals ($\epsilon(t)$) are calculated as the difference between the observed values and the predicted values:

$$\epsilon(t) = Y(t) - \hat{Y}(t)$$

where $\hat{Y}(t)$ is the predicted temperature. The residuals are then plotted to evaluate the fit of the regression model. A well-fitting model will show residuals that are randomly scattered around zero with no discernible pattern.

2.3 SARIMA model

Time series prediction involves predicting future values of a variable that changes over time, based on historical data. The Seasonal AutoRegressive Integrated Moving Average (SARIMA) model is an enhancement of the ARIMA (AutoRegressive Integrated Moving Average) model. Although the ARIMA model is great for capturing trends and cycles, the SARIMA model incorporates additional seasonal components to better account for patterns that repeat at regular intervals, making it especially useful for data with seasonal fluctuations [6].
The general form of the SARIMA model can be written as:

$$Y_t = \mu + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

- Y_t is the time series value at time t,
- μ is the mean of the series,
- $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive (AR) coefficients,

- ϵ_t is the error term (residuals),
- $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average (MA) coefficients.

The seasonal component is added to account for periodic fluctuations (seasonality) in the data. The SARIMA model is represented as:

$$(Y_t - \mu) = \phi_1(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} + \epsilon_t$$

Where the seasonal order is represented as (P, D, Q, s).
So, the model can be written as:

$$\text{SARIMA}(p, d, q)(P, D, Q)_s$$

- p, d, q are the non-seasonal autoregressive order, differencing order, and moving average order, respectively,
- P, D, Q are the seasonal counterparts of p, d, q, respectively, representing the seasonal autoregressive order, seasonal differencing order, and seasonal moving average order,
- s is the length of the seasonal cycle (e.g., 12 for monthly data).

2.4 ETS model

Exponential Smoothing State Space Models (ETS) are useful for data showing trends and seasonality [7]. The ETS model consists of three main components: level (l_t): represents the smoothed value of the time series at time t; trend (b_t): represents the change in the level over time, capturing the underlying trend in the data; seasonality (s_t): represents periodic fluctuations in the time series, capturing seasonal variations.
The general form can be written as:

$$Y_t = \ell_t + b_t + s_t$$

- Y_t is the observed value at time t,
- ℓ_t is the level (or smoothed value),
- b_t is the trend component,
- s_t is the seasonal component.

The ETS model can be represented in terms of error, trend, and seasonality equations as follows:

$$\ell_t = \alpha Y_{t-1} + (1-\alpha)(\ell_{t-1} + b_{t-1})$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}$$

$$s_t = \gamma(Y_t - \ell_t) + (1-\gamma)s_{t-s}$$

- α, β, and γ are the smoothing parameters for the level, trend, and seasonality, respectively,
- s represents the seasonal period,
- Y_t is the actual data value at time t, and
- ℓ_t, b_t, and s_t are the level, trend, and seasonal components at time t.

3 Analysis of Temperature Data for Kazakhstan

In this section, we will analyze the temperature trends for Astana, Kazakhstan, using the graphs. The goal is to explore long-term trends, seasonal patterns, and variations in climate data and to identify potential changes in Kazakhstan's climate over time. The analysis is mainly based on the data for 2000-2024 years from Kazhydromet.

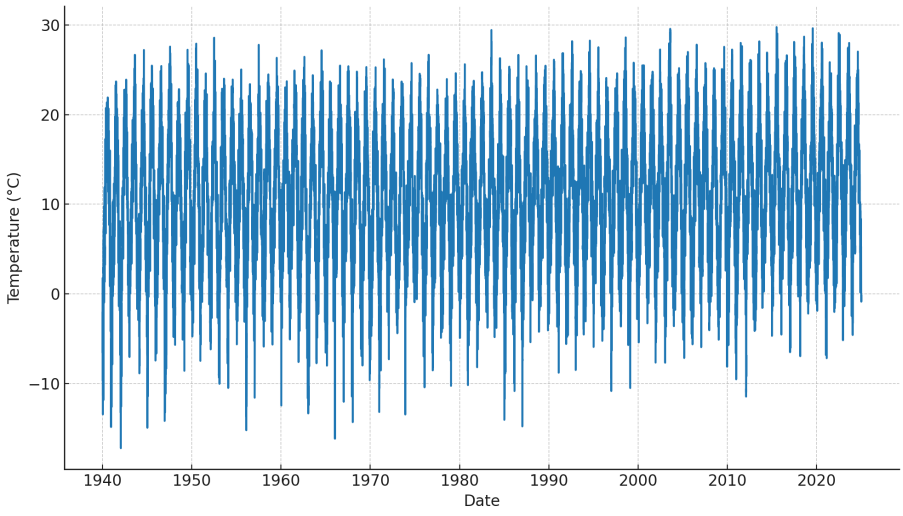

Figure 1: Mean Temperature Trends Over Time (1940-2024)

The graph in Fig. 1 that shows mean temperature trends from 1940 to 2024 highlights a consistent increase in temperatures, pointing to a long-term warming pattern in line with global climate change. Although there are typical seasonal variations—colder winters and warmer summers—the overall trend shows a gradual rise in cold temperatures over the years. This warming trend seems to have accelerated after the 1980s, especially in recent years post-2000. This shift can be considered an effect of climate change, but it is not as extreme as it could be.

Next, we will compare daily maximum and minimum temperatures for 2000 and 2024 years.

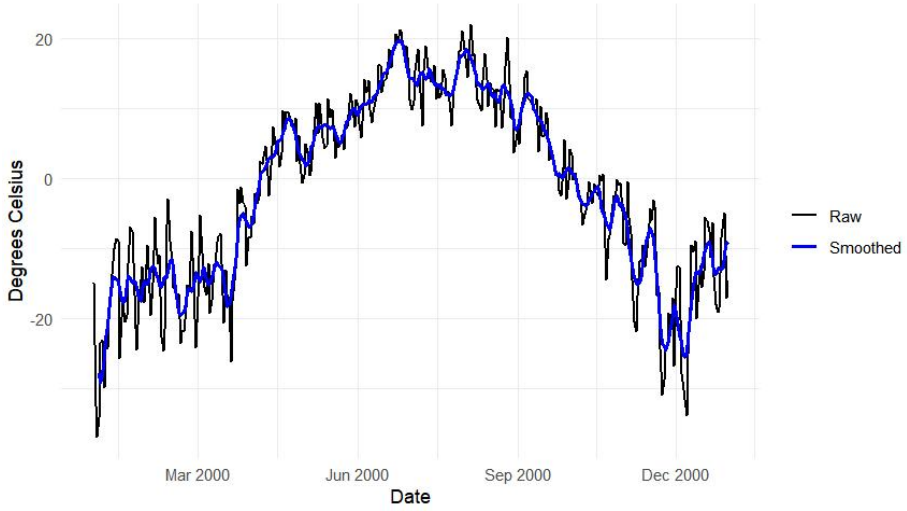

Figure 2: Daily Minimum Temperature for 2000

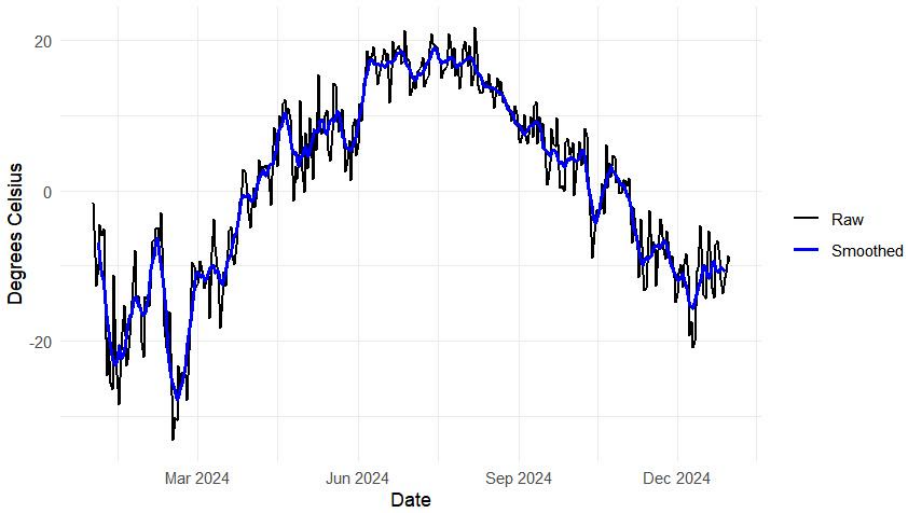

Figure 3: Daily Minimum Temperature for 2024

The graphs in Figures 2 and 3 illustrate the daily minimum temperature in Kazakhstan for two years: 2000 and 2024. They show fluctuations around a seasonal pattern, with noticeable cold periods in the winter months and milder temperatures in summer. The raw data (black line) are compared with the smoothed version (blue line), helping to highlight long-term temperature trends. Seasonal variations and the smoothed line provide a clearer picture of the overall trend of temperature for the year. The graphs suggest that Kazakhstan experiences some seasonal temperature shifts and the overall temperature profile in 2024 appears to be similar to 2000, although slight differences in the magnitude of temperature extremes are noticeable.

The graphs in Figures 4 and 5 show the maximum daily temperature for 2000 and 2024 years. Both years show typical seasonal fluctuations, with colder winters and warmer summers. The raw data (black line) reveal daily variations, while the smoothed data (red line) highlight long-term trends. In both years, the temperature patterns are similar, showing seasonal cycles with higher summer and lower winter temperatures. The main difference between 2000 and 2024 lies in slightly milder winters and possibly more intense summer heat in 2024, but the overall temperature trend remains stable.

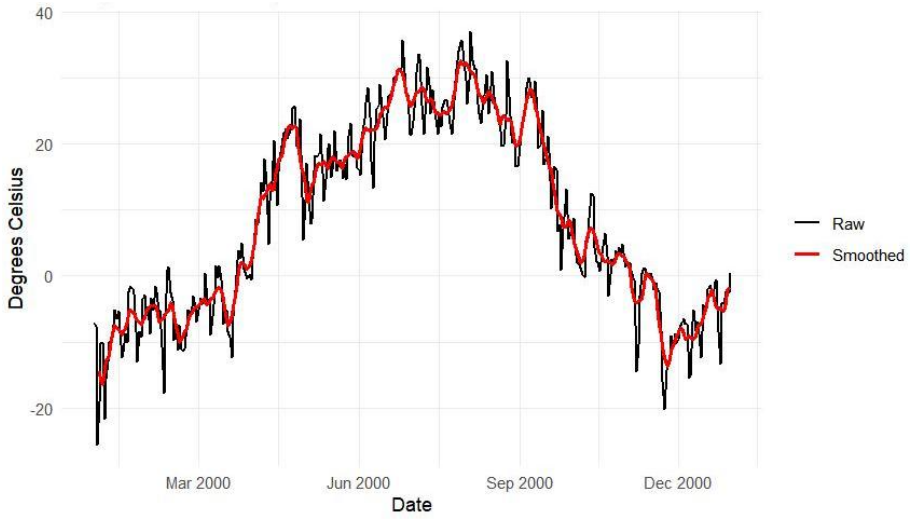

Figure 4: Daily Maximum Temperature for 2000

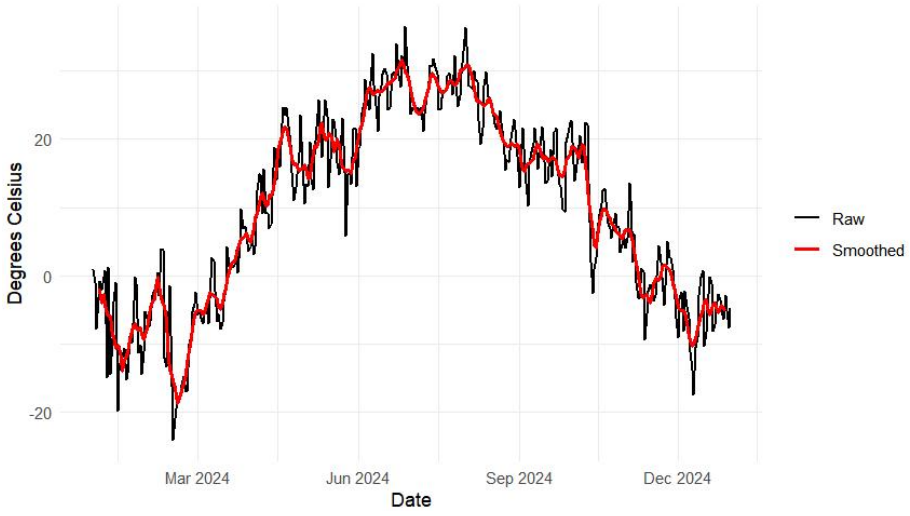

Figure 5: Daily Maximum Temperature for 2024

4 Results and Discussion

4.1 Correlation Matrix

In RStudio, a correlation matrix of various weather variables was created to explore the relationships between seven variables: temperature, total precipita-

tion, relative humidity, wind speed, total cloud cover, duration of sunshine and evapotranspiration. As shown in Fig. 6, the correlation between temperature and sunshine duration is moderately positive at 0.53, which means that warmer temperatures generally coincide with longer periods of sunshine. Another strong positive correlation is between temperature and evapotranspiration (0.76), suggesting that higher temperatures tend to drive higher evapotranspiration rates. In contrast, temperature has a negative correlation with relative humidity (-0.45), indicating that as temperatures rise, relative humidity tends to decrease. Precipitation and relative humidity are moderately positively correlated (0.36), which makes sense because higher humidity often leads to increased rainfall. Wind speed and cloud cover show a very strong positive correlation (0.92), which implies that stronger winds are typically associated with more cloud cover, possibly due to the movement of moisture-laden air. There is also a weak positive linear relationship between wind speed and sunshine duration (0.37), which shows that slightly higher wind speeds may coincide with longer periods of sunshine. Evapotranspiration and the duration of sunshine share a strong positive correlation (0.64), reinforcing the idea that more sunshine leads to higher evapotranspiration. Lastly, there is a very weak correlation (0.02) between temperature and precipitation, indicating that there is no significant linear relationship.

These correlations highlight how interrelated climate variables are and offer valuable insights into how different factors influence each other in the climate system.

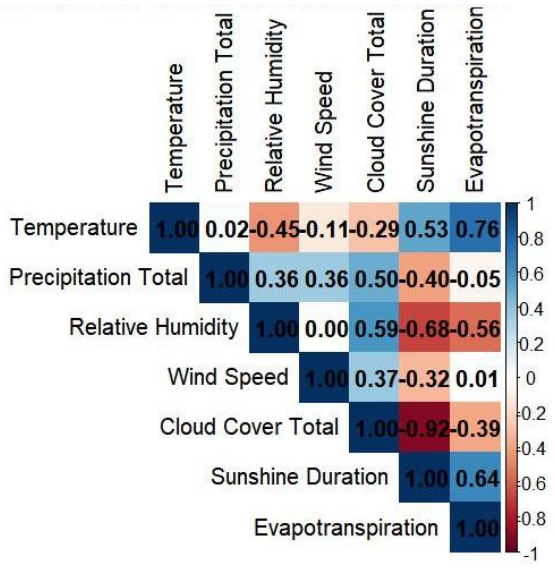

Figure 6: Correlation Matrix of Climate Variables

4.2 B-Spline Smoothing and Functional Regression Model

To see how weak the relationship is between temperature and precipitation, we conducted a regression analysis. We first applied the 3-day moving average method to smooth both the temperature and precipitation data. This method helps to reduce short-term fluctuations and highlight the underlying trends. The part of the smoothed data for temperature and precipitation is shown in Table 1.

Date	**Original Mean Temperature (°C)**	**Smoothed Mean Temperature (°C)**	**Original Precipitation (mm)**	**Smoothed Precipitation (mm)**
2000-01-01	-8.5	-	5.5	-
2000-01-02	-22.9	-21.03	0.6	2.03
2000-01-03	-31.7	-26.63	0	0.2
2000-01-04	-25.3	-24.53	0	0
2000-01-05	-16.6	-19.16	0	0
...	...	...	...	...
2024-12-30	-8.1	-7.7	0.7	0.46
2024-12-31	-7.4	-	0.2	-

Table 1: Smoothed data for temperature and precipitation.

The first step is to calculate the mean values of the smoothed temperature and the smoothed precipitation:

$$\bar{Y} = \frac{-21.03 + (-26.63) + (-24.2) + (-18.9) + (-15.73)}{5} = -21.30$$

$$\bar{X} = \frac{2.03 + 0.2 + 0 + 0 + 0}{5} = 0.446$$

Now, using the formula for the regression coefficient $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

For Jan 2:

$$(X_1 - \bar{X}) = 2.0333 - 0.9721 \approx 1.0612,$$

$$(Y_1 - \bar{Y}) = -21.0333 - 4.41058 \approx -25.4439,$$

$$(X_1 - \bar{X})(Y_1 - \bar{Y}) = 1.0612 \times (-25.4439) \approx -27.0008$$

For Jan 3:

$$(X_2 - \bar{X}) = 0.2 - 0.9721 \approx -0.7721,$$

$$(Y_2 - \bar{Y}) = -26.6333 - 4.41058 \approx -31.04388,$$

$$(X_2 - \bar{X})(Y_2 - \bar{Y}) = (-0.7721) \times (-31.04388) \approx 23.9703$$

Summing all terms:

$$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = -27.0008 + 23.9703 + 28.1376 + \cdots + 6.1216$$
$$\approx 20798.03635$$

Now calculate the denominator:

$$\sum(X_i - \bar{X})^2 = 1.1261 + 0.5962 + 0.9451 + \cdots + 0.2555$$
$$\approx 27260.29062$$

Thus, the regression coefficient is:

$$\hat{\beta} = \frac{20798.03635}{27260.29062}$$
$$\approx 0.762943$$

Now, calculate the intercept using the formula:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \cdot \bar{X}$$

$$\hat{\alpha} = 4.4106 - (0.7629 \times 0.9721) = 4.4106 - 0.7417 = 3.6689$$

For Jan 2:

$$\hat{Y}_2 = 3.6689 + (0.7629 \times 2.0333) = 3.6689 + 1.5513 = 5.2202$$

Residual:

$$\epsilon_2 = -21.0333 - (5.2202) = -26.2535$$

The summing of the residuals for all days follows the same pattern.
For residuals:

$$\begin{aligned}\sum \epsilon_i^2 &= (-26.2535)^2 + (-30.4548)^2 + (-28.2022)^2 + \cdots + (-11.725)^2 \\ &= 689.2486 + 927.4959 + 795.3657 + \cdots + 137.4741 \\ &= 1754825.024\end{aligned}$$

For the observed values:

$$\begin{aligned}\sum (Y_i - \bar{Y})^2 &= 647.3929 + 963.7248 + 837.7504 + \cdots + 146.6662 \\ &= 1770692.732\end{aligned}$$

Thus, R^2 is:

$$\begin{aligned}R^2 &= 1 - \frac{1754825.024}{1770692.732} \\ &\approx 1 - 0.99104 \\ &= 0.00896\end{aligned}$$

This regression analysis reveals a very weak relationship between temperature and precipitation because the value of R^2 is a low 0.00896. Only about 0.9% of the temperature variance is explicable by precipitation. Therefore, the model does not capture an important part of the variability in temperature. Precipitation is not a strong predictor of temperature in this context in that such a low R^2 value indicates even though weather data tend to show complex nonlinear interactions that are influenced by many external factors. Humidity, as well as wind speed, in addition to regional atmospheric patterns, are other variables that may always play a more important role in the determination of temperature fluctuations. This weak model fit suggests that further factors or more advanced modeling techniques may be necessary to better explain temperature trends given the complexity and natural variability in climate data.

4.3 SARIMA and ETS Forecasting

The graph of the time series of the monthly average temperature data for the time range between 2000 and 2024 is visualized in Fig. 7

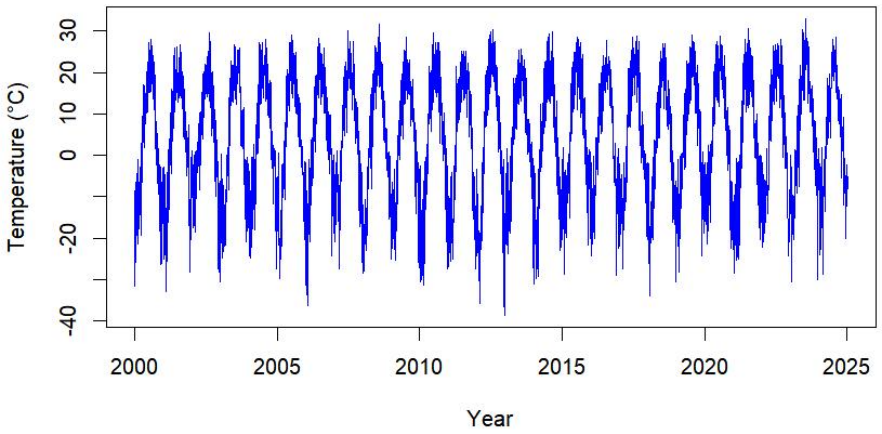

Figure 7: Temperature Time Series for Astana (2000-2024)

We applied the SARIMA and ETS models to historical data to forecast the temperature for the years 2025 and 2026. Since our dataset had seasonal variations, we had plotted the data to check for seasonality, and we did see seasonal patterns, which we will then use to create our time series forecasting models.
Unlike any of the transformations used previously, we fit up the SARIMA model for trend and seasonal component of the data. We first used the Augmented Dickey-Fuller (ADF) test to check the stationarity, and confirmed it was stationary. The results are the ADF statistic: -6.61; p-value: 6.57e-09; critical values: 1%: -3.43; 5%: -2.86; 10%: -2.57. Since the p-value is very small (< 0.05), we can reject the null hypothesis.
The autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the rescaled data were then plotted, as shown in Fig. 8. They help assess whether an AR(p) or MA(q) model is suitable and assist in identifying potential candidate models.

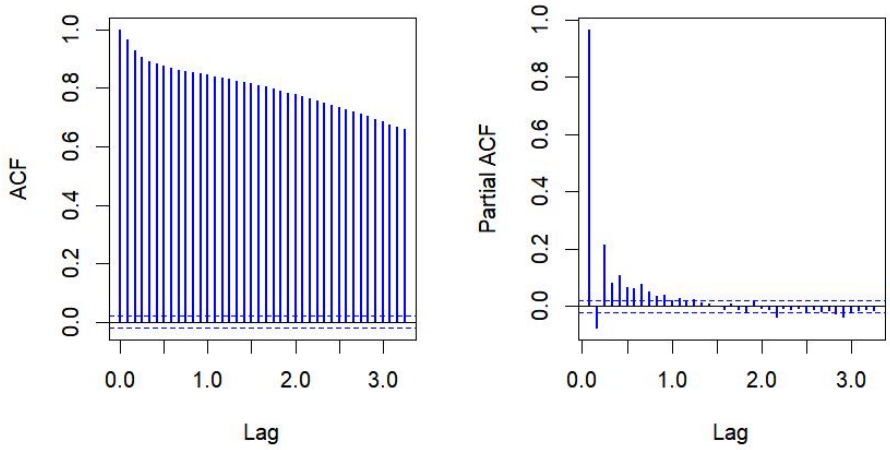

Figure 8: ACF and PACF of Monthly Mean Temperature

The SARIMA model requires us to estimate several parameters. In our case, the model was fitted as:

$$\text{SARIMA}(1,0,1)(1,1,1)_{12}$$

- (1, 0, 1): The non-seasonal part of the model, representing AR(1), differencing (I(0)), and MA(1),
- (1, 1, 1, 12): The seasonal part of the model, with seasonal AR(1), differencing (I(1)), and MA(1) with a seasonal period of 12 months.

Here is the model summary based on the fitting process in Table 2:

Component	Description	Value
AR(1)	Autoregressive term for lag 1	0.1054
MA(1)	Moving Average term for lag 1	0.0872
Seasonal AR(1)	Seasonal autoregressive term for lag 1	-0.1551
Seasonal MA(1)	Seasonal moving average term for lag 1	-1.1613
AIC	Akaike Information Criterion	1340.3
BIC	Bayesian Information Criterion	1358.4

Table 2: Key Results from SARIMA Model Fitting

The model has performed well on the monthly data, and the Ljung-Box test suggests that there is no significant autocorrelation in the residuals (p-value = 0.91). Then a histogram and a Quantile-Quantile (Q-Q) plot were also plotted to assess the normality of the residuals in Figures 9 and 10.

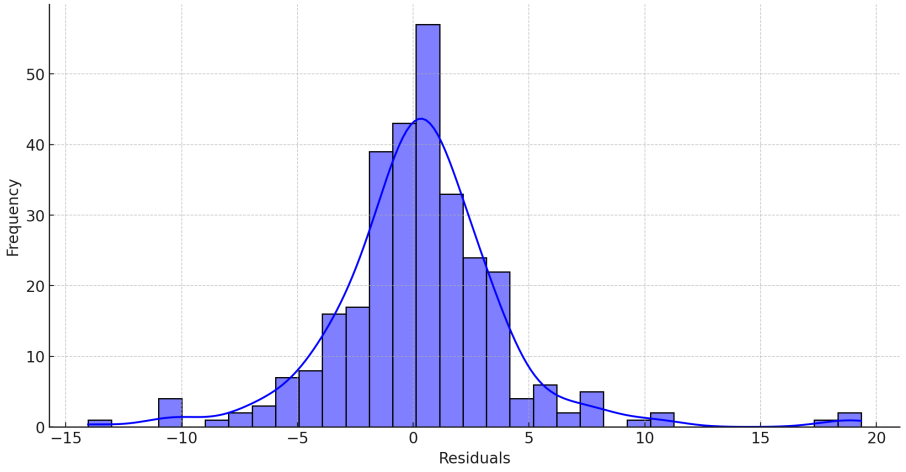

Figure 9: Histogram of SARIMA Model Residuals

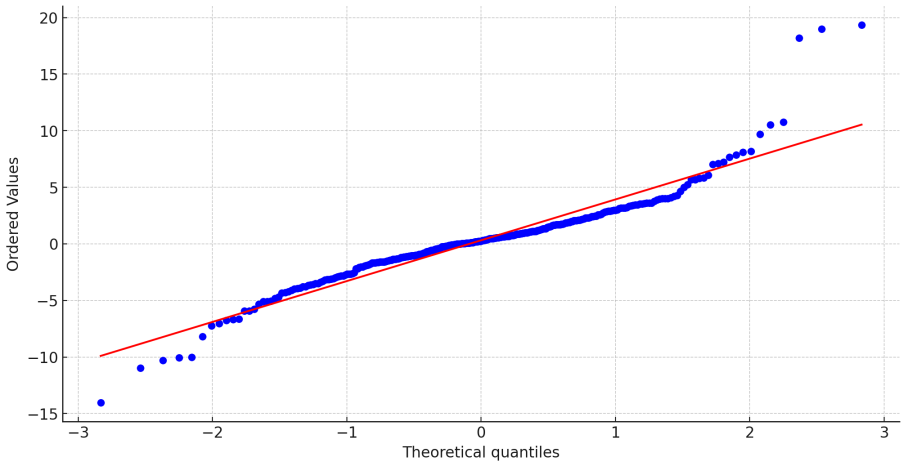

Figure 10: Q-Q Plot of SARIMA Model Residuals

Using the fitted SARIMA model, we forecast the next 24 months (2025 and 2026).

The forecast equation for SARIMA is:

$$\hat{Y}_{t+h} = \mu + \phi_1 \hat{Y}_{t+h-1} + \phi_2 \hat{Y}_{t+h-2} + \cdots + \phi_p \hat{Y}_{t+h-p} + \theta_1 \hat{\epsilon}_{t+h-1} + \cdots + \theta_q \hat{\epsilon}_{t+h-q} + \epsilon_t$$

- $\hat{Y}_{t+h}$ is the forecasted value at time $t+h$,
- $\hat{\epsilon}_{t+h}$ is the forecast error for time $t+h$.

The forecast temperatures for the next 24 months are shown in Tables 3 and 4.

Date	**SARIMA Forecast**
2025-01-01	-13.4126
2025-02-01	-11.9548
2025-03-01	-4.5438
2025-04-01	7.3421
2025-05-01	15.6846
2025-06-01	19.7878
2025-07-01	21.7268
2025-08-01	19.8244
2025-09-01	13.0403
2025-10-01	5.3256
2025-11-01	-4.3892
2025-12-01	-11.5506

Table 3: SARIMA Forecast for 2025

Date	**SARIMA Forecast**
2026-01-01	-13.6888
2026-02-01	-12.2438
2026-03-01	-4.5321
2026-04-01	7.6172
2026-05-01	15.1596
2026-06-01	20.1055
2026-07-01	21.7331
2026-08-01	19.6260
2026-09-01	12.7642
2026-10-01	5.3592
2026-11-01	-4.2544
2026-12-01	-11.1175

Table 4: SARIMA Forecast for 2026

The SARIMA model was successfully fitted to the monthly temperature data to forecast the next 24 months (2025-2026) representing the historical temperature data (blue line) and the forecast trend (red line) in Fig. 11. Like historical data, the forecast follows a similar seasonal path, but the confidence interval 95% (pink shaded area) increases as the forecast progresses. This growing uncertainty is what we expect in a time series forecasting model and indicates that the accuracy of the model decreases further into the future.

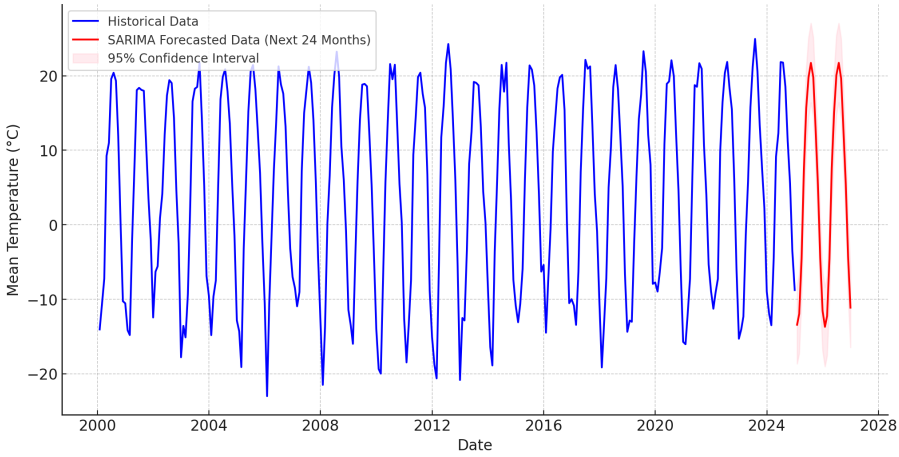

Figure 11: SARIMA Model Forecast (Next 24 months) with 95% Confidence Interval

Next, for our analysis, we fit the ETS model to the temperature data from 2000 to 2024, and the forecast was made for the next 24 months. We used an additive model for both trend and seasonality, with a seasonal period of 12 months. Here is the model summary based on the fitting process in Table 5:

Component	**Description**	**Value**
α	Level Smoothing Parameter	0.1765
β	Trend Smoothing Parameter	0.1314
γ	Seasonality Smoothing Parameter	0.3487
AIC	Akaike Information Criterion	607.56
BIC	Bayesian Information Criterion	666.81
Seasonal Period	Monthly data with yearly seasonality	12

Table 5: Key Results from ETS Model Fitting

The histogram and a Quantile-Quantile (Q-Q) plot of the residuals indicated

that they followed the normal distribution with slight skewness in Figures 12 and 13.

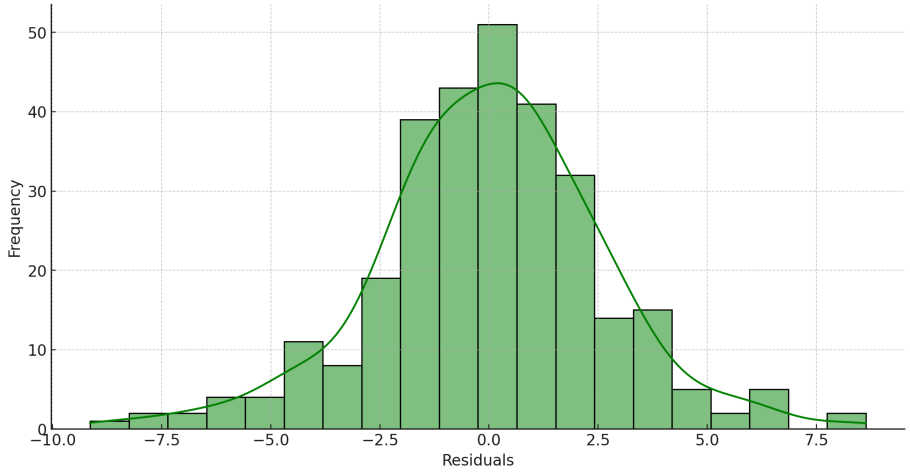

Figure 12: Histogram of ETS Model Residuals

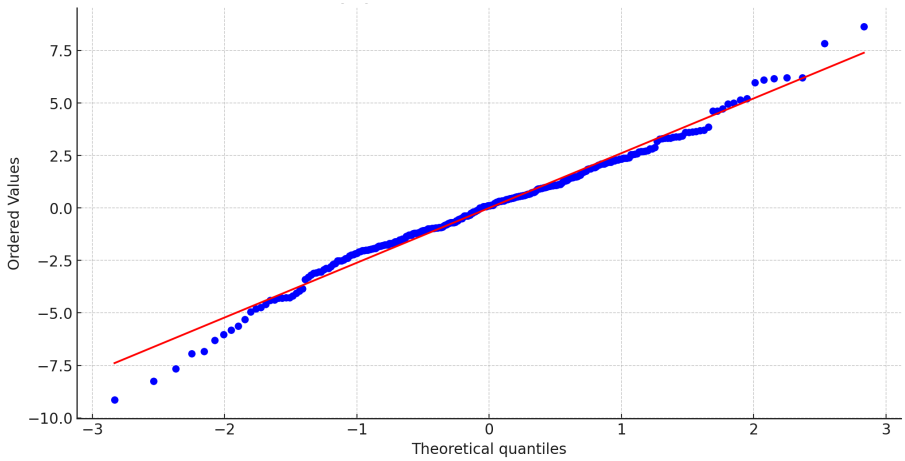

Figure 13: Q-Q Plot of ETS Model Residuals

The forecast equation for ETS is:

$$\hat{Y}_{t+h} = \ell_t + b_t + s_t$$

- $\hat{Y}_{t+h}$ is the forecasted value at time $t + h$,
- ℓ_t is the level component,

- b_t is the trend component,
- s_t is the seasonal component.

Using the fitted ETS model, we forecast the temperature for the next 24 months (2025 and 2026). The forecast values are shown in Tables 6 and 7.

Date	ETS Forecast
2025-01-01	-13.8685
2025-02-01	-12.3493
2025-03-01	-4.0060
2025-04-01	7.7228
2025-05-01	15.5344
2025-06-01	20.4001
2025-07-01	21.4235
2025-08-01	20.0672
2025-09-01	13.5380
2025-10-01	5.5258
2025-11-01	-3.8052
2025-12-01	-11.0864

Table 6: ETS Forecast for 2025

Date	ETS Forecast
2026-01-01	-13.8226
2026-02-01	-12.3034
2026-03-01	-3.9601
2026-04-01	7.7687
2026-05-01	15.5804
2026-06-01	20.4460
2026-07-01	21.4694
2026-08-01	20.1131
2026-09-01	13.5839
2026-10-01	5.5718
2026-11-01	-3.7593
2026-12-01	-11.0405

Table 7: ETS Forecast for 2026

The ETS model was successfully fitted to the monthly temperature data from 2000 to 2024 in Figure 14. The forecast, illustrated with historical data (green) and forecast (blue), tracks similarly to the seasonal trends in the historical data. The 95% confidence interval (gray shaded area) also expands as time passes for 2025-2026, indicating less certainty in the prediction.

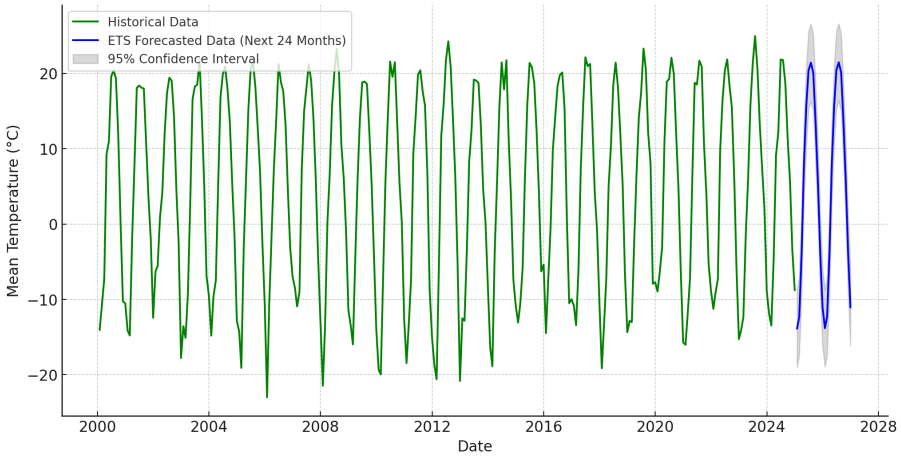

Figure 14: ETS Model Forecast (Next 24 months) with 95% Confidence Interval

To compare the performance of both models, we calculated metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) for each model based on forecast values for 2025 and 2026 and listed in Table 8.

Metric	**SARIMA**	**ETS**
MAE	1.58	1.52
RMSE	2.13	2.06
MAPE	32.63%	29.65%

Table 8: Calculated Metrics

The ETS model performed slightly better in terms of accuracy, as it had lower values of MAE, RMSE, and MAPE. Furthermore, residual diagnostics for ETS did not show significant autocorrelation, while SARIMA showed significant autocorrelation, suggesting that ETS captured the underlying patterns in the data more effectively.

5 Conclusion

In this work, we performed regression and time series analysis for the Kazakhstan, Astana's climate data. Linear relationships between different variables were discussed using the correlation matrix and future temperature values predicted. Forecasting models were effective in capturing seasonal patterns in the

data. The ETS model, with its straightforward approach of breaking down trend and seasonality, produced smoother forecasts and narrower confidence intervals, which made it a more reliable option for long-term predictions. While the SARIMA model, though more complex, showed a greater range in its predicted values. As the forecast period stretched, the confidence intervals for SARIMA grew significantly larger. This analysis serves as an initial step to understand the climate trends in Kazakhstan. Further improvements may include analyzing each region of Kazakhstan, not only Astana.

References

[1] V. Salnikov, Y. Talanov, S. Polyakova, A. Assylbekova, A. Kauazov, N. Bultekov, G. Musralinova, D. Kissebayev, and Y. Beldeubayev, "An assessment of the present trends in temperature and precipitation extremes in kazakhstan," *Climate*, vol. 11, no. 2, p. 33, 2023.

[2] B. Fallah, I. Didovets, M. Rostami, and M. Hamidi, "Climate change impacts on central asia: Trends, extremes, and future projections," *International Journal of Climatology*, 2024.

[3] W. Bank, "Kazakhstan climate change knowledge portal," 2021.

[4] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*. New York, NY: McGraw-Hill, 5th ed., 2005.

[5] C. de Boor, *A Practical Guide to Splines*. New York, NY: Springer, revised ed., 2001.

[6] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ: Wiley, 5th ed., 2015.

[7] R. J. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin, Germany: Springer, 2008.