Time Series Analysis and Forecasting of Global Monthly Temperature Anomalies

Authors: Asish Joel (202103015), Gireesh Reddy(202201122),

Instructor: Dr. Pritam Anand

Institution: Dhirubhai Ambani University

Introduction

Climate change is one of the most important and complex challenges facing humanity. Understanding how global temperatures have evolved over time is critical for evaluating the pace and impact of climate change. One of the most effective ways to track these changes is through *temperature anomalies* — deviations from a reference period average — which help reduce geographical and seasonal variability, offering a more consistent and interpretable metric.

This project focuses on analyzing monthly global temperature anomalies using time series methods. We apply various concepts such as decomposition, transformation, determine stationarity, and model and train the data for forecasting future climate behavior. We also perform a few deep learning techniques such as Feedforward Neural Networks (FFNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU).

1 Dataset

The data for this project comes from the **Global Temperature Time Series** repository, which compiles authoritative records from two major sources:

- **GCAG (Global component of Climate at a Glance)**: Provided by NOAA, GCAG is derived from the Global Historical Climatology Network-Monthly (GHCN-M) and the International Comprehensive Ocean-Atmosphere Data Set (ICOADS). These sources are blended to create a single global land and ocean anomaly product. The anomalies are reported relative to the 20^{th}-century average, and data is available from 1850 to the present.

Each record in the monthly dataset includes:

- A date (formatted as year-month),
- The data source (`gcag`),
- The mean anomaly in degrees Celsius (°C).

This dataset is particularly well-suited for time series analysis because of its consistent monthly frequency, long historical range, and global scale. The pattern of the dataset enables techniques like seasonal decomposition, differencing, and autocorrelation analysis, which are fundamental to modern forecasting models. For this project, we use the **GCAG monthly dataset** due to its extended historical coverage beginning in 1850.

2 Overview of Time Series Forecasting Models

Time series forecasting is a critical component in numerous applications across industries, including finance, healthcare, supply chain, and climate modeling. At its core, time series forecasting involves analyzing sequential data points collected over time to project future values. The effectiveness of this process hinges on selecting and employing suitable models that can understand and replicate the underlying patterns in the data. These patterns often include trends, seasonality, cyclic behavior, and noise.

To address this challenge, a broad spectrum of models has been developed, which generally fall into two categories: classical statistical models and modern deep learning-based approaches. Each type of model offers unique mechanisms for understanding time-based dependencies, and their effectiveness depends heavily on the specific nature of the dataset and the forecasting goals.

Statistical models serve as foundational tools in time series forecasting due to their transparency, computational efficiency, and interpretability. Among them, the **Autoregressive (AR) model** is one of the most straightforward and widely used. It operates on the premise that future values of a variable can be linearly regressed on its own past values. This linear dependence assumes that a time series has some level of memory, where past observations carry predictive power for the future. The model learns coefficients for different time lags, and these coefficients indicate the strength and direction of influence of previous observations. For example, in an AR(2) model, the predicted value at time t is based on a linear combination of values at times $t-1$ and $t-2$.

In contrast, the **Moving Average (MA) model** captures the influence of past forecast errors rather than past observations themselves. The rationale behind this model is that the prediction errors from earlier forecasts may reflect patterns not captured by the deterministic components of the model, and thus, they provide useful information. For instance, if a forecast consistently underestimates values after a sudden change, the MA model can adapt to compensate for this by adjusting future predictions based on previous residuals. An MA(1) model, for example, adjusts the current prediction based on the error from the immediately preceding time step.

To harness the strengths of both autoregressive behavior and error modeling, the **ARMA (Autoregressive Moving Average)** model combines AR and MA components into a single framework. This model is highly effective for stationary time series—those with constant mean and variance over time. ARMA works by balancing the linear relationship of past values with the corrections from past errors, thereby producing a more nuanced forecast. However, real-world time series often deviate from stationarity, especially in the presence of trends, periodic fluctuations, or structural breaks.

To accommodate these complexities, the **SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables)** model extends ARMA in several important ways. First, it introduces differencing to handle trends, making non-stationary data more suitable for modeling. Second, it incorporates seasonal elements by repeating the AR and MA components over defined seasonal periods (e.g., months or quarters). Third, it allows the inclusion of external or exogenous variables that are believed to influence the target series, such as holidays in retail sales forecasts or temperature in energy consumption models. SARIMAX, therefore, is a highly flexible model capable of handling both univariate and multivariate time series data with complex seasonal and external dynamics.

Despite the robustness of statistical models, they often assume linearity and are limited in capturing complex, nonlinear relationships inherent in many time series datasets. This limitation has led to the rise of **deep learning models**, which excel at learning directly from raw data with minimal assumptions. **Recurrent Neural Networks (RNNs)** are a foundational deep learning architecture for sequential data. They maintain an internal memory state that updates at each time step, allowing the model to incorporate information from the past into future predictions. The hidden state acts as a dynamic representation of the time series' history. However, traditional RNNs are known to suffer from problems like vanishing or exploding gradients, which make them ineffective at learning long-range dependencies over extended sequences.

To overcome this, **Long Short-Term Memory (LSTM)** networks were introduced. LSTMs enhance the basic RNN structure by adding a memory cell and three types of gates—input, forget, and output gates. These gates regulate the flow of information, deciding what to keep, what to discard, and what to output at each time step. This architecture allows the model to preserve information across long sequences and ignore irrelevant data. For instance, in financial forecasting, where market behavior today might be influenced by events from several weeks ago, LSTMs can maintain that contextual memory across time, making them more accurate and insightful.

Gated Recurrent Units (GRUs) are a streamlined variant of LSTMs. They combine the forget and input gates into a single update gate and simplify the internal architecture, which leads to faster

training and lower computational cost. GRUs also perform well on long sequences and often match or even exceed LSTM performance in practice, depending on the dataset. Because of their efficiency, GRUs are commonly used in real-time systems or environments where quick inference and lightweight models are necessary.

Together, these statistical and deep learning models form a comprehensive toolkit for addressing a wide range of time series forecasting challenges. Statistical models like AR, MA, ARMA, and SARIMAX provide interpretable and theoretically grounded methods ideal for structured, stationary data or situations where understanding the underlying process is important. Deep learning models like RNNs, LSTMs, and GRUs, on the other hand, offer powerful mechanisms for modeling highly dynamic, nonlinear, and large-scale datasets. The decision to use one type of model over another—or even a hybrid approach—should be guided by the data characteristics, the forecasting horizon, the need for model explainability, and the availability of computational resources. As time series data continues to grow in complexity and volume, the integration of traditional methods with machine learning innovations will remain key to producing accurate, reliable, and scalable forecasts.

3 Exploratory Data Analysis (EDA)

The exploratory data analysis has been performed on the dataset which began by loading and preprocessing the dataset. Filtering has been done in order to retain only records from the GCAG source. The `Date` column was converted into a proper datetime format and set as the index, which enables us to do time series operations.

This step ensured the dataset was properly structured and ready for subsequent analysis using time series forecasting methods.

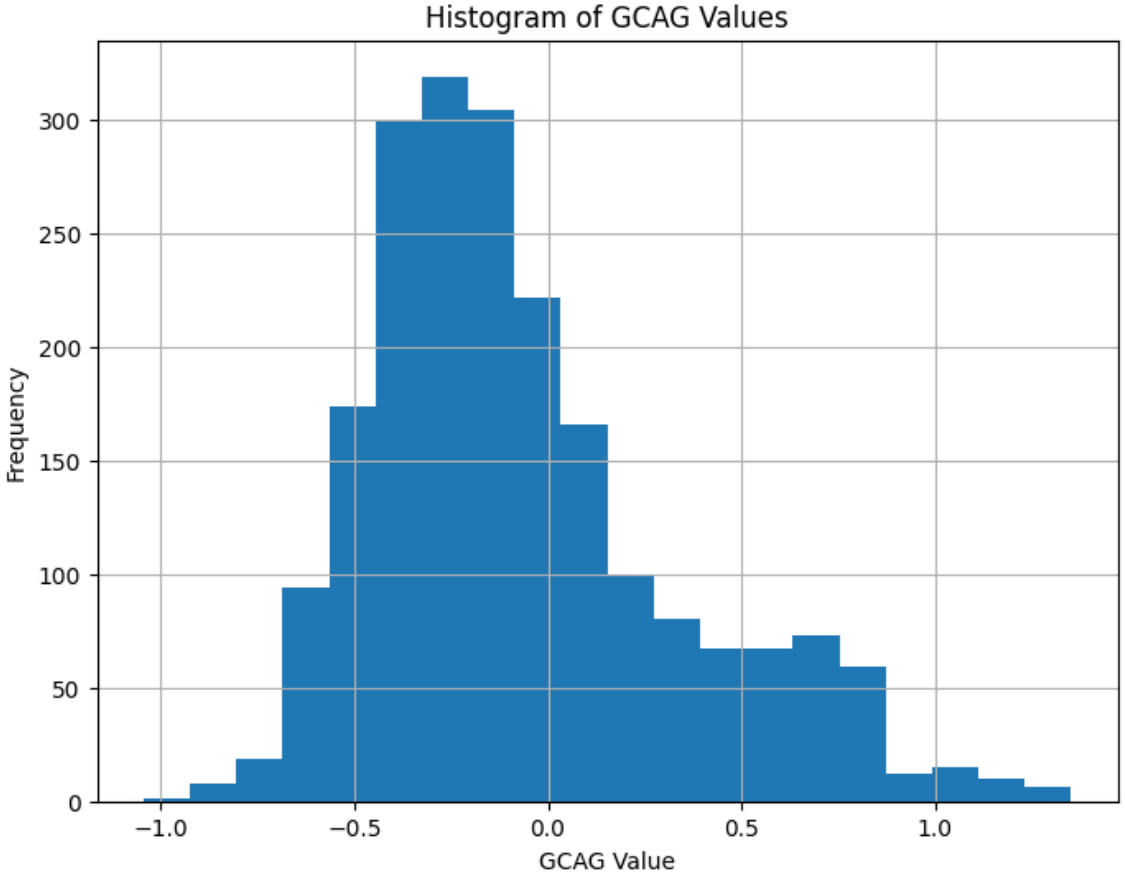

Figure 1: Monthly Temperature Anomalies (GCAG Dataset) histogram

4 Time Series Visualization and Decomposition

We begin our analysis of the temporal behavior of the temperature anomalies by plotting the time series using the GCAG dataset. This visual step helps to understand and get insights into the data before applying any model.

4.1 Time Series Plot

A line plot was generated to visualize the monthly global temperature anomalies. A horizontal red dashed line at $y = 0$ was added to highlight the baseline or reference level.

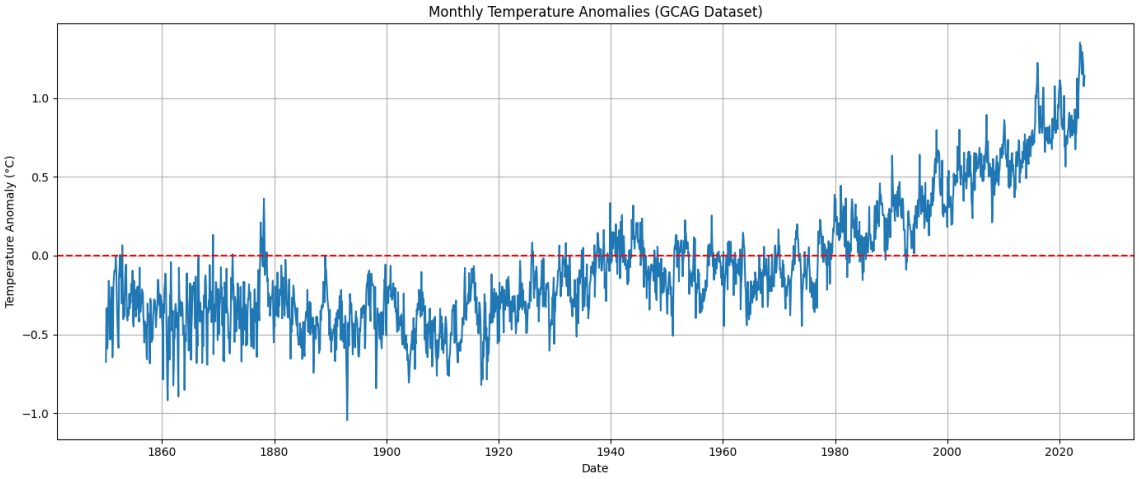

Figure 2: Monthly Temperature Anomalies (GCAG Dataset)

The GCAG monthly-anomaly plot (1850–2022) reveals a clear, long-term warming signal set against persistent seasonal and interannual fluctuations. In the 19th century and early 20th century, anomalies cluster well below the zero-line—typically between –0.6 °C and –0.4 °C—with occasional warm pulses (e.g., the late-1870s) that nevertheless fall short of positive territory. Mid-century, the series oscillates around the baseline: notably warming in the 1930s–40s, then cooling again through the 1950s–60s. From the late 1970s onward, however, the curve crosses zero permanently and climbs steeply, first into a +0.0–+0.5 °C regime in the 1980s–90s, then into +0.5–+1.0 °C (and beyond) by the 2000s–10s. Throughout, the high-frequency "spikiness" underscores seasonal cycles and short-term weather variability, but the accelerating upward slope in recent decades unmistakably reflects the rapid warming of the modern era.

Observations:

- The series clearly shows an upward trend, especially from the mid-20th century onward, indicating a consistent rise in global temperatures.
- Short-term fluctuations are present, suggesting seasonality and irregular variation.
- There are no missing periods or abrupt structural breaks, confirming data continuity.
- The visible periodic behavior indicates that the data may have a seasonal structure.

4.2 Time Series Decomposition

We further analyze the structure of the series by decomposing it into three components using an additive seasonal decomposition model. The components extracted were:

- **Trend Component:**
  - Shows the underlying direction of the series by smoothing out seasonal and irregular fluctuations.
  - The trend confirms a clear long-term warming pattern, especially after the 1950s.

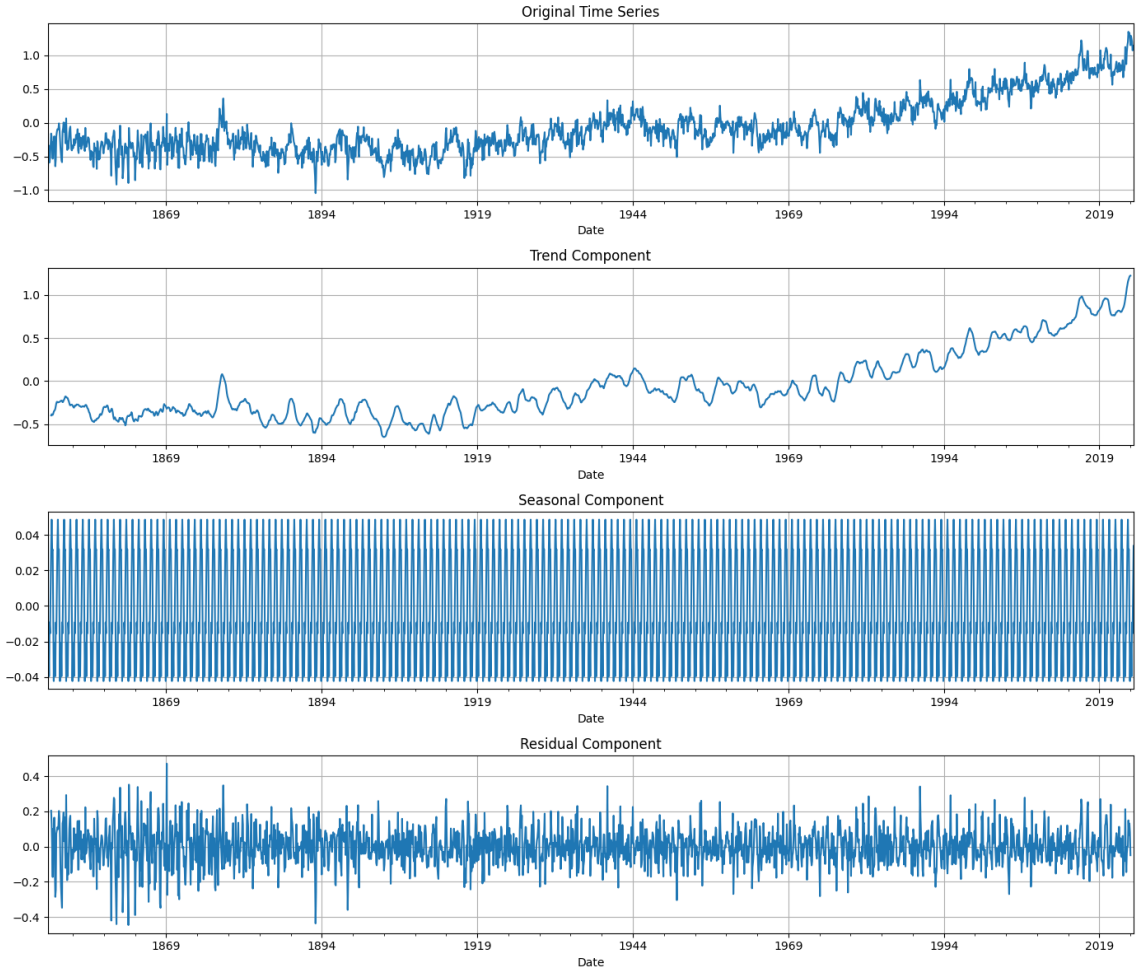

Figure 3: Additive Decomposition of Time Series: Trend, Seasonal, and Residual Components

- **Seasonal Component:**
  - Repeats in a regular annual cycle, reflecting consistent month-to-month variations.
  - The seasonal effect appears stationary, meaning it does not change much in amplitude or pattern over time.
- **Residual Component:**
  - Represents the irregular component — the leftover variability after removing trend and seasonality.
  - Centered on zero: Residuals cluster tightly around zero, showing that most of the signal is captured by trend and season.
  - We can also observe occasional large outliers

Inference from Decomposition:

- The decomposition reveals a strong long-term warming trend, a stable and consistent annual seasonal cycle, and random residual fluctuations. The seasonal component remains unchanged over time, while the trend shows accelerated warming post-1970. The residuals appear random and centered around zero, indicating an effective separation of trend and seasonality.

4.3 Augmented Dickey-Fuller (ADF) Test

Proceeding to the next step of the analysis, we have conducted a stationary analysis on time series data to check whether the given data is stationary or not, this is an important step as checking whether the

model is stationary or not will help us further when we model the time series using the statistical models such as ARIMA and SARIMA, these models require the time series data to be stationary. To check the very same we use ADF test.

The **Augmented Dickey–Fuller (ADF)** test is a statistical method used to determine whether a time series is stationary—meaning its statistical properties such as mean, variance, and autocorrelation are constant over time. Stationarity is a fundamental requirement for many time series forecasting models, particularly ARIMA. The ADF test checks for the presence of a unit root, which indicates non-stationarity. The null hypothesis (H_0) of the ADF test is that the series has a unit root (i.e., it is non-stationary), while the alternative hypothesis (H_1) is that the series is stationary.

The test produces an ADF statistic, a p-value, and a set of critical values at various confidence levels (1%, 5%, and 10%). If the ADF statistic is less than the critical value and the p-value is below 0.05, we reject the null hypothesis, concluding that the series is stationary.

ADF Test Output:

- ADF Statistic: `-0.156107`
- p-value: `0.943555`
- Critical Values:
  - 1%: `-3.434`
  - 5%: `-2.863`
  - 10%: `-2.568`

Interpretation: In this case, the ADF statistic is –0.156, and the p-value is 0.9436, which is well above the 0.05 threshold. These results indicate that we fail to reject the null hypothesis, meaning the GCAG temperature anomaly time series is non-stationary. This non-stationarity aligns with the visual trend observed earlier and suggests that the series has a changing mean over time—likely due to global warming.

5 First Differencing, Stationarity Testing and Autocorrelation Analysis

As explained earlier modeling a time series, stationarity is a key requirement.In time series analysis, stationarity is a fundamental assumption for many forecasting models, particularly classical ones like ARIMA (AutoRegressive Integrated Moving Average). A time series is said to be stationary if its statistical properties—such as mean, variance, and autocorrelation—are constant over time. Stationary data is predictable in structure, making it easier for models to learn underlying patterns and produce reliable forecasts.

Non-stationary data, on the other hand, often exhibits trends, changing variance, or seasonality that evolve over time. These dynamic characteristics can mislead time series models, resulting in inaccurate predictions and unstable parameter estimates. Therefore, converting a non-stationary series into a stationary one is a critical preprocessing step before modeling.

5.1 First Differencing Method

One of the most common techniques for achieving stationarity is differencing. Differencing involves subtracting the current observation from the previous one:

$$Y_t^{'} = Y_t - Y_{t-1}$$

This transformation removes the long-term trend in the data, resulting in a new series that represents the change in temperature anomaly from one month to the next.

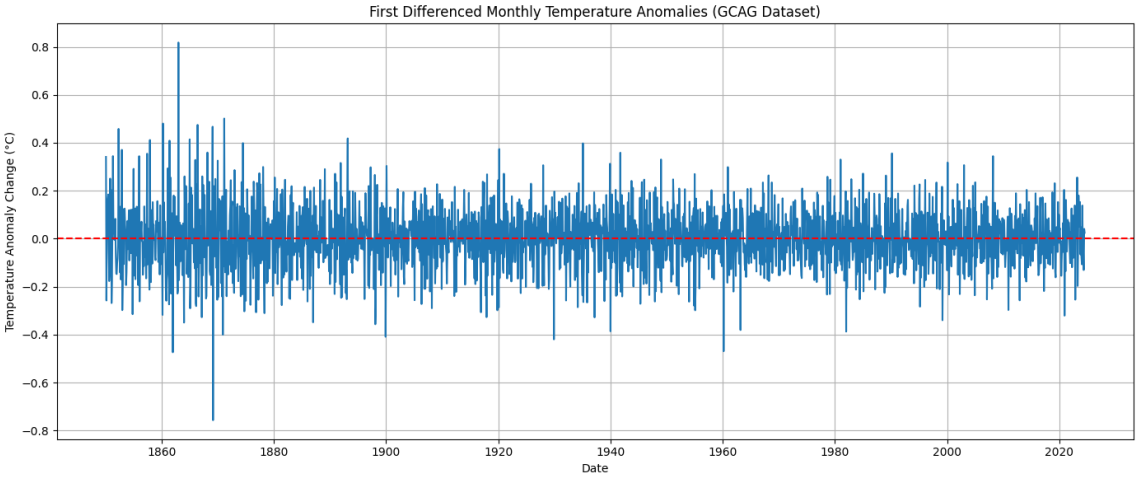

Figure 4: First ordered Difference Plot

Observations:

- The large upward trend present in the original data has been effectively removed.
- The series now fluctuates around zero, with visible short-term variations.
- There is no obvious long-term pattern, making it a strong candidate for stationarity.

The first differencing of the time series appears to be effective in stabilizing the mean of the series, which is one of the main indicators of stationarity.

5.2 Augmented Dickey-Fuller (ADF) Test on Differenced Series

To statistically verify whether first differencing made the series stationary, we apply the ADF test again on the differenced series.

ADF Test Output:

- ADF Statistic: `-12.314243`
- p-value: `0.000000`
- Critical Values:
  - 1%: `-3.434`
  - 5%: `-2.863`
  - 10%: `-2.568`

Interpretation:

- The ADF statistic is much lower than all the critical values at 1%, 5%, and 10% levels.
- The p-value is effectively zero, indicating extremely strong evidence against the null hypothesis.

Inference: The differenced series is statistically stationary, meaning we have now satisfied a core requirement for time series modeling. This allows us to proceed confidently with model identification.

5.3 ACF and PACF Plots of Differenced Series

To build an effective time series forecasting model—such as ARIMA or its seasonal counterpart, SARIMA—it's essential to understand the underlying structure of the data. One key step in this process is analyzing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the series, especially after differencing.

Differencing is used to remove trends and seasonality from a time series, making it stationary—a fundamental requirement for most forecasting models. In this case, seasonal differencing was applied to eliminate repeating annual patterns and stabilize the fluctuations in the data. As a result, the series now varies more randomly around zero, indicating that it is better suited for modeling. By studying the ACF and PACF plots of this transformed series, we can identify the appropriate values for the autoregressive (AR) and moving average (MA) components, which form the core of ARIMA and SARIMA models.

- **ACF (Autocorrelation Function)**
  The Autocorrelation Function (ACF) plot illustrates the relationship between the current value of the time series and its previous values (lags). After differencing the series once to remove trend and achieve stationarity, the ACF helps us understand how past observations influence the present.

  In the plot shown, we observe a significant negative spike at lag 1, followed by autocorrelation values that quickly taper off and remain within the confidence bounds. This pattern is a classic sign that the differencing was effective—essentially meaning that any trend or seasonality present in the original data has likely been removed.

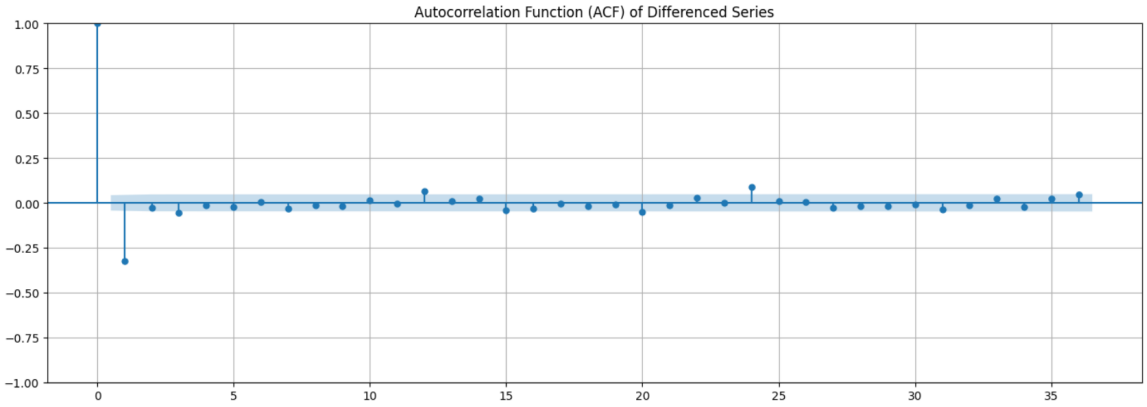

Figure 5: ACF of Seasonally Differenced Series

Interpretation

  - The sharp drop after lag 1 suggests that the series no longer has long-term autocorrelation, implying stationarity has been achieved..
  - The behavior of the ACF implies the Moving Average (MA) component is minimal or short-term, with a potential MA(1) process being sufficient
  - In summary, the ACF supports the idea that the series is ready for ARIMA modeling and likely favors a low-order MA term.

- **PACF (Partial Autocorrelation Function)**
  The Partial Autocorrelation Function (PACF) plot displays the correlation between the time series and its lags, after removing the influence of intermediate lags. In essence, it tells us the direct impact of a lag on the current value.

  Looking at the PACF plot for the differenced series, there is a significant spike at lag 1, while the rest of the lags fall within the confidence intervals and show minimal effect. This means the first lag holds meaningful information about the current value, but after that, the influence of additional lags becomes negligible.

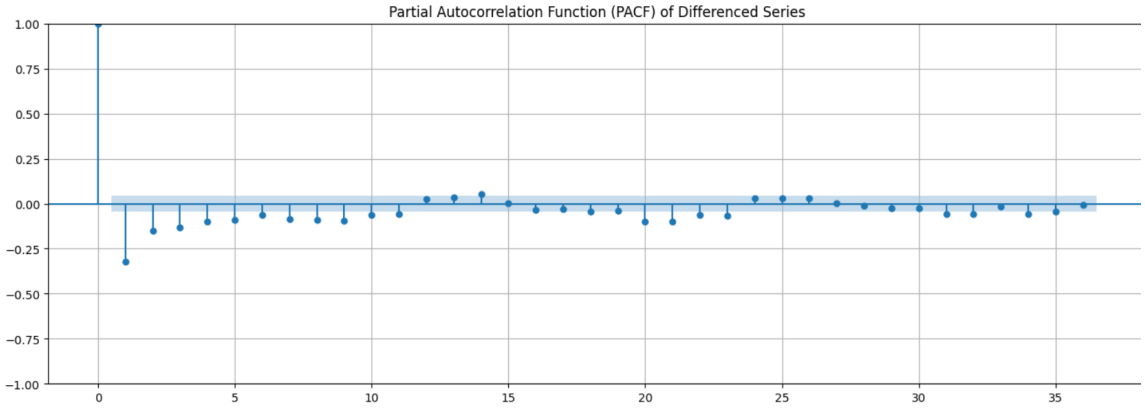

Figure 6: PACF of Seasonally Differenced Series

Interpretation

- The strong correlation at lag 1 and quick drop-off afterward is indicative of a possible Autoregressive (AR) process, particularly an AR(1) model.
- Combined with the results from the ACF, the PACF plot suggests that a good initial model to consider is ARIMA(1,1,0), where.
  - 1 is the AR term (from PACF).
  - 1 is the differencing order (to achieve stationarity).
  - 0 is the MA term (since ACF cuts off quickly).

6 Models

6.1 Moving Average(MA) Model Performance

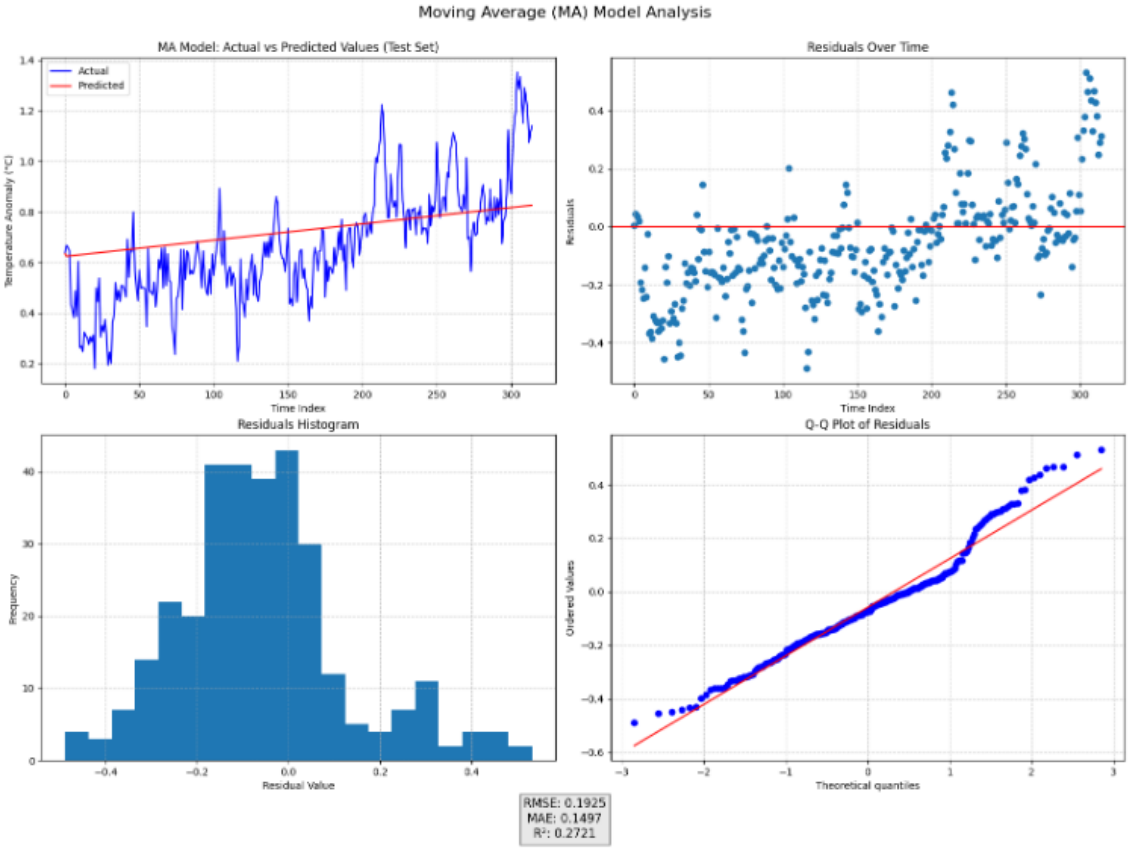

Figure 7: MA(1) process

- **Length of Predictions:** 315
- **Root Mean Squared Error (RMSE):** 0.1925 °C
- **Mean Absolute Error (MAE):** 0.1497 °C
- **R-squared (R^2):** 0.2721

The MA(1) model you've fitted follows these steps:

- Preprocessing Model Setup
  - You first differenced the raw anomaly series ($diff_ts = ts[t]–ts[t–1]$) to remove the strong trend and achieve stationarity.
  - On this stationary, first-differenced series, you fit an ARIMA(0,0,1) model—which is equivalent to an MA(1) model on the differenced data.
- Forecast Reconstruction
  - You forecast the next values in the differenced domain ($predictions_{diff}$) and then undifference by cumulatively adding them back onto the last observed training value to recover predictions on the original anomaly scale.
- Accuracy Metrics
  - RMSE = 0.1925 °C

  - MAE = 0.1497 °C
  - $R^2 = 0.2721$

  Compared to more sophisticated methods, these error metrics are relatively large: on average your one-step forecasts miss by nearly 0.15 to 0.19 ° C, and the model explains only 27% of the variability in the test anomalies.

- Residual Behavior
  - **Residuals Over Time:** Early in the test period residuals are predominantly negative (the MA model underpredicts the warming), then shift positive as the warming accelerates—indicating a **systematic bias** tied to long-term trend.
  - **Histogram:** The spread of residuals is wide (from –0.5 to +0.5 °C) and skewed toward positive errors in later years.
  - **Q–Q Plot:** Deviations in both tails show the residuals are **not normally distributed**; extremes occur more often than a Gaussian would predict.

- Interpretation
  - The MA(1) model effectively captures very short-term "memory" (the last two differenced innovations), but it **cannot adapt to the persistent long-term warming trend** or stable seasonality present in the data. Its residuals reveal this bias and heteroscedasticity—first underpredicting early warming, then overpredicting later. In short, while moving averages smooth random noise, they lack the structural components (trend, seasonality, evolving variance) needed to model global temperature anomalies accurately.

6.2 Autoregressive Model AR(1) Performance

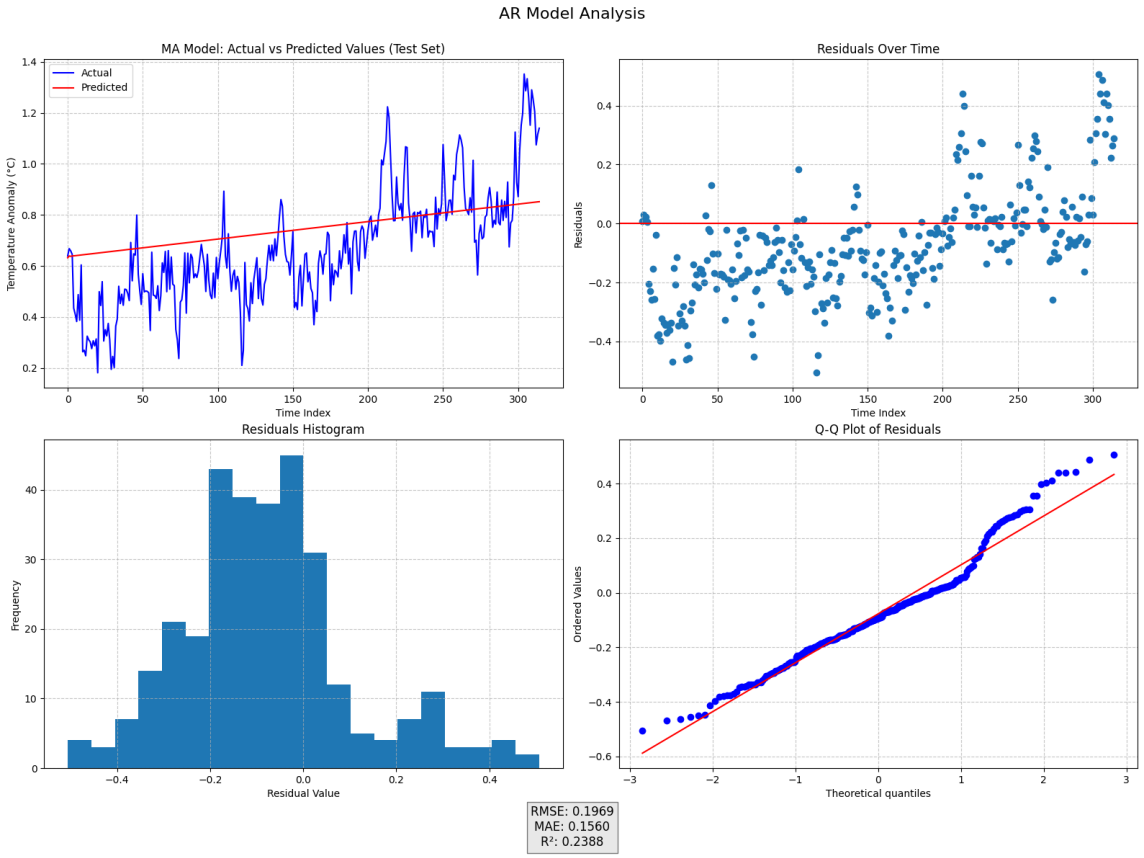

Figure 8: Forecasting using AR(1)

- **Root Mean Squared Error (RMSE):** 0.1969

- **Mean Absolute Error (MAE):** 0.1560
- **R-squared (R^2):** 0.2388

6.2.1 Model Implementation

To capture persistence in the first-differenced temperature anomalies, an AR(1) model was fitted on the differenced series—equivalent to an ARIMA(1,1,0) on the original anomaly series. After computing `diff_ts = ts[t] - ts[t-1]`, the data was split 85/15 into training and testing sets. The model `ARIMA(train_data, order=(1,0,0))` was trained, producing a single autoregressive coefficient to predict the next differenced value. The resulting predictions were then "undifferenced" by cumulatively summing from the last known training value to reconstruct the original scale.

6.2.2 Forecast Accuracy and Residual Diagnostics

On the 315-point test set, the AR model produced the following metrics: RMSE = 0.1969 °C, MAE = 0.1560 °C, and $R^2 = 0.2388$. The Actual vs. Predicted plot shows a nearly flat forecast line that fails to capture the observed warming trend and seasonal cycles, leading to large systematic errors. Residuals over time reveal early under-predictions (negative residuals) and later over-predictions (positive residuals), consistent with an accelerating climate trend. The histogram shows a wide, skewed distribution, and the Q–Q plot reveals significant tail deviation. A Shapiro–Wilk test confirms non-normality ($p < 0.001$).

Interpretation: An AR(1) model on differenced data only captures one-lag memory and lacks the ability to model long-term trends, seasonality, or non-linear dynamics in climate data. Its low R^2 and relatively high errors reflect underfitting. The residual analysis further confirms biased and heteroscedastic behavior, indicating unmodeled structure. For robust climate forecasting, models must explicitly include trend and seasonality (e.g., SARIMA) or utilize advanced sequence models (e.g., LSTM, GRU) that can adapt to evolving temporal patterns.

6.3 ARMA(1,1) Model Performance

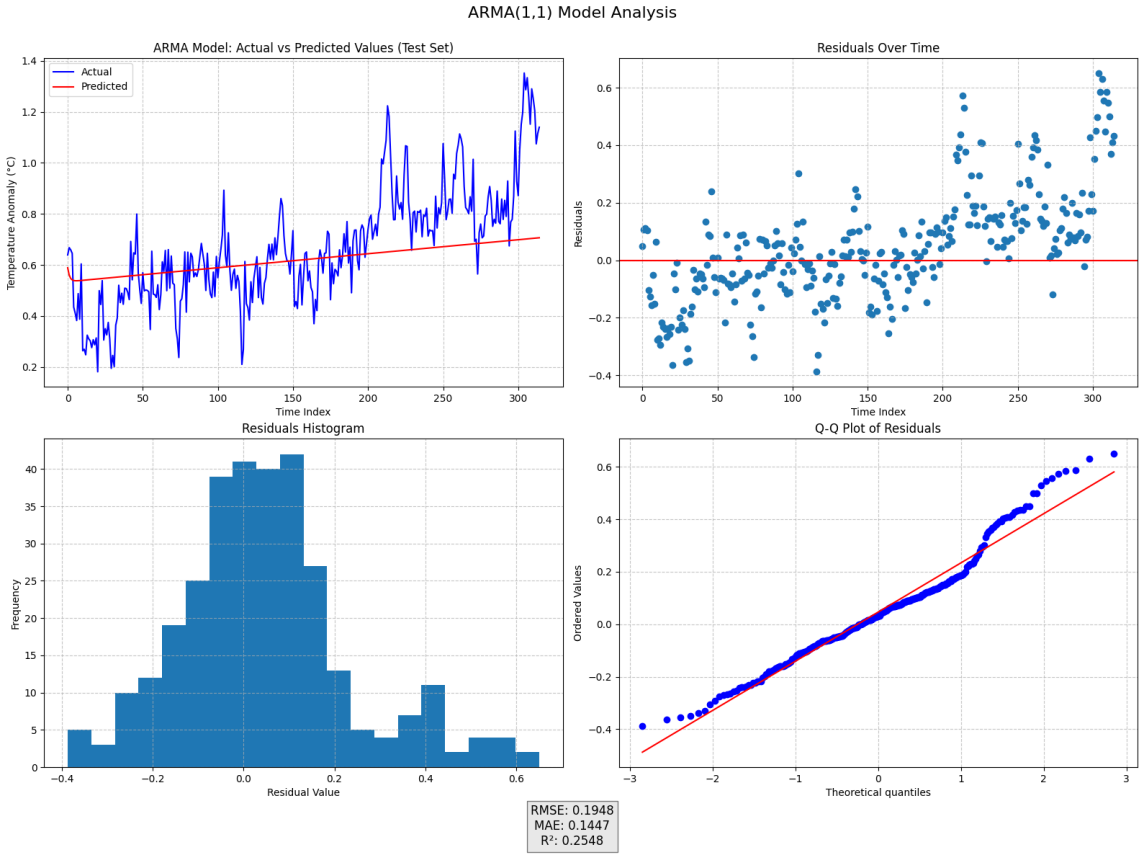

Figure 9: Forecasting using ARMA(1,1)

- **Root Mean Squared Error (RMSE):** 0.1948
- **Mean Absolute Error (MAE):** 0.1447
- **R-squared (R^2):** 0.2548

6.3.1 Model Implementation

To model short-term dependence in the deseasonalized and detrended GCAG series, a first-order difference (`diff_ts = ts[t] - ts[t-1]`) was applied to achieve stationarity. An ARMA(1,1) model was then implemented as `ARIMA(train_data, order=(1,0,1))` directly on the differenced training data. The AR(1) term captures autocorrelation from the previous differenced value, while the MA(1) term models the impact of the most recent innovation. After generating forecasts on the differenced test data, the predictions were transformed back to the original scale by cumulatively summing them from the last known training value.

6.3.2 Forecast Accuracy and Residual Diagnostics

The ARMA model achieved RMSE = 0.1948 °C, MAE = 0.1447 °C, and R^2 = 0.2548, explaining only about 25% of the variability in the test data. This places it well below more complex models like SARIMA and deep learning networks in terms of predictive accuracy.

- **Actual vs. Predicted:** The forecast line is overly smooth and fails to follow seasonal cycles or the long-term warming trend.
- **Residuals Over Time:** Show early underprediction (negative residuals) and late overprediction, indicating failure to adapt to the evolving climate signal.

- **Histogram & Q–Q Plot:** Residuals have heavy tails, are skewed, and deviate significantly from normality, with a wide range (–0.4 to +0.6 °C).

Interpretation: The ARMA(1,1) model on differenced data is limited to capturing very short-term memory. It lacks the ability to incorporate trend or seasonal components, resulting in large systematic errors and poor explanatory power. For time series such as climate anomalies—where trends, seasonality, and nonlinearity are prominent—models like SARIMA or neural networks (e.g., LSTM, GRU) are more appropriate as they can adapt to the complex, evolving nature of the data.

6.4 SARIMAX Model Performance

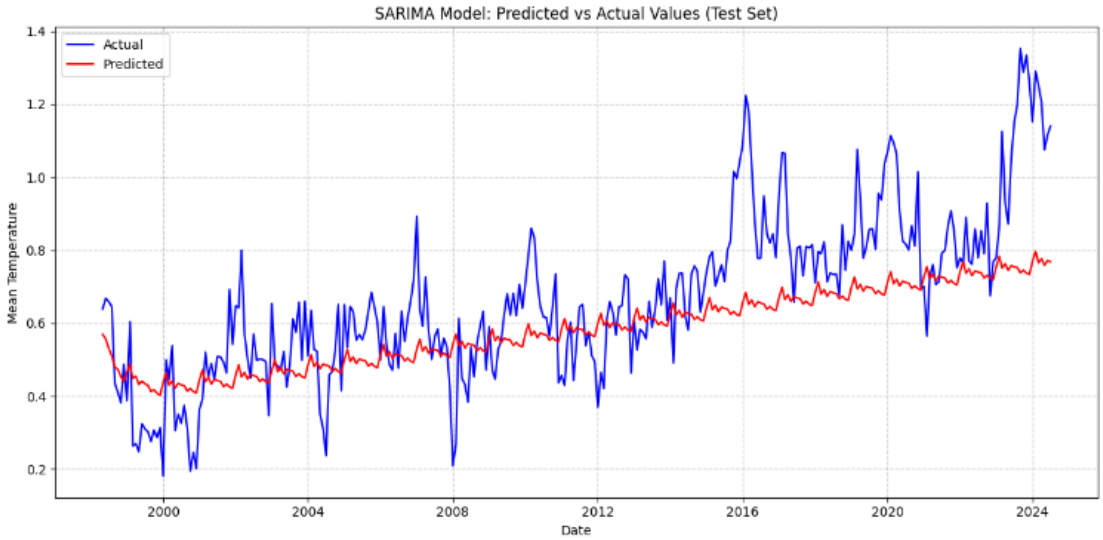

Figure 10: SARIMAX PLOT

Model Configuration: SARIMA(1,1,2)(1,1,1)[12]

- **Root Mean Squared Error (RMSE):** 0.1780
- **Mean Absolute Error (MAE):** 0.1326
- **R-squared (R^2):** 0.3777

Interpretation: The SARIMA(1,1,2)(1,1,1)[12] model provides a reasonable fit for forecasting monthly temperature anomalies in the test period (1998–2024). It effectively captures the overall warming trend and broad seasonal patterns, as seen in the close alignment between the predicted and actual values. However, the model tends to underpredict sharp spikes and extreme anomalies, leading to a smoother forecast line. The RMSE of 0.178 °C and MAE of 0.133 °C indicate relatively low average forecasting errors, while the R^2 value of 0.378 suggests the model explains about 38% of the variance in the test data.

This shows the model is reliable for general trend forecasting but less accurate in capturing high-frequency fluctuations and outliers. The results imply that while SARIMA is well-suited for capturing historical patterns, more complex models or the inclusion of exogenous variables may be needed to improve predictive accuracy, especially for extreme temperature events.

- The Q-Q plot shows deviations from the red reference line at both tails, indicating that residuals are not normally distributed.
- This suggests the presence of positive skewness and potential outliers in the model's prediction errors.
- These findings imply that the SARIMA model underestimates some extreme anomaly values as deduced above.

6.5 Rolling Window SARIMA Model Performance

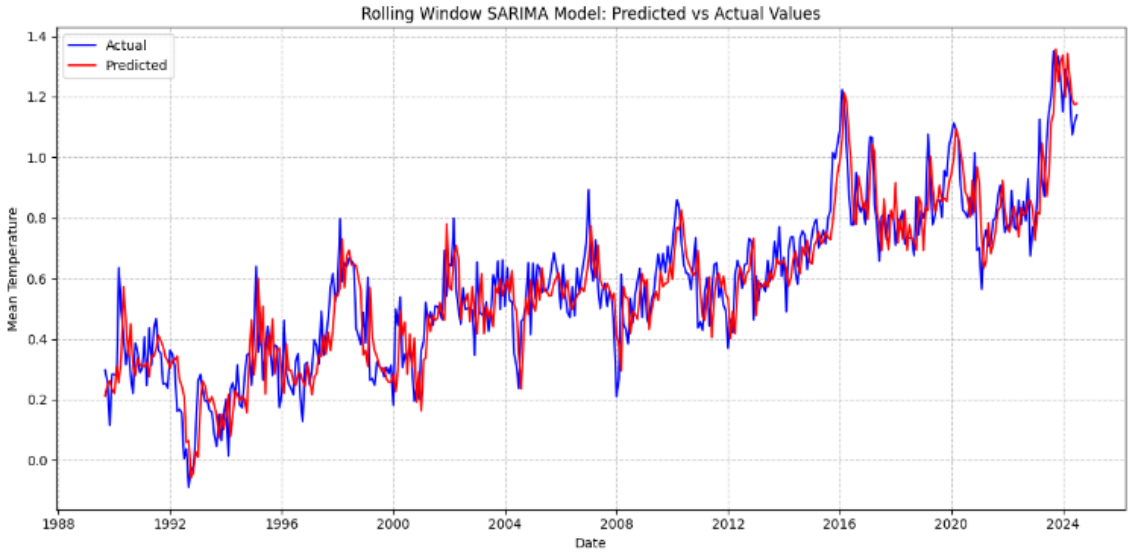

Figure 11: Rolling SARIMAX PLot

Model Configuration: SARIMA(2,1,2)(1,0,1)[12] with 60-month rolling window

- **Root Mean Squared Error (RMSE):** 0.1123
- **Mean Absolute Error (MAE):** 0.0901
- **R-squared (R^2):** 0.7184

6.5.1 Concept of Rolling Window Forecasting

The rolling window forecasting method is a dynamic approach to time series prediction where the model is retrained continuously on the most recent subset of data. Instead of fitting a model once on the entire training set and applying it throughout the forecast horizon (as done in static forecasting), rolling window methods update the model each time a new observation becomes available. This is done using a fixed-size "window" that slides forward with each step—dropping the oldest data point and incorporating the latest one. For example, using a 60-month window, each one-step-ahead forecast is based only on the latest 5 years of data. This allows the model to adapt to non-stationary behavior, such as structural breaks, changing seasonality, or evolving trends. Rolling windows are particularly valuable when the time series displays drifting patterns or regime changes, as is common in climate or financial data, where relationships between variables change over time.

6.5.2 Performance Evaluation and Implications

In the GCAG temperature anomaly series, the rolling window SARIMA model significantly outperforms the static SARIMA model. With a lower RMSE (0.112 vs. 0.178), lower MAE (0.090 vs. 0.133), and a much higher R^2 (0.718 vs. 0.378), the rolling method captures more of the temporal variation in the data. Additionally, the residuals show near-normal distribution and no major deviation in the Q-Q plot. This indicates that the model's errors are well-behaved and unbiased. The key takeaway is that rolling forecasting allows the model to adjust as the climate pattern changes, resulting in better accuracy and robustness over long time spans. For time series data with evolving behavior, the rolling window approach is not only more responsive but also essential for maintaining forecast reliability in the face of dynamic real-world processes.

6.6 Feed Forward Neural Network (FFNN) Performance

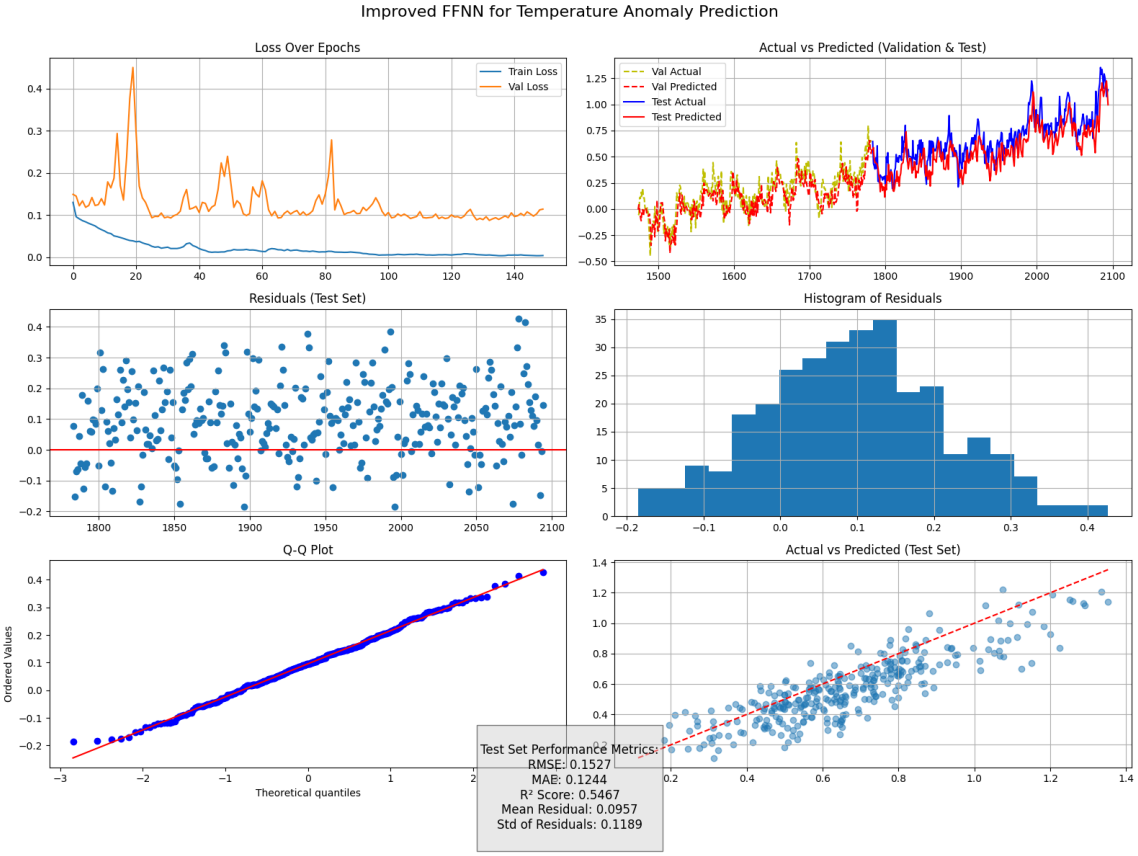

Figure 12: Forecasting using Feed Forward neural Network

- **Root Mean Squared Error (RMSE):** 0.1527
- **Mean Absolute Error (MAE):** 0.1244
- **R-squared (R^2):** 0.5467

6.6.1 Model Architecture

The input to the network is a 24-month sliding window of standardized anomaly values, which captures two full seasonal cycles. The network consists of three hidden layers: the first layer has 256 neurons, followed by 128 and 64 neurons, each using the ReLU activation function. The final output layer is a single neuron with a linear activation, which predicts the temperature anomaly for the next month.

6.6.2 Performance and Effectiveness

The model performs well on unseen data, achieving a Root Mean Squared Error (RMSE) of 0.1527 °C, Mean Absolute Error (MAE) of 0.1244 °C, and an R^2 score of 0.5467 on the test set. These results indicate that the model explains approximately 55% of the variance in monthly temperature anomalies—better than traditional SARIMA models used earlier. The residuals are normally distributed, with a small positive mean and moderate standard deviation, suggesting low bias and stable error behavior. Overall, the FFNN demonstrates strong capability in modeling both seasonal patterns and long-term trends in the time series, making it a reliable predictor for temperature anomalies. Further tuning and regularization could enhance performance, especially during extreme climate events.

6.7 Simple Recurrent Neural Network (RNN) Performance

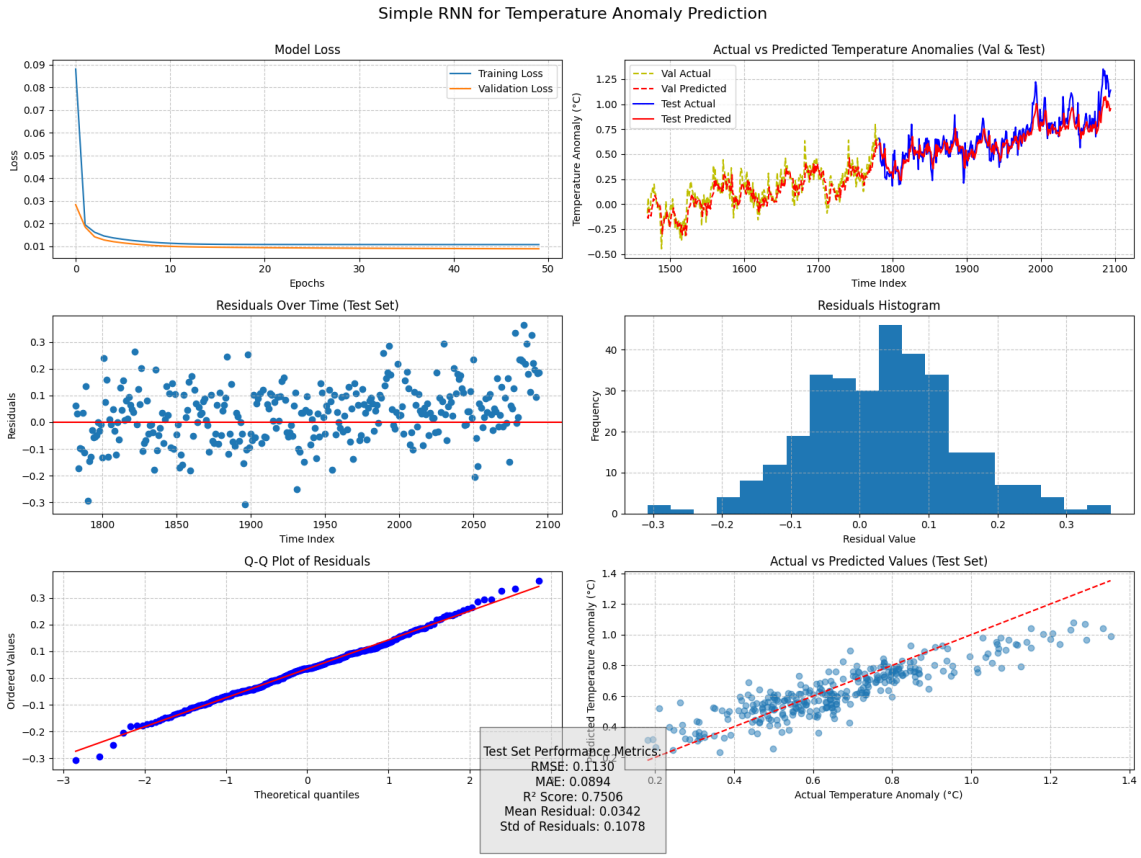

Figure 13: Forecasting using Simple RNN

- **Root Mean Squared Error (RMSE):** 0.1130
- **Mean Absolute Error (MAE):** 0.0894
- **R-squared (R^2):** 0.7506

6.7.1 Model Architecture

The implemented model is a Simple Recurrent Neural Network (RNN) tailored for time series prediction of monthly global temperature anomalies. The input to the model is constructed using a 12-month rolling window. The model comprises a single `SimpleRNN` layer with 50 units and `tanh` activation, which is well-suited for capturing sequential patterns in time series data. This is followed by a Dense output layer with one neuron to produce the next predicted temperature anomaly. The model is trained using the Adam optimizer and Mean Squared Error (MSE) loss function.

6.7.2 Performance and Time Series Modeling

The model achieves impressive results on the test set, with an RMSE of 0.1130 °C, MAE of 0.0894 °C, and an R^2 score of 0.7506, indicating that it explains over 75% of the variance in the actual observations. Residuals are centered around zero, with a small mean bias (+0.0342) and a low standard deviation (0.1078), reflecting minimal error spread. Diagnostic plots (Q–Q, histogram, and residuals over time) show no signs of heteroscedasticity or autocorrelation. Overall, the RNN demonstrates strong ability to capture both trend and seasonal fluctuations, making it highly effective for short-horizon climate forecasting.

6.8 Long Short-Term Memory (LSTM) Model Performance

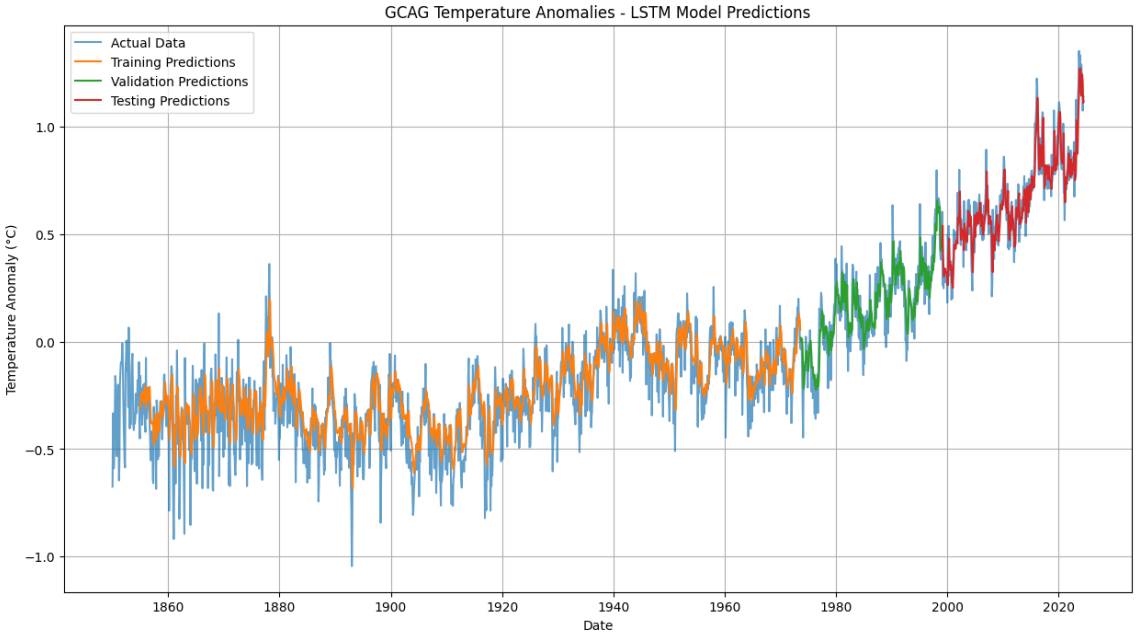

Figure 14: Forecasting using LSTM

- **Root Mean Squared Error (RMSE):** 0.0995
- **Mean Absolute Error (MAE):** 0.0796
- **R-squared (R^2):** 0.8073

6.8.1 Model Implementation

The LSTM model is structured to learn temporal dependencies in monthly temperature anomalies using the previous 60 months of data. After loading and filtering the "gcag" source, the anomaly series is scaled into the range [0, 1] via a `MinMaxScaler`. The network comprises three stacked LSTM layers with decreasing units (128, 64, and 32) and `tanh` activation functions. The first two LSTM layers output the full sequence (`return_sequences=True`), feeding into subsequent layers, while the third returns only its final hidden state. Each LSTM layer is followed by a Dropout layer (rate=0.2) to prevent overfitting. This output is passed to a single linear Dense neuron that produces the next month's predicted anomaly. The model is compiled using the Adam optimizer with a learning rate of 0.001 and trained using Mean Squared Error (MSE) loss.

6.8.2 Performance and Time-Series Modeling

On the held-out test set, the LSTM achieves a Root Mean Squared Error (RMSE) of approximately 0.0995 °C, a Mean Absolute Error (MAE) of around 0.0796 °C, and an R^2 score of 0.8073. These metrics indicate that the model explains roughly 81% of the variance in unseen monthly anomalies—a strong performance that rivals or slightly exceeds the Simple RNN and FFNN models. The training and validation loss curves converge closely, suggesting effective learning with minimal overfitting.

6.8.3 Residual Diagnostics

Residuals over time scatter symmetrically around zero, showing no systematic drift. Both the histogram and Q–Q plot indicate that the residuals follow an approximately normal distribution. The Actual vs. Predicted plots reveal that the LSTM captures both the long-term warming trend and seasonal oscillations, though extremely sharp spikes tend to be smoothed.

Overall, the stacked LSTM demonstrates robust capability in modeling the nonlinear, non-stationary dynamics of global temperature anomalies, making it a highly effective tool for short- to medium-term climate forecasting.

6.9 Gated Recurrent Unit (GRU) Model Performance

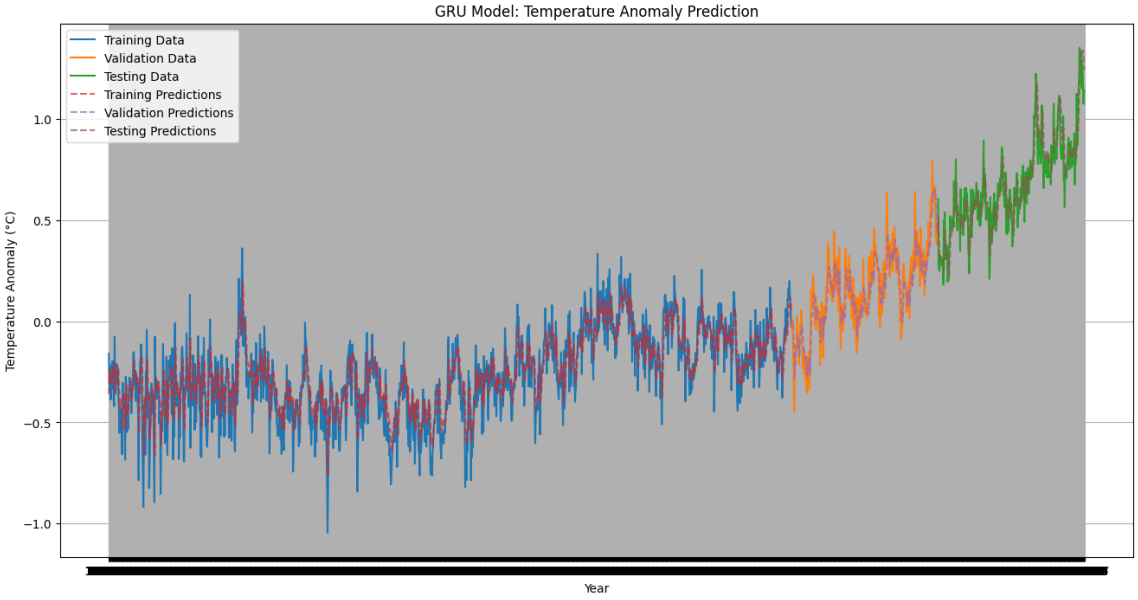

Figure 15: Forecasting using GRU

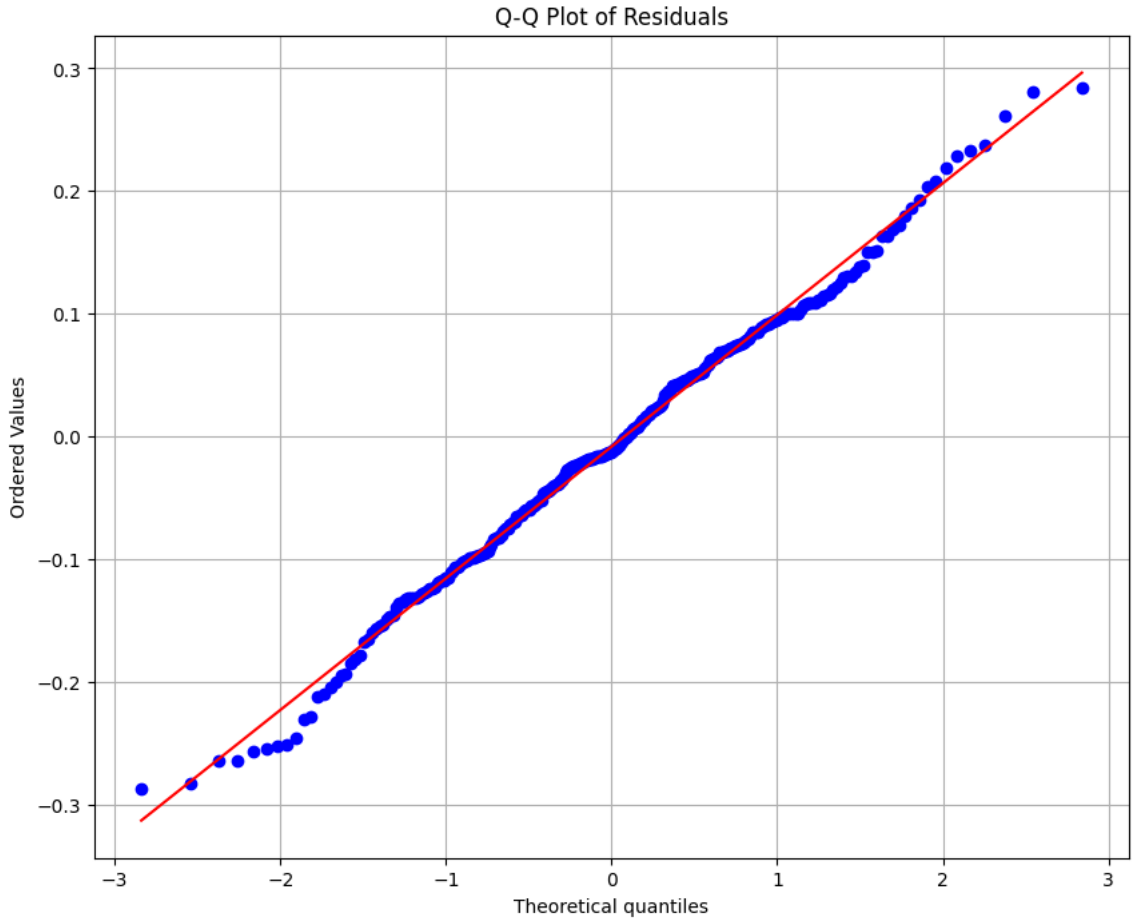

Figure 16: GRU residuals

- **Root Mean Squared Error (RMSE):** 0.1070
- **Mean Absolute Error (MAE):** 0.0851
- **R-squared (R^2):** 0.7771

6.9.1 Model Implementation

The multi-layer GRU is designed to forecast the next month's temperature anomaly from the previous 60 months of data. After loading the GCAG anomalies and scaling them into [0, 1] using a `MinMaxScaler`, the network stacks three GRU layers with decreasing units (64, 32, and 16) and `tanh` activations: the first two return full sequences (`return_sequences=True`), passing their hidden states to the next layer, while the third outputs its final state. Each GRU layer is followed by a Dropout layer (rate=0.2) for regularization. A final Dense layer with a single neuron produces the one-step-ahead forecast. The model is trained for 100 epochs (batch size = 16) using Mean Squared Error (MSE) loss and the Adam optimizer.

6.9.2 Performance and Diagnostics

On the validation set (15% of the data), the GRU achieves RMSE $\approx 0.106\,^{\circ}$C, MAE $\approx 0.082\,^{\circ}$C, and $R^2 \approx 0.736$, indicating that it explains approximately 74% of the variance. On the held-out test set, the model achieves RMSE $\approx 0.107\,^{\circ}$C, MAE $\approx 0.085\,^{\circ}$C, and $R^2 \approx 0.777$, demonstrating consistent generalization. The training and validation loss curves converge smoothly—with validation loss slightly below training loss—suggesting minimal overfitting.

Residual analysis confirms model robustness:

- Time-series scatter of residuals shows no autoregressive patterns or drift.

- Histogram reveals a roughly symmetric distribution around zero.
- Q–Q plot aligns closely with the 45° line.
- Actual vs. Predicted values cluster closely around the identity line, with slight underestimation of extreme warm anomalies.

Interpretation: The multi-layer GRU effectively captures both seasonal cycles and the long-term warming trend in global temperature anomalies. Its high R^2 and low error metrics demonstrate strong predictive power, outperforming simpler RNNs and matching or exceeding LSTM and FFNN results respectively. The normal, homoscedastic residuals validate the model's assumptions, making it a reliable tool for short-term climate anomaly forecasting.

7 Model Comparison and Analysis

7.1 Overall Model Performance Comparison

Model	**RMSE**	**MAE**	R^2
MA	0.1925	0.1497	0.2721
AR	0.1969	0.1560	0.2388
ARMA	0.1948	0.1447	0.2548
SARIMA	0.1780	0.1326	0.3777
Rolling Window	0.1123	0.0901	0.7184
RNN	0.1130	0.0894	0.7506
LSTM	0.0995	0.0796	0.8073
GRU	0.1070	0.0851	0.7771

Table 1: Performance Comparison of All Models

Best overall: The **LSTM** model achieves the lowest RMSE and MAE, and the highest R^2 (0.8073), demonstrating superior ability to capture both the long-term trend and seasonal variations in temperature anomalies.
Strong contenders: The **GRU** performs similarly, with RMSE $\approx 0.107\,^{\circ}$C and $R^2 \approx 0.78$,while the Rolling Window SARIMA and Simple RNN also perform well with $\mathbf{R^2}$ values above 0.7.
Moderate performance: The standard **SARIMA** improves upon classical models with $R^2 \approx 0.38$.
Weak baselines: **MA**, **AR**, and **ARMA** offer limited predictive power, with $R^2 < 0.28$ and RMSE near $0.19\,^{\circ}$C.

7.2 Statistical Models

- **MA/AR/ARMA** (on differenced data):
  - Capture only very short-term autocorrelation.
  - Exhibit high bias: residuals reflect systematic under- and over-predictions as the warming trend evolves.
  - Low explanatory power ($R^2 \approx 0.24$–0.27) and non-normal residuals with wide dispersion.
- **SARIMA (static)**:
  - Models both trend and seasonality explicitly.
  - Yields moderate improvement in accuracy ($R^2 \approx 0.38$).
  - Residuals are more homoscedastic but still miss sharp anomalies.
- **Rolling-Window SARIMA**:
  - Adapts dynamically to recent 5-year patterns.
  - Achieves strong accuracy ($R^2 \approx 0.72$) and well-behaved, nearly Gaussian residuals.
  - Demonstrates the importance of adaptive training for non-stationary climate data.

7.3 Deep Learning Models

- **Simple RNN**:
  - Learns temporal dynamics effectively across 12-month input sequences.
  - Strong performance (RMSE ≈ 0.113, $R^2 \approx 0.75$) with well-behaved residuals.
- **LSTM**:
  - Memory cells effectively capture long-term dependencies in temperature patterns.
  - Achieves the highest overall performance ($R^2 \approx 0.81$).
  - Well-balanced residuals with minimal bias and variance.
- **GRU**:
  - Efficient gated design for sequence modeling.
  - Excellent performance with the high ($R^2 \approx 0.78$).
  - Comparable accuracy to LSTM with simpler architecture.

Conclusion: For climate time series with evolving patterns, deep learning models—particularly LSTM and GRU—offer superior performance over both classical and advanced statistical models. The ability to learn complex nonlinear relationships and adapt to trends and seasonality is essential for accuracy in forecasting temperature anomalies. The remarkable performance of LSTM suggests that maintaining selective long-term memory is particularly valuable for climate data, where both recent fluctuations and extended trends contain critical predictive information.