Department of Mathematics, IIT Bombay

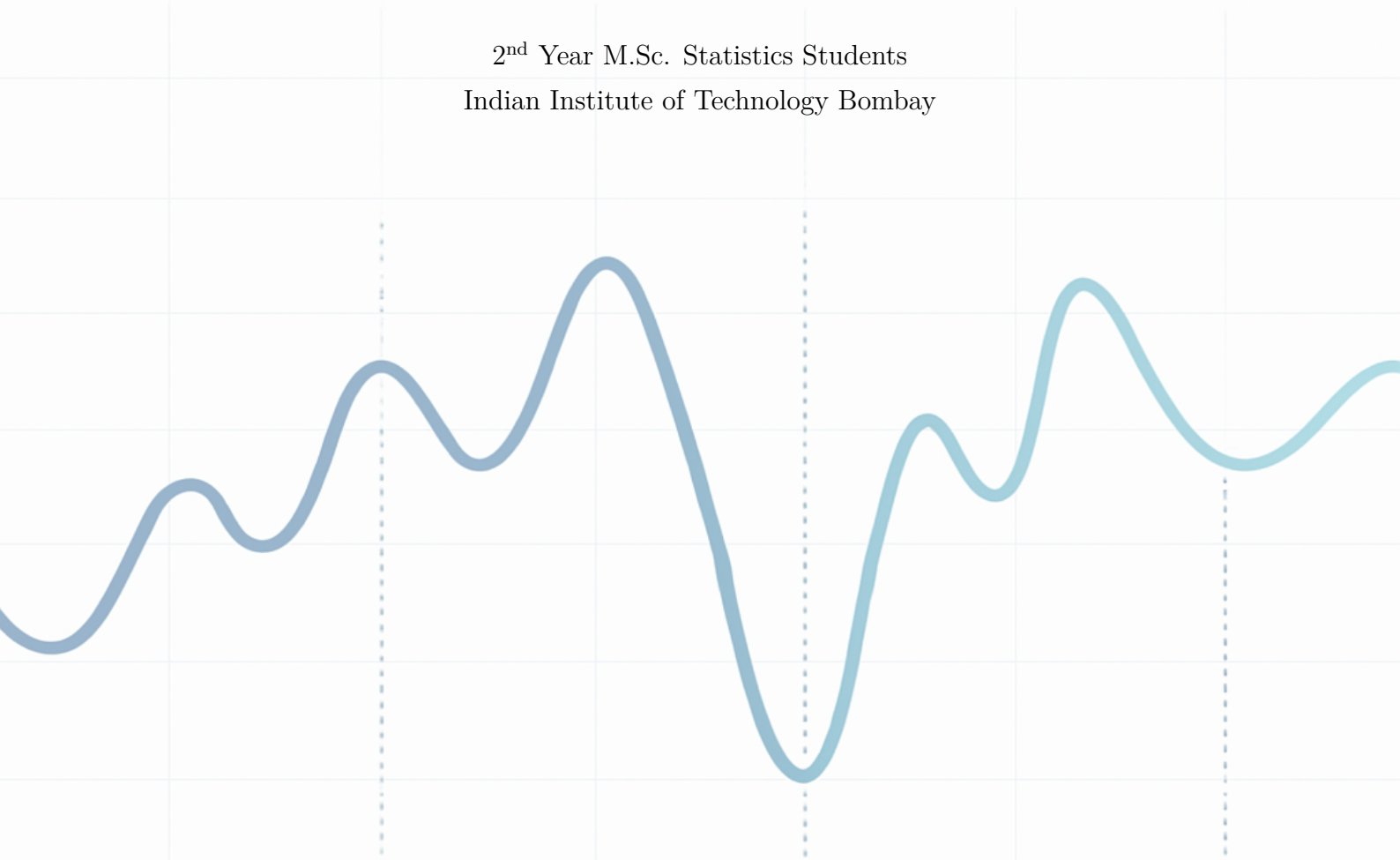
Change Point Detection and Time Series Analysis

A guided project under Prof. Monika Bhattacharjee

Team Members

Amit Patange — 24N0046
Sarthak Game — 24N0047
Tathagat Pandey — 24N0070

2nd Year M.Sc. Statistics Students
Indian Institute of Technology Bombay



Acknowledgement

We are deeply grateful to **Prof. Monika Bhattacharjee** for her invaluable guidance and mentorship throughout the course of this project. Her unwavering support, insightful suggestions, and constructive feedback were instrumental in shaping the direction and depth of our work.

We especially thank her for guiding and helping us understand the fundamental concepts of **Change Point Detection**, and for introducing us to various methodologies to address the challenges we encountered during the project. Her patience and willingness to engage in detailed discussions enabled us to gain clarity and move forward with confidence at every stage.

This project would not have been possible without her continued encouragement, academic insight, and generous investment of time. We feel fortunate to have worked under her supervision.

Introduction

Understanding structural changes in time series data is critical for generating meaningful insights and reliable forecasts. In long-term observations, particularly those spanning multiple decades or centuries, the underlying dynamics of the process often evolve. These shifts—known as change points—mark locations in time where the statistical properties of the series (such as mean, variance, or trend) undergo significant alteration. Failing to account for these changes can lead to misleading models and inaccurate predictions.

In many real-world scenarios, including environmental, economic, and social data, early portions of a time series may behave differently than more recent ones. For instance, initial observations may lack a strong directional trend, while later segments exhibit systematic behavior due to technological, behavioral, or natural changes. Treating such a series as a single homogeneous process can obscure valuable information. Therefore, segmenting the series into structurally consistent intervals is a crucial first step in effective time series modeling.

This study investigates a variety of Change Point Detection (CPD) techniques to identify significant transitions in the structure of time series data. We implement Binary Segmentation, a recursive partitioning method; PELT (Pruned Exact Linear Time), known for its optimal detection in linear time; and CUSUM (Cumulative Sum), which is sensitive to shifts in the mean. In addition, we incorporate Bottom-Up Segmentation, a hierarchical method that starts with fine-grained partitions and merges segments based on similarity, and Window-Based approaches, which scan the data locally for potential structural breaks. These methods are complementary in their sensitivity to different types of changes and provide a rich foundation for analyzing temporal dynamics.

Once change points are identified, the segmented time series is treated as a collection of shorter, more homogeneous intervals. Instead of building a global model, we adopt a time-weighted ensemble approach where each segment contributes to the forecast based on its temporal relevance. More recent segments are assigned higher weights, allowing the model to emphasize recent patterns while still incorporating long-term trends. This ensemble strategy not only enhances forecast accuracy but also adapts to nonstationarities and structural shifts in the underlying process.

In summary, this work highlights the importance of structural analysis in time series forecasting. By leveraging robust change point detection methods and integrating them into a weighted modeling framework, we aim to improve interpretability, adaptiveness, and predictive performance—particularly in long, evolving datasets where temporal context is key.

Background

Time series analysis has long served as a cornerstone in understanding how values evolve over time, especially in fields such as economics, environmental studies, and signal processing. Foundational models like ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal ARIMA) have been widely used due to their ability to capture linear dependencies and seasonal structures in time series data. These models assume stationarity or allow for transformation into a stationary process, making them well-suited for many classical forecasting applications.

However, real-world data often exhibit structural changes — sudden shifts in the underlying distribution caused by events such as policy changes, technological interventions, or, in the case of climate studies, major environmental shifts. These changes challenge the assumptions of traditional time series models and call for techniques that can explicitly detect such shifts.

This is where Change Point Detection (CPD) becomes a valuable analytical tool. CPD aims to identify locations in the time series where the statistical properties — such as mean, variance, or correlation structure — undergo significant changes. Among the earliest methods developed for this purpose is the Cumulative Sum (CUSUM) technique introduced by E. S. Page in 1954, which monitors cumulative deviations from a target value. Binary Segmentation, another classical approach, recursively identifies multiple change points by dividing the data into segments. The Bottom-Up segmentation strategy starts by considering the finest granularity and iteratively merges segments to minimize a given cost. More recent and efficient techniques such as PELT (Pruned Exact Linear Time), introduced by Killick et al., allow for optimal detection of multiple change points with linear computational complexity under certain conditions.

These techniques are particularly useful in analyzing long-term climate data, which span over a century and are susceptible to natural variability, measurement adjustments, and true shifts in environmental patterns. By identifying and understanding structural changes in such datasets, analysts can partition the series into more homogeneous segments, allowing for more reliable modeling and forecasting. This segmentation not only improves statistical inference but also enhances interpretability — enabling analysts to associate each segment with meaningful historical or environmental events.

Overall, the combination of traditional time series modeling and modern change point detection provides a powerful framework to understand complex temporal dynamics, especially in the context of long-spanning datasets like global temperature records.

Data Source and Structure

The dataset utilized in this analysis represents a comprehensive reconstruction of Earth's surface temperature anomalies spanning more than 170 years, from the mid-19th century to the present. This long historical range allows for the exploration of climate variability across multiple time scales, including decadal, multi-decadal, and centennial trends. The data integrates land and oceanic temperature records, offering a global perspective on how surface temperatures have evolved over time.

This temperature dataset is produced by combining two authoritative sources: the **Berkeley Earth land-surface temperature dataset**, which aggregates temperature records from thousands of weather stations globally, and the **HadSST4 dataset**, which provides sea surface temperature estimates derived from historical ship-based measurements, buoys, and satellites. The two are seamlessly merged to form a monthly time series of global surface temperature anomalies.

Temperature values are reported as **monthly anomalies in degrees Celsius (°C)**, representing deviations from a defined climatological baseline, specifically the average temperature during the period **January 1951 to December 1980**. This baseline is widely adopted in climate science to normalize historical trends and reduce seasonal bias. The anomaly-based reporting provides a consistent framework for comparing temperatures across centuries and locations, regardless of absolute differences in regional climate.

Special attention is paid to areas covered by sea ice. In this study, the dataset version that extrapolates values using **air temperatures above sea ice** is employed. This approach is considered more reflective of climatic conditions and avoids the underestimation of variability that may occur if using ocean water temperatures beneath the ice.

Structure of the Dataset:

Column Name	Description
`Year`	Indicates the year of observation. The data begins in the 1850s and extends through the current decade.
`Month`	Represents the month of the observation, ranging from 1 (January) to 12 (December).
`Monthly_Anomaly`	The temperature anomaly in °C, calculated relative to the 1951–1980 global average for that specific month. This metric reflects the deviation from expected climatological norms.

The clean and uniform structure of this dataset makes it particularly well-suited for time series analysis. Its monthly frequency allows for both high-resolution seasonal assessments and long-term trend evaluations. Additionally, the anomaly format helps mitigate the influence of geographical and seasonal variation, which is critical for robust applications such as change point detection, trend forecasting, and statistical decomposition.

Methodology

The methodological framework adopted in this study integrates both classical time series analysis and modern change point detection (CPD) techniques. This hybrid strategy allows for robust forecasting while accounting for non-stationarities and structural changes in the data, particularly relevant given the temporal span of over 150 years. In this section, we discuss each component in detail and explain how they are synthesized in the overall analytical pipeline.

0.1 Classical Time Series Models

Classical models such as ARIMA (AutoRegressive Integrated Moving Average) and its seasonal extension SARIMA (Seasonal ARIMA) are widely employed for time series forecasting. These models assume that the underlying data is stationary or can be made stationary through differencing.

An ARIMA(p, d, q) model is mathematically represented as:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t, \tag{1}$$

where B is the backshift operator, $\phi(B)$ and $\theta(B)$ are the polynomials of order p and q, respectively, and ε_t is white noise. The model accounts for autocorrelation and moving average patterns in the residuals and captures the intrinsic memory structure of the series.

For seasonal data, the SARIMA$(p, d, q)(P, D, Q)_s$ model is used:

$$\Phi(B^s)\phi(B)(1-B)^d(1-B^s)^D y_t = \Theta(B^s)\theta(B)\varepsilon_t, \tag{2}$$

where s denotes the seasonal period, and capitalized polynomials (Φ, Θ) handle seasonal terms.

These models form the baseline forecasting tools. However, due to their assumption of consistent statistical structure over time, they often perform poorly when the series exhibits abrupt changes in level, trend, or variance. This limitation motivates the use of change point detection techniques.

0.2 Change Point Detection (CPD)

Change point detection refers to identifying points in time where the statistical properties of a sequence change. CPD is crucial for long time series data, especially when modeling historical records that may reflect regime shifts, climate anomalies, or methodological changes in measurement.

Let $y_1, y_2, \ldots, y_T$ be a sequence of observations. The goal is to detect a set of time indices $\tau_1, \tau_2, \ldots, \tau_m$ such that the data can be partitioned into segments $[1, \tau_1), [\tau_1, \tau_2), \ldots, [\tau_m, T]$ where each segment exhibits statistical homogeneity.

0.2.1 Binary Segmentation

Binary Segmentation is one of the earliest and simplest approaches to CPD. Initially proposed by Scott and Knott (1974), the method recursively tests for a single change point in the sequence. If a change is detected, the sequence is split, and the procedure is

repeated on the resulting sub-segments. Although computationally efficient, the method may miss multiple closely spaced change points.

0.2.2 CUSUM (Cumulative Sum)

CUSUM, introduced by Page (1954), is based on the cumulative sum of deviations from the mean.

$$C_t = \sum_{i=1}^{t} (y_i - \bar{y}), \tag{3}$$

where $\bar{y}$ is the mean of the series. Significant shifts in C_t indicate a change in the underlying distribution. CUSUM is particularly sensitive to mean changes and is often used in quality control and industrial process monitoring.

0.2.3 PELT (Pruned Exact Linear Time)

Proposed by Killick et al. (2012), the PELT algorithm provides an exact and efficient solution to the problem of multiple change point detection by minimizing a cost function:

$$\min_{m,\tau_1,\ldots,\tau_m} \left[\sum_{i=0}^{m} \mathcal{C}(y_{\tau_i+1:\tau_{i+1}}) + \beta m \right], \tag{4}$$

where $\mathcal{C}$ is a cost function for each segment (for example, ne.g.ative logarithmic likelihood), and β is a penalty term controlling the number of change points. PELT guarantees optimal segmentation under certain regularity conditions and has linear computational complexity.

0.2.4 Window-Based Segmentation

In window-based segmentation, a sliding window approach is used to compare statistical characteristics between two adjacent windows. If the divergence (for example, in mean or variance) exceeds a threshold, a change point is declared. This method is non-parametric and flexible, often relying on statistical distances such as the Kolmogorov-Smirnov or Kullback-Leibler divergence.

0.2.5 Bottom-Up Segmentation

Bottom-up segmentation is a hierarchical approach that starts with the finest possible segmentation (each point being its own segment) and merges segments iteratively based on similarity measures. Originally proposed for piecewise linear modeling, this approach is useful in settings where fine granularity is preferred at the start.

0.3 Integrating CPD with Forecasting

To effectively use CPD to improve time-series forecasting, we follow a hybrid strategy. First, change points are identified to divide the series into statistically homogeneous regions. Then individual ARIMA/SARIMA models are fitted to each segment. For future forecasting, more recent segments are weighted more heavily as their dynamics are expected to be more relevant. This ensemble-style forecasting can be formalized as follows.

$$\hat{y}_{T+h} = \sum_{i=1}^{m} w_i \cdot \hat{y}_{T+h}^{(i)}, \tag{5}$$

where $\hat{y}_{T+h}^{(i)}$ is the forecast from the i-th segment model and w_i is a weight assigned based on recency (e.g. $w_i \propto \exp(-\lambda(T - \tau_i))$).

This layered methodology leverages the interpretability of classical models and the adaptability of CPD, providing a robust framework for long-range forecasting in the presence of nonstationarities.

0.4 Modeling Strategy and Segmentation Flow

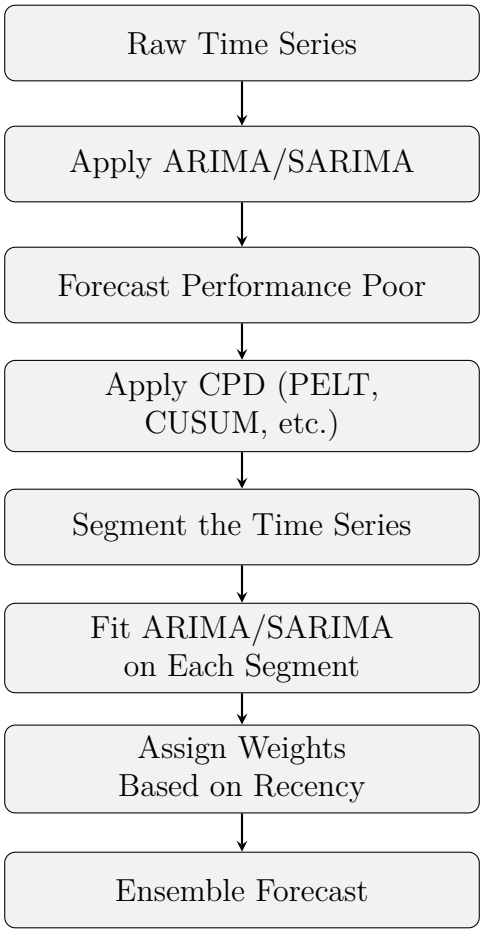


Figure 1: Pipeline for Forecasting with Change Point Detection and Ensemble Modeling

Results and Discussion

Exploratory Data Analysis (EDA)

We began with an exploratory data analysis (EDA) of the global surface temperature anomaly dataset spanning over 170 years. The goal was to uncover key temporal features such as long-term trends, seasonality, and possible structural shifts. Figure 2 shows the monthly temperature anomaly time series.

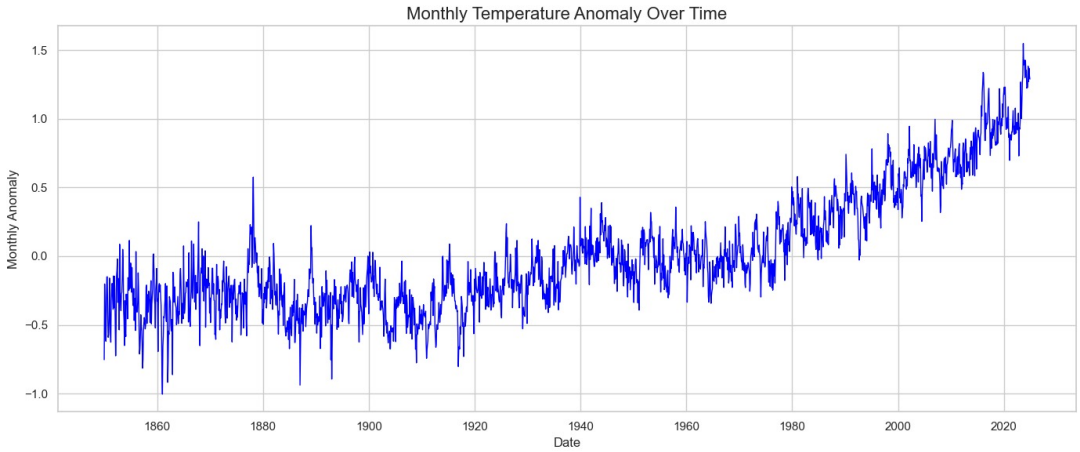


Figure 2: Global Temperature Anomalies (1850–Present)

In this plot, the early part of the series—up to around 1940—shows no clear trend, with anomalies fluctuating near the baseline. From the 1940s onward, a sustained upward trend becomes evident, likely reflecting structural changes driven by industrialization and rising greenhouse gas emissions.

This evolving behavior motivated the use of Change Point Detection (CPD) techniques to identify shifts in the statistical properties of the series. Segmenting the data accordingly enables more accurate and interpretable modeling by accounting for these structural changes.

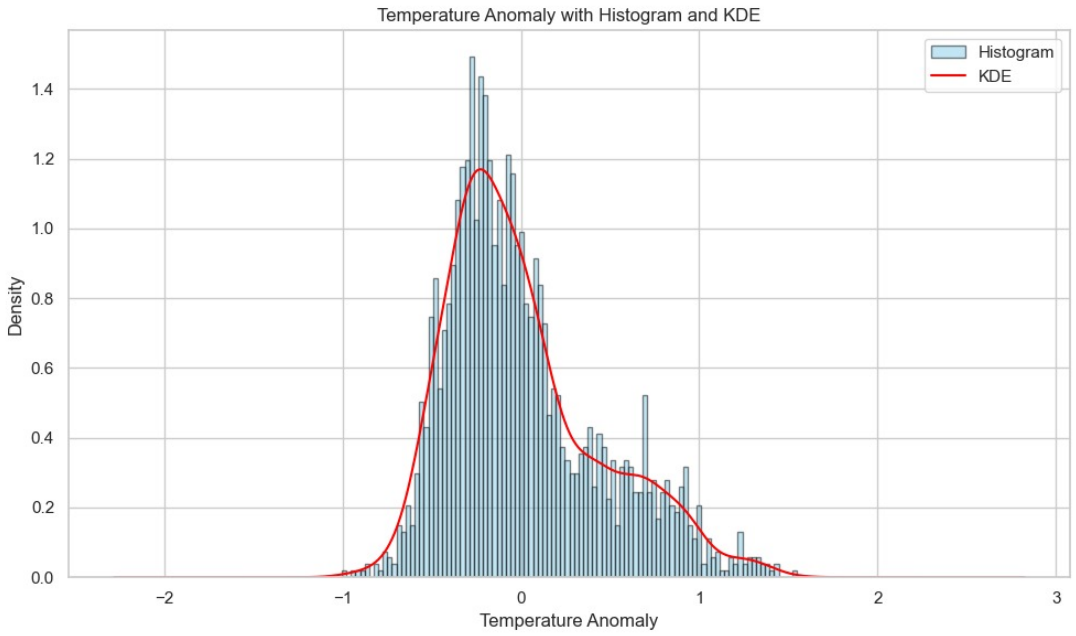


Figure 3: Distribution of Temperature Anomalies

This histogram Figure 3 with KDE overlay shows the distribution of global temperature anomalies. The data is approximately unimodal and right-skewed, with most anomalies clustered around 0°C, indicating small deviations from the long-term average. The long right tail suggests occasional high positive anomalies in recent decades.

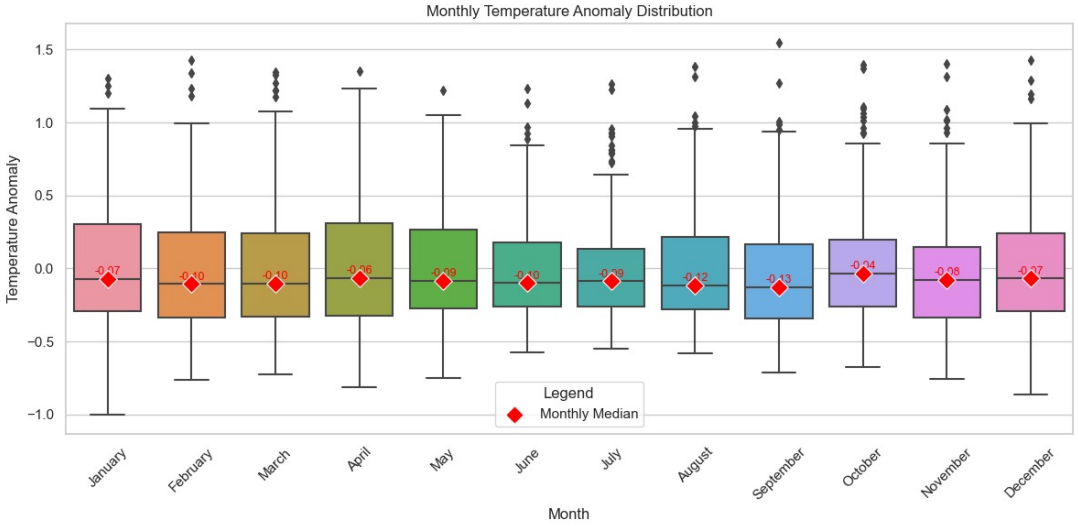


Figure 4: Box-plot of Monthly Temperature Anomalies

This boxplot shows the monthly distribution of temperature anomalies across all years. While anomalies vary widely each month, the medians (in red) are consistently below zero, with slightly more negative values during mid-to-late months like August and September. Outliers are present in every month, highlighting occasional extreme temperature deviations.

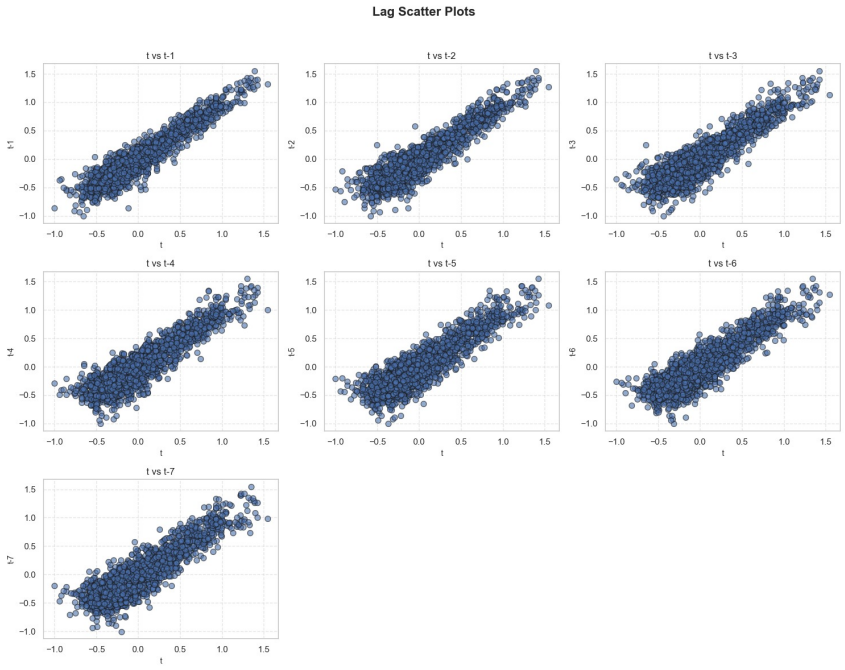


Figure 5: Lag Scatterplot

The lag scatter plots exhibit strong positive autocorrelation, particularly at lower lags. This indicates that the temperature anomaly series has a persistent structure,

where past values strongly influence future ones. Such consistent patterns are ideal for time series modeling, suggesting that predictive models can effectively leverage these temporal dependencies for accurate forecasting.

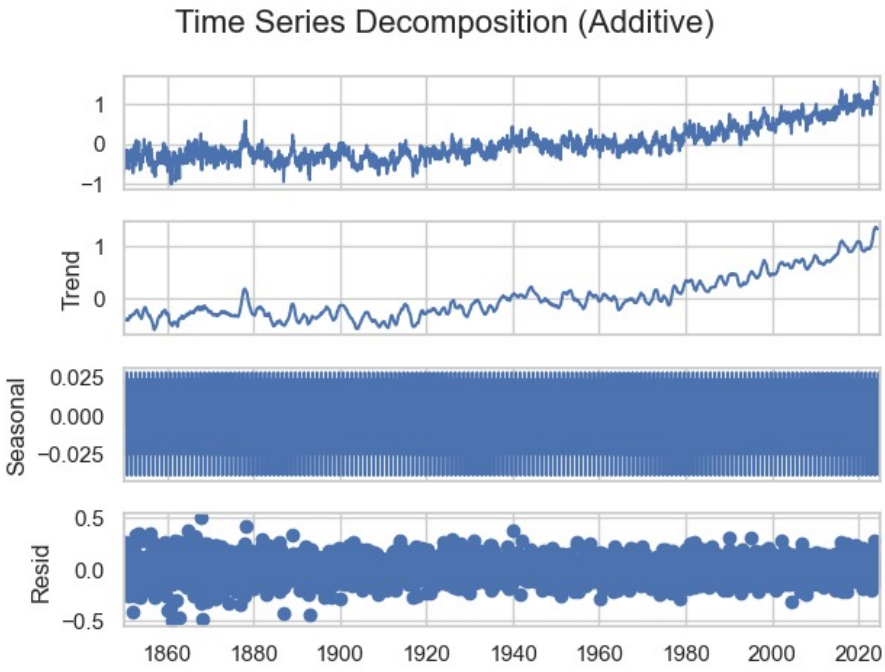


Figure 6: STL Decomposition

Figure 6 illustrates the seasonal decomposition of the global temperature anomaly series. The observed series (top panel) displays an overall upward trajectory with visible fluctuations. The trend component (second panel) captures a consistent increase over time, aligning with global warming patterns. The seasonal component (third panel) remains nearly flat, indicating negligible seasonality in the monthly anomalies. Finally, the residual (bottom panel) resembles random noise, confirming that the trend and seasonal patterns have been effectively extracted from the original series.

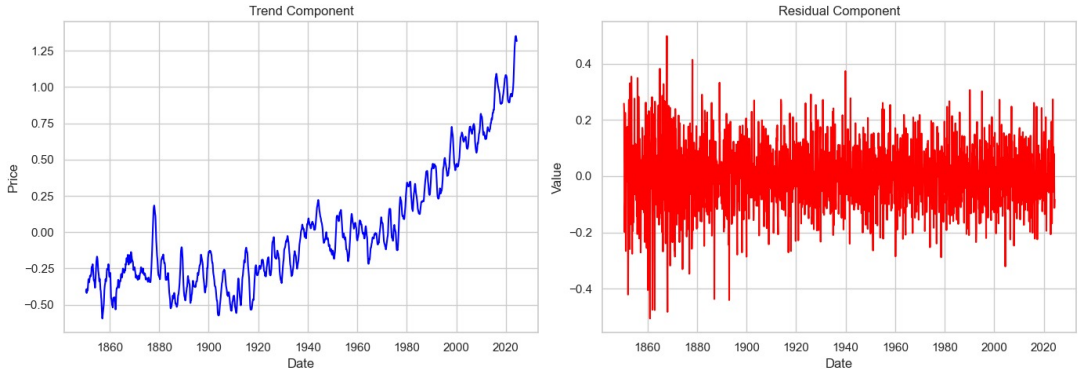


Figure 7: Trend and Residual Components

Figure 7 shows the trend and residual components of the global temperature anomaly series. The trend (left) displays a strong upward pattern, especially after the 1970s, reflecting long-term warming. The residual (right) fluctuates randomly around zero, indicating that the primary structure has been effectively captured by the trend component.

Augmented Dickey-Fuller (ADF) Test

The Augmented Dickey-Fuller (ADF) test is a commonly used statistical test for checking the presence of a unit root in a univariate time series. The presence of a unit root implies that the series is non-stationary, meaning its statistical properties such as mean and variance change over time. Stationarity is a crucial assumption for many time series models, including ARIMA, and must be ensured before proceeding with modeling.

Unlike the basic Dickey-Fuller test, the ADF test includes lagged differences of the series to account for higher-order autocorrelation, making it more robust for practical applications. The test adjusts for serial correlation by adding lag terms, thereby improving the reliability of inference.

Hypotheses:

- H_0: The series has a unit root (non-stationary)
- H_1: The series is stationary

Test Results:

- ADF Statistic: **-0.2591**
- p-value: **0.9311**
- Lags used: 24, Observations: 2075
- Critical values:
  - 1%: -3.4335, 5%: -2.8629, 10%: -2.5675

Conclusion: Since the ADF statistic is greater than all critical values and the p-value is substantially high, we fail to reject H_0. This indicates that the series is **non-stationary**, which motivated the application of first-order differencing to achieve stationarity for reliable time series modeling.

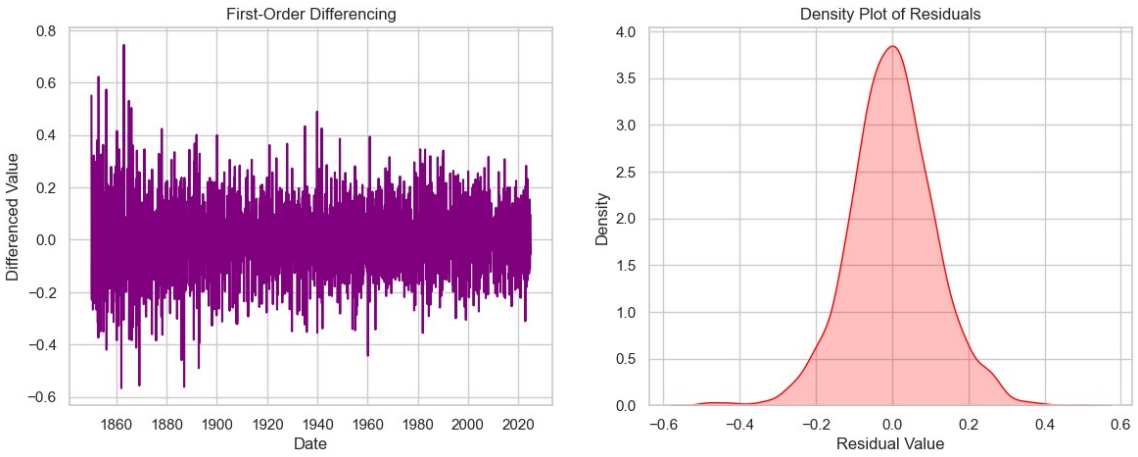


Figure 8: First-Order Differencing and Residual Distribution

Figure 8 illustrates the effect of first-order differencing on the temperature anomaly series (left) and the distribution of residuals (right). Differencing stabilizes the mean and removes long-term trends, yielding a more stationary series. The density plot shows that the residuals are approximately normally distributed, centered around zero. This transformation helps satisfy key assumptions for modeling, such as constant variance and independence. Moreover, it prepares the series for techniques like ARIMA that require stationarity as a prerequisite.

ADF Test After First-Order Differencing

To address the non-stationarity observed in the original series, first-order differencing was applied. The Augmented Dickey-Fuller (ADF) test was then re-conducted on the transformed series to evaluate its stationarity.

Test Results:

- ADF Statistic: **-13.1549**
- p-value: **1.34e-24**
- Lags used: 23, Observations: 2075
- Critical values:
  - 1%: -3.4335, 5%: -2.8629, 10%: -2.5675

Conclusion: The ADF statistic is significantly lower than all critical values, and the p-value is effectively zero. Thus, we reject the null hypothesis and conclude that the differenced series is **stationary**. This confirms that first-order differencing has successfully removed the unit root and stabilized the series for further analysis.

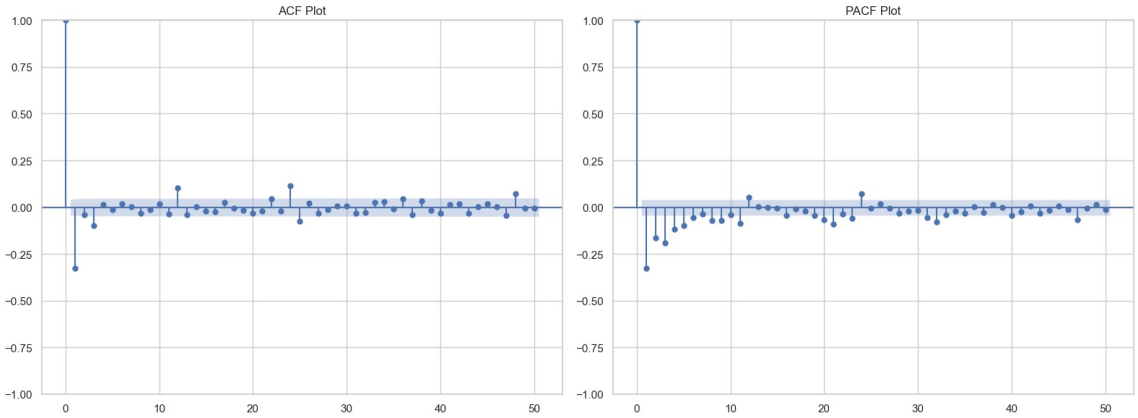


Figure 9: ACF and PACF Plots of Differenced Series

Figure 9 displays the ACF (left) and PACF (right) plots of the first-order differenced temperature anomaly series. In the ACF plot, there is a strong spike at lag 1 followed by a rapid drop-off, with all subsequent lags falling within the confidence bounds. This pattern indicates that the series is now stationary, confirming that differencing was effective.

The PACF plot shows a significant spike at lag 1 and no notable partial autocorrelations beyond that point. This suggests the presence of only one significant lag in the autoregressive structure, pointing toward an AR(1) model as a potential candidate for modeling the differenced series.

0.5 Model Fit and Forecasting

0.5.1 ARIMA

We fitted an Autoregressive Integrated Moving Average (ARIMA) model to the time series data to capture both autoregressive (AR) and moving-average (MA) components, as well as any necessary differencing to ensure stationarity. The general form of the ARIMA(p, d, q) model is:

$$\phi(L)\,(1-L)^d y_t \;=\; \theta(L)\,\varepsilon_t,$$

where L is the lag operator, d is the order of differencing, $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ is the AR polynomial, $\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$ is the MA polynomial, and ε_t is white noise. Model selection was guided by the Akaike Information Criterion (AIC) and examination of residual autocorrelation.

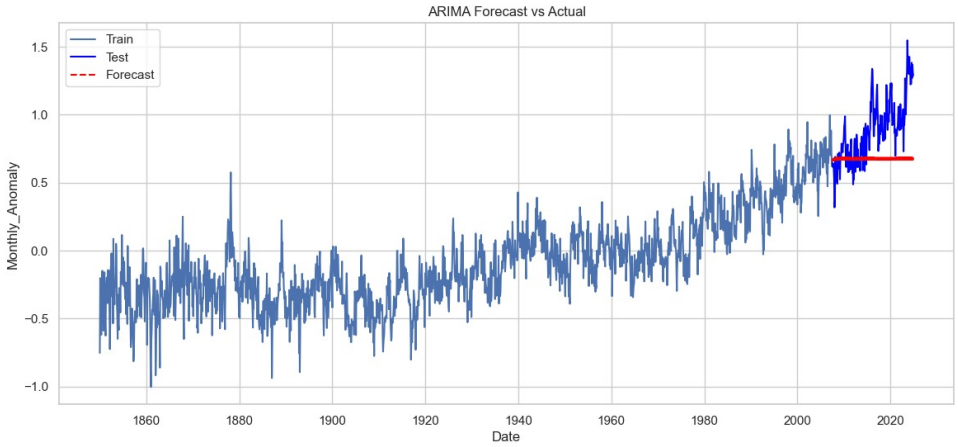


Figure 10: ARIMA model in-sample fit and out-of-sample forecast.

The ARIMA forecasts, shown in Figure 10, fail to capture the underlying dynamics and exhibit large prediction errors in the holdout sample. Overall, the ARIMA model demonstrates poor performance on this dataset.

0.5.2 SARIMA

A Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the ARIMA framework by incorporating seasonal autoregressive and moving-average terms to account for periodic patterns in the data. The general SARIMA$(p, d, q) \times (P, D, Q)_s$ specification is given by:

$$\Phi(L^s)\,\phi(L)\,(1-L)^d\,(1-L^s)^D\,y_t \;=\; \Theta(L^s)\,\theta(L)\,\varepsilon_t,$$

where:

- L is the lag operator,
- s is the seasonal period (e.g., $s = 12$ for monthly data with annual seasonality),
- d and D are the nonseasonal and seasonal orders of differencing, respectively,
- $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ is the nonseasonal AR polynomial,
- $\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$ is the nonseasonal MA polynomial,
- $\Phi(L^s) = 1 - \Phi_1 L^s - \ldots - \Phi_P L^{Ps}$ is the seasonal AR polynomial,

- $\Theta(L^s) = 1 + \Theta_1 L^s + \ldots + \Theta_Q L^{Qs}$ is the seasonal MA polynomial,
- ε_t is white noise.

Model orders (p, d, q) and (P, D, Q) were selected by minimizing the Akaike Information Criterion (AIC), supplemented by inspection of the seasonal and nonseasonal residual autocorrelation functions to ensure no remaining structure.

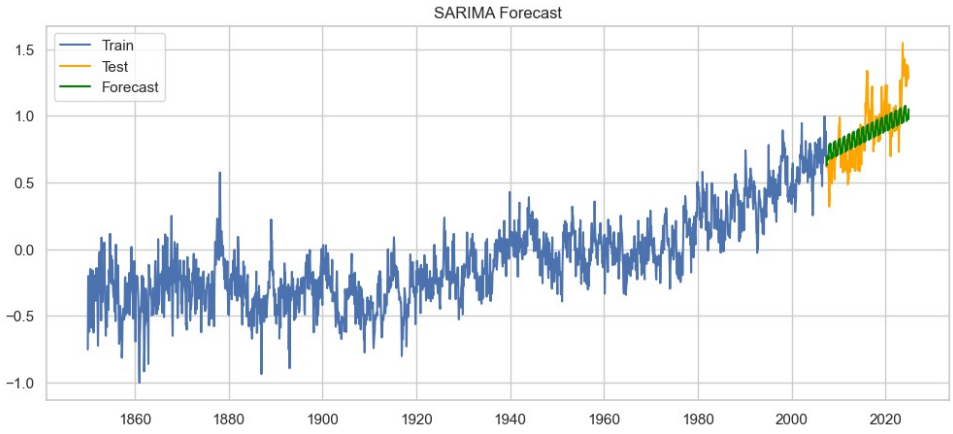


Figure 11: SARIMA model in-sample fit and out-of-sample forecast.

Figure 11 displays the SARIMA in-sample fit and out-of-sample forecasts. Compared to the ARIMA model, the SARIMA specification captures the seasonal cycle more effectively. However, while seasonal peaks are better aligned, the forecasts still exhibit considerable errors in the holdout period, indicating limitations in the linear seasonal structure for this dataset.

0.6 Structural Changes and Change Point Analysis

The suboptimal forecasting performance observed for both the ARIMA and SARIMA models can be attributed to underlying structural changes in the time series that violate the stationarity and linearity assumptions. To address these regime shifts, we propose a two-pronged approach: (1) change point detection to identify and segment distinct regimes, and (2) an ensemble modeling framework that fits specialized models on each detected segment and aggregates their forecasts.

0.6.1 Change Point Detection via PELT

We employ the Pruned Exact Linear Time (PELT) algorithm to detect abrupt shifts in the mean and variance of the series. PELT optimizes a penalized cost function for partitioning the data:

$$\min_{m,\{\tau_i\}} \sum_{i=0}^{m} \Big[C(y_{(\tau_i+1):\tau_{i+1}}) + \beta \Big],$$

where m is the number of change points, $\{\tau_i\}$ are their locations, $C(\cdot)$ is the segment cost (e.g., Gaussian likelihood), and β is the penalty term.

Figure 12 illustrates the series with PELT-identified change points, highlighting key regime shifts.

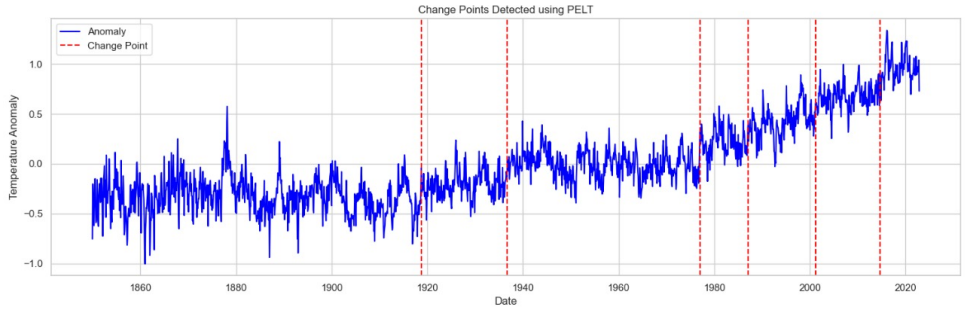


Figure 12: Detected change points using PELT.

0.6.2 Change Point Detection via Binary Segmentation

The Binary Segmentation method iteratively tests for a change point across the entire series, splits at the most significant point, and then repeats within each resulting segment. It maximizes a similar cost reduction criterion but follows a greedy, top-down approach:

$$BS : find\tau^* = \arg\max_{\tau}[\, C(y_{1:\tau}) + C(y_{(\tau+1):T}) - C(y_{1:T})],$$

and then apply recursively on subsegments.

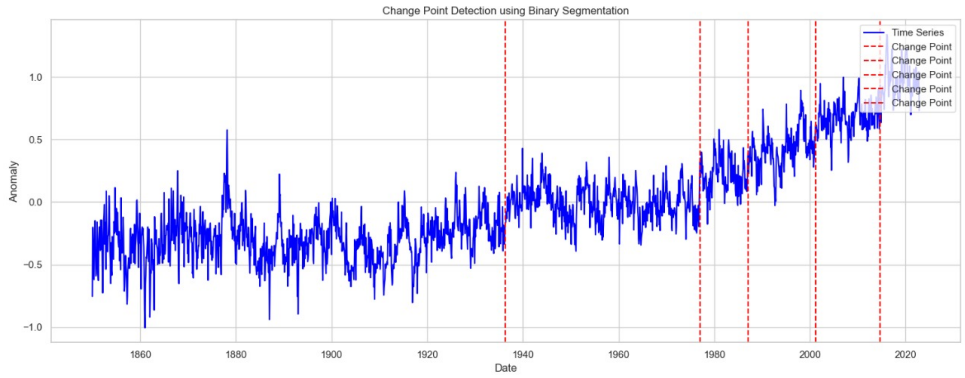


Figure 13: Detected change points using Binary Segmentation.

Figure 13 shows the change points located by Binary Segmentation, offering an alternative decomposition of the series' structural breaks.

0.7 Ensemble Framework Overview

After detecting change points, the time series is partitioned into homogeneous segments. For each segment, we fit the most appropriate univariate model (e.g., ARIMA, SARIMA, Exponential Smoothing) based on information criteria and residual diagnostics. Final forecasts are obtained by concatenating segment-level predictions and applying smoothing at segment boundaries to mitigate discontinuities.

0.7.1 Ensemble Forecasting with Exponential Weights

To improve forecast accuracy, we combine multiple model forecasts using an exponentially weighted ensemble. Suppose we have forecasts $\hat{y}_t^{(k)}$ from K different models for each time point t in the forecast horizon. We assign weights w_k to each model such that more recent models receive higher importance according to an exponential decay governed by a parameter $\lambda \in (0, 1)$:

$$w_k = \frac{\exp(-\lambda(K-k))}{\sum_{j=1}^{K} \exp(-\lambda(K-j))},$$

where $k = 1, 2, \ldots, K$ indexes the models, with $k = K$ representing the most recent model.

The ensemble forecast at time t is then a weighted sum of the individual forecasts:

$$\hat{y}_t^{\text{ens}} = \sum_{k=1}^{K} w_k \hat{y}_t^{(k)}.$$

This scheme emphasizes more recent models while still leveraging information from older models, producing a combined forecast that tends to balance bias and variance more effectively than any single model alone.

0.8 Improved Forecasts and Visualization

The ensemble approach, leveraging exponentially weighted combinations of individual segment forecasts, yields substantially improved predictive accuracy compared to single-model results. By adaptively integrating information across detected regimes and emphasizing recent structural changes, the final forecasts better capture both short-term fluctuations and long-term trends in the data.

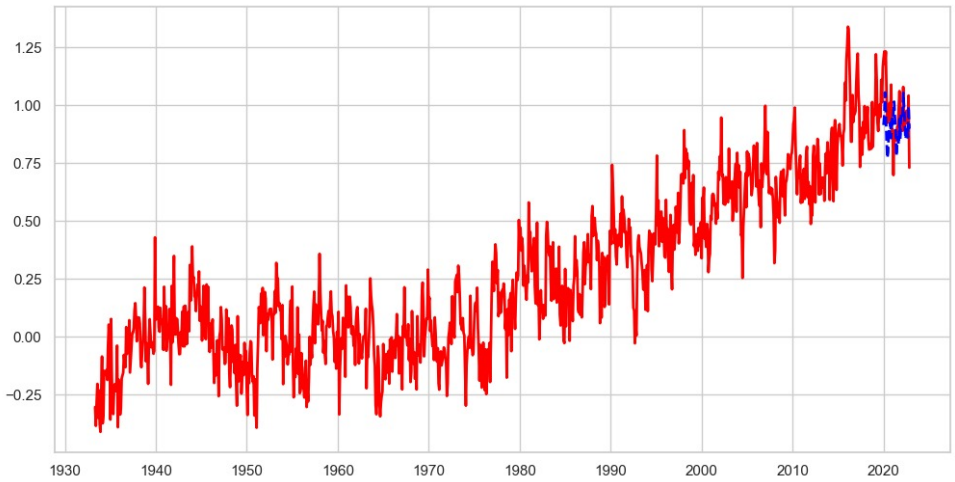


Figure 14: Final ensemble forecast compared to actual values and individual model predictions. The ensemble achieves higher fidelity to the observed test data.