Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Bayesian CUSUM Framework for Integrated Breakpoint Detection, Outlier Reconstruction, and Missing-Value Estimation in Climate Time Series Homogenization

Zohreh Javanshiri

javanshirizohreh@gmail.com

Research Institute of Meteorology and Atmospheric Sciences (RIMAS)

Research Article

Keywords: Climate time series homogenization, Bayesian inference, Breakpoint detection, CUSUM algorithm, Outlier reconstruction, Missing-value estimation

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11360813/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Bayesian CUSUM Framework for Integrated Breakpoint Detection, Outlier Reconstruction, and Missing-Value Estimation in Climate Time Series Homogenization

Zohreh Javanshiri[1*]

[1*]Climate Research Center (CRC), Research Institute of Meteorology and Atmospheric Sciences (RIMAS), Mashhad, Razavi Khorasan Province, Iran.

Corresponding author(s). E-mail(s): javanshirizohreh@gmail.com;

Abstract

Homogenization of climate time series is essential for reliable climate change assessment, yet distinguishing structural breakpoints from transient anomalies remains challenging when outliers and missing observations coexist with artificial shifts. Existing homogenization methods mainly focus on breakpoint detection, whereas the combined effects of structural breaks, outliers, and missing observations have received comparatively little attention. This study proposes a Bayesian Cumulative Sum (BYS-CUSUM) framework that integrates breakpoint verification, outlier identification, missing-value reconstruction, and homogenization within a unified probabilistic workflow. Candidate breakpoints are first identified using the CUSUM algorithm applied to regression residuals derived from neighboring reference stations. Bayesian inference is then used to estimate breakpoint locations and shift magnitudes while quantifying their uncertainty. Outliers are identified from the same residuals, and both detected outliers and missing observations are reconstructed using the regression model before breakpoint adjustments are applied. The proposed framework was evaluated against RHtestsV4 and Climatol using four synthetic benchmark scenarios with progressively increasing breakpoint complexity, outlier contamination, and missing observations. While all methods performed similarly under ideal conditions, the benchmark methods became increasingly vulnerable as contamination increased. Climatol frequently missed subtle breakpoints, whereas RHtestsV4 produced false or displaced detections. In contrast, BYS-CUSUM consistently achieved

more accurate breakpoint localization and superior reconstruction of contaminated observations, reducing the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) by up to 83% and 82%, respectively. These results demonstrate that BYS-CUSUM provides a robust and effective framework for climate time series homogenization under realistic observational conditions.

Keywords: Climate time series homogenization, Bayesian inference, Breakpoint detection, CUSUM algorithm, Outlier reconstruction, Missing-value estimation

1 Introduction

Climate change is one of the most significant environmental challenges of the twenty-first century, with profound impacts on ecosystems, water resources, agriculture, public health, and socio-economic systems worldwide. Reliable assessments of long-term climate variability and change fundamentally depend on the availability of high-quality, temporally consistent meteorological observations, particularly long-term records of surface air temperature and precipitation. These observations constitute the foundation for climate trend analysis, model evaluation, extreme-event assessment, and the development of evidence-based adaptation and mitigation strategies. Consequently, ensuring the temporal homogeneity of climate time series has become an indispensable prerequisite for producing reliable climate information and supporting both scientific research and operational climate services ([1–5]).

Despite rigorous observational protocols, long-term meteorological records are rarely free from artificial inhomogeneities. Non-climatic factors, including station relocation, instrument replacement, sensor recalibration, changes in observing practices, urbanization surrounding monitoring sites, and modifications to data processing procedures, frequently introduce abrupt artificial shifts into climate records. Such discontinuities violate the assumption of temporal consistency by producing systematic changes that are unrelated to genuine climate variability. If these artificial disturbances remain uncorrected, they may substantially bias estimates of long-term climate trends, distort the frequency and intensity of climate extremes, and ultimately compromise the reliability of regional and global climate assessments ([1, 6, 7]).

To minimize these effects, climate data homogenization aims to detect, characterize, and correct artificial discontinuities before climate analyses are performed. During the past three decades, homogenization has evolved into a standard preprocessing procedure in climatological research, and international initiatives coordinated by the World Meteorological Organization (WMO) have consistently emphasized its importance for developing reliable climate datasets. Comparative benchmark studies have repeatedly demonstrated that homogenized observations provide considerably more accurate representations of long-term climate variability than raw measurements, highlighting the critical role of homogenization in modern climate research ([3, 7–12]).

Artificial disturbances in climate records rarely occur in a single, isolated form. Instead, observational series are commonly affected by a combination of persistent

structural shifts, transient anomalous observations, and missing values, each originating from different sources and requiring distinct statistical treatment. Structural breakpoints generally reflect permanent changes in observing conditions, whereas outliers arise from temporary sensor malfunctions, transmission errors, recording mistakes, or short-lived environmental disturbances. Missing observations further complicate the analysis by reducing the amount of available information for statistical inference. Although these disturbances differ in both origin and duration, they frequently coexist within the same climate record, making reliable homogenization substantially more challenging than the detection of structural changes alone.

Distinguishing among these different forms of artificial disturbance remains one of the principal challenges in climate time series homogenization. Persistent structural changes and transient anomalous observations may produce similar statistical signatures, particularly when mean shifts are relatively small or when anomalous observations occur in close temporal proximity to structural changes. Conversely, high background variability and incomplete observations may obscure genuine discontinuities, increasing the likelihood of missed detections. Because these disturbances influence one another, reliable homogenization requires statistical procedures capable of separating persistent changes from transient anomalies while simultaneously accounting for observational uncertainty and incomplete records.

To address these challenges, numerous statistical methods have been proposed for climate homogenization. Classical hypothesis-testing procedures, including the Standard Normal Homogeneity Test (SNHT), the Buishand Range Test, Pettitt's test, and likelihood-ratio-based approaches, remain widely used because of their conceptual simplicity and computational efficiency. However, these methods generally assume relatively ideal observational conditions and often experience reduced detection accuracy when multiple artificial disturbances coexist within the same observational record ([13–15]).

More recently, operational homogenization software has substantially improved the robustness of climate data quality control. RHtests employs penalized maximal t- and F-tests together with likelihood-based procedures to detect multiple change points while accounting for temporal autocorrelation ([16]). In contrast, Climatol implements an iterative SNHT-based approach that exploits information from neighboring reference stations to identify and adjust multiple artificial discontinuities ([17]). Both methods have demonstrated strong performance across a wide range of climatological applications and are now widely adopted for the operational homogenization of climate observations.

Nevertheless, current methodologies are primarily designed to detect and adjust structural discontinuities, while the interactions among breakpoints, transient outliers, and missing observations receive comparatively less attention. In practice, these disturbances seldom occur independently. Outliers located near structural changes may complicate breakpoint identification, while inaccurate breakpoint estimation may subsequently influence outlier reconstruction and missing-value estimation. Consequently, errors introduced during one stage of the homogenization procedure may propagate to

subsequent stages, ultimately affecting the quality of the homogenized series. Developing statistical methodologies capable of jointly accounting for these interacting disturbances therefore remains an important challenge in climate homogenization.

Among sequential monitoring techniques, the Cumulative Sum (CUSUM) algorithm has received considerable attention because of its computational efficiency and high sensitivity to persistent mean shifts. Originally developed for industrial statistical process control, CUSUM has evolved into one of the most widely used sequential change-point detection techniques and has been extensively studied in the statistical process control and change detection literature ([18–20]). Its ability to rapidly identify candidate change points makes it particularly attractive for large observational datasets.

Bayesian inference provides a natural probabilistic framework for addressing these challenges by explicitly accounting for uncertainty throughout the homogenization process. Unlike classical hypothesis-testing approaches, Bayesian methods combine prior knowledge with observed data to estimate posterior probability distributions of unknown parameters, thereby enabling simultaneous inference on breakpoint location, shift magnitude, and associated uncertainty. Rather than producing a binary decision regarding the presence or absence of a breakpoint, Bayesian models quantify the probability of alternative explanations for the observed data, offering greater flexibility when structural changes coexist with anomalous observations or incomplete records. These characteristics make Bayesian inference particularly well suited to climate homogenization, where multiple sources of uncertainty frequently arise simultaneously ([21–23]).

Recent advances in probabilistic programming have substantially improved the practical implementation of Bayesian hierarchical models. In particular, Stan has become one of the most widely adopted platforms for Bayesian computation because it employs Hamiltonian Monte Carlo (HMC), an efficient gradient-based Markov Chain Monte Carlo algorithm capable of sampling from complex posterior distributions with high computational efficiency and robust convergence properties ([24–26]). These developments have made Bayesian approaches increasingly feasible for large-scale climate applications requiring both statistical rigor and computational efficiency.

Despite substantial advances in climate homogenization, important methodological challenges remain when multiple artificial disturbances coexist within the same climate record. Most existing methods employ sequential processing strategies in which breakpoint detection, outlier treatment, and missing-value estimation are performed in separate stages. Although these approaches have demonstrated satisfactory performance in many practical applications, their accuracy may deteriorate when different disturbances interact, motivating the development of more robust homogenization strategies.

Furthermore, breakpoint detection itself becomes considerably more difficult when structural shifts are small, occur near the beginning or end of a time series, or are masked by transient outliers and missing observations. Under such conditions, deterministic decision rules may fail to adequately distinguish persistent structural changes from short-lived anomalies or to quantify the uncertainty associated with breakpoint

location and magnitude. These challenges motivate the use of a probabilistic framework that explicitly characterizes uncertainty while providing robust estimation of structural breakpoints together with consistent treatment of outliers and missing observations.

To address these challenges, this study proposes a novel Bayesian Cumulative Sum (BYS-CUSUM) framework for climate time series homogenization. The proposed methodology integrates the computational efficiency of CUSUM with Bayesian posterior inference implemented through HMC within a unified probabilistic framework. Candidate breakpoints are first identified using CUSUM applied to regression residuals derived from neighboring reference stations. Bayesian inference is subsequently employed to estimate the posterior distributions of breakpoint location and shift magnitude while explicitly accounting for parameter uncertainty. Following probabilistic breakpoint verification, transient outliers and missing observations are reconstructed, and the estimated shifts are used to produce the final homogenized climate series. The principal contributions of this study are summarized as follows:

1. A unified workflow integrating Bayesian breakpoint verification, residual-based outlier identification, regression-based reconstruction of missing observations, and final climate data homogenization.
2. Computationally efficient breakpoint estimation through CUSUM-based candidate screening followed by local Bayesian refinement, avoiding Bayesian inference over the entire time series.
3. Robust Bayesian breakpoint verification through a Student-t likelihood, reducing the influence of transient outliers on posterior estimation of breakpoint locations and shift magnitudes.
4. Probabilistic estimation of breakpoint locations and shift magnitudes, providing explicit uncertainty quantification for structural changes in climate time series.

The remainder of this paper is organized as follows. Section 2 presents the proposed BYS-CUSUM framework, including CUSUM-based candidate breakpoint detection, Bayesian breakpoint verification, residual-based outlier identification, regression-based reconstruction of missing observations, and the final homogenization procedure. Section 3 describes the experimental design and the simulation framework used to generate synthetic climate time series for evaluating the proposed methodology. Section 4 presents the benchmark experiments and comparative performance evaluation against RHtestsV4 and Climatol. Section 5 discusses the experimental results, analyzes the strengths and limitations of the proposed framework, and highlights its performance under different contamination scenarios. Finally, Section 6 summarizes the main findings and outlines directions for future research.

2 Proposed BYS-CUSUM Framework

To address the methodological challenges outlined in the Introduction, we developed the BYS-CUSUM framework, which combines efficient candidate screening with probabilistic breakpoint verification. The framework integrates structural breakpoint

estimation, residual-based outlier detection, regression-based reconstruction of outliers and missing observations, and final breakpoint adjustment within a unified workflow (Fig. 1).

The procedure begins by constructing a regression model between the target station and its neighboring reference stations, from which regression residuals are obtained. These residuals constitute the common information source for two complementary branches of the framework. In the first branch, candidate breakpoints are efficiently identified using the CUSUM algorithm and subsequently refined through Bayesian inference to estimate breakpoint locations, shift magnitudes, and their associated uncertainty. In parallel, the same regression residuals are analyzed to identify transient outliers. The detected outliers, together with genuine missing observations, are then reconstructed using the regression relationship. Finally, the estimated breakpoint adjustments are applied to the reconstructed series to produce the homogenized climate record.

2.1 Stage I: Automatic Candidate Screening via Residual CUSUM

The first Stage of the BYS-CUSUM framework focuses on the automated identification of potential breakpoint candidates. This stage is designed to maximize sensitivity, thereby minimizing the risk of overlooking subtle shifts before the rigorous Bayesian evaluation performed in Stage II.

2.1.1 Composite Reference and Residual Generation

To isolate the non-climatic signal, a composite reference series (X_{comp}) is constructed by averaging standardized data from neighboring stations. Standardization removes differences in absolute temperature levels among neighboring stations, allowing the composite reference to represent the common regional climate signal rather than station-specific variability. The target series (y_t) is then regressed against this reference and seasonal harmonics:

$$y_t = \alpha + \beta_{ref} X_{comp,t} + \beta_{s1} \sin\left(\frac{2\pi t}{12}\right) + \beta_{s2} \cos\left(\frac{2\pi t}{12}\right) + e_t \tag{1}$$

The residuals (e_t) represent the target station's local deviations. To maintain the continuity of the cumulative summation in the presence of missing observations, any missing values (NAs) in e_t are temporarily imputed with 0, yielding a modified residual series denoted as e'_t.

2.1.2 CUSUM Transformation and Kernel Smoothing

The CUSUM transformation provides an efficient mechanism for accumulating persistent deviations in the regression residuals, thereby enhancing the visibility of structural changes relative to short-term random variability. The CUSUM series (S_i) is calculated

as the running sum of the modified residuals:

$$S_i = \sum_{j=1}^{i} e'_j \tag{2}$$

To suppress high-frequency meteorological noise and seasonal artifacts, a kernel smoother employing a Gaussian kernel with a 12-month bandwidth is applied directly to the S_i series, yielding the smoothed cumulative sum series ($\hat{S}_i$).

2.1.3 Detection of Turning Points and Local Dominance

Breakpoint candidates are identified by analyzing the geometric properties of the smoothed CUSUM curve ($\hat{S}_i$). The process involves three sequential filtering steps:

- Turning Point Identification: We identify points where the sign of the first derivative of $\hat{S}_i$ changes, representing local maxima (peaks) or minima (valleys).
- Local Dominance Filter: To prevent over-detection due to small fluctuations, a local search window with a radius of 48 months (yielding a total window length of 96 months, equivalent to an 8-year span) is deployed. A turning point at time tp is only retained if it represents the absolute extremum (maximum or minimum) within its local neighborhood ($tp \pm 48$ months). This ensures that only globally significant shifts within an 8-year period are considered.
- Boundary Constraint: To avoid unstable estimations near the start and end of the record, candidates within the first or last 24 months of the series are excluded.

The resulting candidates are passed to Stage II as initial breakpoint candidates for Bayesian refinement. Posterior estimates of breakpoint magnitude and uncertainty are subsequently used to compute the Signal-to-Noise Ratio (SNR), which determines whether each candidate is retained as a confirmed breakpoint.

2.2 Stage II: Bayesian Refinement and Magnitude Estimation

In the second phase, each candidate identified in Satge I is rigorously evaluated using a Bayesian State-Space model. This phase transitions from a global screening to a local, robust verification to precisely estimate the timing and magnitude of shifts.

2.2.1 Local Window Extraction and Standardization

For each candidate c, a local analysis window is defined with a radius of 50 months ($c \pm 50$). To ensure numerical stability and prior consistency within the BYS-CUSUM framework, the target series (y_{win}) and the reference matrix (X_{refs}) within this window are standardized (Z-score normalization). A minimum requirement of 60 valid (non-missing) observations is enforced to ensure sufficient information for posterior convergence.

2.2.2 Bayesian Hierarchical State-Space Model

The core of the refinement process is a Bayesian regression model implemented in `Stan`. Rather than assuming a Gaussian likelihood, the observation model is specified using a Student-t distribution. Owing to its heavier tails, the Student-t likelihood down-weights extreme residuals, thereby reducing the influence of transient outliers on posterior inference for breakpoint location and shift magnitude without requiring explicit outlier removal. The model is formulated as:

$$y_t \sim \text{student_t}(\nu, \mu_t, \sigma_y) \tag{3}$$

where ν denotes the degrees-of-freedom parameter controlling the tail heaviness of the Student-t distribution, μ_t represents the latent mean, and σ_y denotes the residual scale parameter.

The latent mean μ_t is decomposed into climatic and non-climatic components:

$$\mu_t = \alpha + \beta_{ref} X_{ref,t} + \text{Seasonal}_t + \text{Trend}_t + \text{Step}_t \tag{4}$$

- Soft-Step Transition: To accommodate uncertainty in breakpoint timing while maintaining a smooth approximation to an abrupt transition, the breakpoint is modeled using a logistic transition function: $\text{Step}_t = \gamma \cdot \text{logit}^{-1}(\kappa(t - t_0))$, where t_0 is the estimated breakpoint time, γ is the magnitude, and κ (constrained between 2 and 30) controls the transition's sharpness.
- Stochastic Trend: A random walk component ($\text{Trend}_t = \sum \epsilon_{rw}$) is integrated to capture low-frequency climate variability and thereby reducing the risk that gradual climatic variability is erroneously attributed to an abrupt structural breakpoint.
- MCMC Configuration: We employ the No-U-Turn Sampler (NUTS) with 4 parallel chains, each consisting of 1,000 warmup iterations and 2,000 sampling iterations. To handle the complex posterior geometry, the *adapt_delta* parameter is set to 0.99.

2.2.3 Prior Specifications

Since the model is estimated using standardized data (Z-score scale), weakly informative priors are assigned to the model parameters to facilitate stable sampling and reduce the risk of overfitting:

$$\alpha \sim \mathcal{N}(0, 0.5^2) \tag{5}$$

$$\beta_{ref} \sim \mathcal{N}(0.5, 0.5^2) \tag{6}$$

$$\sigma_y \sim \mathcal{N}^+(0, 1^2) \tag{7}$$

$$\sigma_{rw} \sim \mathcal{N}^+(0, 0.05^2) \tag{8}$$

$$\beta_{season} \sim \mathcal{N}(0, 1^2) \tag{9}$$

$$\nu \sim \text{Gamma}(2, 0.1) \tag{10}$$

The Gamma prior on the degrees-of-freedom parameter allows the likelihood to adapt between heavy-tailed and approximately Gaussian behavior according to the information contained in the data.

$$\kappa \sim \mathcal{N}(5, 2^2) \quad \text{truncated to } [2, 30] \tag{11}$$

To impose regularizing shrinkage on the step-change magnitude and reduce the risk of detecting spurious jumps, a zero-centered prior is assigned to the breakpoint magnitude parameter:

$$\gamma \sim \mathcal{N}(0, 2^2) \tag{12}$$

For the breakpoint timing parameter t_0, a continuous uniform prior is specified over the local analysis window, excluding the boundary edges to improve numerical stability:

$$t_0 \sim \mathcal{U}(12, N - 12) \tag{13}$$

where N is the number of observations in the local analysis window.

2.2.4 Candidate Validation and and Magnitude Back-Transformation

Posterior inference provides the final estimates of the breakpoint location t_0 and and breakpoint magnitude γ. A candidate is officially confirmed only if its SNR meets the following criterion:

$$\text{SNR} = \frac{|\mu_\gamma|}{\sigma_\gamma} \geq 1.0 \tag{14}$$

where μ_γ and σ_γ are the posterior mean and standard deviation of the jump magnitude, respectively. Confirmed magnitudes are then back-transformed to the original scale (e.g., °C) using the local standard deviation.

2.2.5 Deduplication via Hierarchical Clustering

To address the "double-detection" of a single event (which may occur due to overlapping windows in Stage I), a final deduplication step is performed. We apply Agglomerative Hierarchical Clustering (HAC) to the confirmed breakpoints. Confirmed breakpoints separated by less than 24 months are assigned to the same cluster. For each cluster, the representative breakpoint location is defined as the temporal mean of the clustered candidates, while the corresponding breakpoint magnitude is obtained by averaging their posterior estimates.

2.3 Stage III: Climate Series Reconstruction and Homogenization

The final stage of the proposed BYS-CUSUM framework is dedicated to reconstructing and homogenizing the climate time series. This stage integrates the information obtained from the previous stages to generate the final corrected series. First, transient outliers are identified from the regression residuals and treated as missing observations.

Both detected outliers and originally missing values are subsequently reconstructed using the global regression model developed in Stage I. Finally, the posterior estimates of breakpoint magnitudes obtained in Stage II are sequentially applied to correct the identified structural discontinuities, yielding the final homogenized climate time series.

2.3.1 Outlier Detection and Missing-Value Reconstruction

Following Bayesian validation of breakpoint candidates, the regression residuals are re-examined to identify transient anomalous observations. Outliers are detected using a residual-based thresholding criterion. Specifically, observations whose absolute residual exceeds three times the residual standard deviation are classified as outliers:

$$|e_t| > 3\sigma_e \tag{15}$$

where e_t denotes the regression residual and σ_e is its standard deviation. Detected outliers are temporarily treated as missing observations by replacing their values with `NA`. Subsequently, both originally missing observations and detected outliers are reconstructed using the global regression model developed in Satge I. The regression model predicts the expected climatic value from the composite reference series and seasonal harmonic components:

$$\hat{y}_t = \hat{\alpha} + \hat{\beta}_{ref} X_{comp,t} + \hat{\beta}_{s1} \sin\left(\frac{2\pi t}{12}\right) + \hat{\beta}_{s2} \cos\left(\frac{2\pi t}{12}\right) \tag{16}$$

Predicted values replace all missing observations and detected outliers. If isolated missing values remain because of unavailable predictor information, a one-dimensional linear interpolation is applied to ensure a complete time series before the homogenization stage.

2.3.2 Homogenization of the Climate Series

Consistent with standard climatological homogenization practice, the most recent homogeneous segment is adopted as the reference period. Therefore, all observations preceding each confirmed breakpoint are adjusted using the corresponding posterior estimate of the breakpoint magnitude to produce the final homogenized climate series.

Confirmed breakpoints are first sorted in reverse chronological order to prevent successive corrections from affecting previously adjusted segments. For each confirmed breakpoint occurring at time t_b with estimated magnitude γ_b, all observations preceding the breakpoint are adjusted according to

$$y_t^{(h)} = y_t^{(r)} + \gamma_b, \qquad t \le t_b \tag{17}$$

where $y_t^{(r)}$ denotes the reconstructed series after outlier and missing-value correction, and $y_t^{(h)}$ represents the final homogenized series.

Applying breakpoint corrections sequentially from the most recent breakpoint toward the beginning of the record ensures that multiple structural shifts are accumulated consistently without altering previously corrected segments. The resulting series simultaneously incorporates reconstructed outliers, estimated missing observations, and cumulative breakpoint adjustments, yielding the final homogenized climate record.

3 Synthetic Climate Data Generation

To provide a rigorous and realistic evaluation of the proposed BYS-CUSUM framework, a synthetic climate dataset was generated using statistical characteristics extracted from an actual meteorological network. Rather than relying on idealized white-noise simulations, the proposed simulation framework preserves the temporal persistence, seasonal variability, long-term warming trend, and spatial dependence observed in real climate records. The empirical parameters were extracted from three meteorological stations located in Northeast Iran, namely Mashhad (target station), Sabzevar (reference station 1), and Torbat-e Heydarieh (reference station 2), using monthly mean temperature observations spanning the period 1959–2025. This design enables a representative assessment of homogenization performance under conditions that closely resemble operational climatological datasets.

3.1 Empirical Parameter Extraction via STL Decomposition

Let $y_{i,t}$ represent the observed monthly mean temperature of station $i \in \{1, 2, 3\}$ at time t. To isolate the underlying statistical parameters, each empirical series is decomposed using Seasonal-Trend decomposition based on Loess (STL):

$$y_{i,t} = T'_{i,t} + S'_{i,t} + R'_{i,t} \tag{18}$$

where $T'_{i,t}$ represents the low-frequency trend, $S'_{i,t}$ is the seasonal component, and $R'_{i,t}$ is the stationary remainder. Prior to STL decomposition, missing observations were interpolated using Kalman smoothing after seasonal adjustment to ensure complete time series. STL decomposition was employed solely to estimate the statistical characteristics required for the simulation framework and was not involved in the proposed homogenization methodology. From these empirical components, four key climate parameters are estimated:

- Long-Term Trend (b): A coherent long-term trend is computed by averaging the trend components across all stations, $\bar{T}_t = \frac{1}{3}\sum_{i=1}^{3} T'_{i,t}$. A linear regression is then fitted over the time index:

$$\bar{T}_t = a + b \cdot t + e_{trend} \tag{19}$$

  where the slope b denotes the baseline monthly global warming rate.
- Seasonal Amplitude (A_i) and Baseline (L_i): For each station, the empirical baseline mean temperature level is computed as $L_i = \mathbb{E}[y_{i,t}]$. The historical seasonality amplitude is defined as $A_i = \max_t(S'_{i,t})$.
- Regional Climate Signal (ϕ_{reg}, σ_{reg}): To capture large-scale, low-frequency atmospheric oscillations shared across the region, a regional remainder series is constructed as the spatial average of the residuals: $\bar{R}_t = \frac{1}{3}\sum_{i=1}^{3} R'_{i,t}$. An Autoregressive model of order 1, AR(1), is fitted to this series:

$$\bar{R}_t = \phi_{reg}\bar{R}_{t-1} + \epsilon_{reg,t}, \quad \epsilon_{reg,t} \sim \mathcal{N}(0, \sigma^2_{reg}) \tag{20}$$

where ϕ_{reg} is the autoregressive coefficient and σ_{reg} is the residual standard deviation.

- Spatially Correlated Local Residuals ($\mathbf{\Sigma}_{noise}$): The localized microclimatic variations are captured through the spatial covariance matrix ($\mathbf{\Sigma}_{noise} \in \mathbb{R}^{3\times3}$) computed directly from the detrended and deseasonalized empirical remainder matrix $\mathbf{R}'$.

3.2 Generative Mathematical Framework

Using the parameters extracted above, we simulate $H = 3$ synthetic stations over a horizon of $N = 804$ months (equivalent to 67 years). The synthetic temperature series $Y_{i,t}$ for station i at month t is formulated as the superposition of five physically-meaningful components:

$$Y_{i,t} = L_i + T_t + \Omega_t + S_{i,t} + \eta_{i,t} \tag{21}$$

where:

- Linear Trend Component (T_t): Represents the global climate change forcing, defined as:
$$T_t = b \cdot t \tag{22}$$
- Coherent Regional Signal (Ω_t): Models regional climate variability and is generated as an AR(1) process scale-adjusted by 50% to prevent over-dominance:
$$\Omega_t = \phi_{reg}\Omega_{t-1} + \zeta_t, \quad \zeta_t \sim \mathcal{N}\left(0, (0.5\sigma_{reg})^2\right) \tag{23}$$
- Station-Specific Seasonality ($S_{i,t}$): Models annual cycles with minor, stochastic phase shifts (θ_i) and amplitude variations (A'_i) to mimic microclimatic discrepancies:
$$S_{i,t} = \sin\left(\frac{2\pi t}{12} + \theta_i\right) \cdot A'_i \tag{24}$$
where $\theta_i \sim \mathcal{U}(-0.1, 0.1)$ and $A'_i \sim A_i \cdot \mathcal{U}(0.95, 1.05)$.
- Spatially and Temporally Correlated Local Noise ($\eta_{i,t}$): Generating realistic climate noise requires incorporating both spatial covariance (neighbor similarity) and temporal inertia (persistence). First, a spatially correlated multivariate white noise vector is generated:
$$\mathbf{w}_t \sim \mathcal{MVN}(\mathbf{0}, \mathbf{\Sigma}_{noise}) \tag{25}$$
Then, a univariate AR(1) filter is applied element-wise to propagate the temporal dependency:
$$\eta_{i,t} = \phi_{reg}\eta_{i,t-1} + w_{i,t} \tag{26}$$

This simulation framework preserves the principal statistical characteristics of the observed meteorological network, including temporal persistence, seasonal variability,

and spatial dependence, thereby providing a realistic benchmark for evaluating climate homogenization methods. The synthetic temperature series generated using this framework constitute the baseline homogeneous dataset. Different experimental scenarios were subsequently produced by introducing controlled structural breakpoints, missing observations, and outliers with predefined magnitudes and locations, as described in Section 4.

4 Experimental Evaluation and Benchmark Examples

The comparative study is structured through four experimental scenarios, progressing from simple mean shifts to complex cases involving smart outliers and high missing data ratios (up to 25 instances per series). In all experiments, the primary objective is to recover the latent "True" series from the contaminated "Observed" data. To ensure a rigorous evaluation, the proposed BYS-CUSUM framework is compared against Climatol (v3.1.1) and RHtestsV4, configured to leverage their maximum capabilities.

4.1 Configuration of Benchmark Methods

For Climatol, the `homogen` function was executed with the standardization parameter set to `std=3` (appropriate for temperature series) and the SNHT detection threshold set to `snht1=snht2=20`. The default value of the outlier tolerance parameter (`dz.max=6`) was used for Examples 1 and 4. For Examples 2 and 3, larger anomalies required increasing `dz.max` to 9 to ensure stable execution. This relaxation, however, reduced the algorithm's ability to identify and remove extreme outliers.

For RHtestsV4, the analysis was performed using the `FindU.wRef` and `FindUD.wRef` functions. To fulfill the requirement of a reference series, a composite reference series was constructed by averaging two neighboring stations ($Ref = [R_1 + R_2]/2$), mirroring the information availability of the BYS-CUSUM model. The detection was based on a penalized maximal F-test with a significance level of $p = 0.05$ (`p.lev=0.95`). The final homogenization was performed using the Quantile Matching (QM) adjustment (`UD_data[, 11]`), which corrects both the mean and the distribution of the series.

4.2 Example 1: Pure Mean Shifts (No Outliers, No Missing Data)

Example 1 serves as a baseline experiment designed to evaluate breakpoint detection and series reconstruction under ideal conditions, where structural shifts constitute the only source of inhomogeneity. No synthetic outliers or missing observations were introduced, allowing the comparison to focus exclusively on the ability of each method to identify and correct mean shifts.

4.2.1 Breakpoint Detection and Magnitude Analysis

As presented in Table 1 and Fig. 2, all three homogenization frameworks successfully identified the three imposed discontinuities under this ideal scenario. The proposed

BYS-CUSUM framework and RHtestsV4 pinpointed the exact locations of the breakpoints ($t = 70, 180, 300$). Climatol also demonstrated high sensitivity, successfully capturing all three breaks at indices 71, 181, and 299, representing a minor deviation of only ± 1 month from the true locations.

While the Bayesian method directly provides precise magnitude estimates for these shifts with minimal error (absolute error < 0.05), estimating the magnitude for RHtestsV4 and Climatol requires separate post-processing steps (e.g., calculating differences in mean values before and after the flagged change-points).

Table 1: Breakpoint detection, magnitude estimation, and outlier performance for Example 1.

Method	Detected	Magnitude	Outl.	Notes
True	70, 180, 300	+1.5, -2.0, +1.2	0/0	–
BYS-CUSUM	70, 180, 300	+1.54, -1.95, +1.23	1/0	Precise detection
RHtest	70, 180, 300	*(Calc.)*	0/0	Precise detection
Climatol	71, 181, 299	*(Calc.)*	0/0	Precise detection (± 1 month accuracy)

4.2.2 Outlier Detection Analysis

Since no synthetic outliers were introduced in this experiment, the expected outcome is that no observations should be classified as anomalous. Both Climatol and RHtestsV4 correctly reported zero detected outliers. The proposed BYS-CUSUM framework identified one isolated outlier at month 222 (Fig. 3), representing a single false-positive detection.

Although this isolated false alarm slightly increases the reconstruction error, it has negligible influence on breakpoint localization. More importantly, it indicates that the proposed framework exhibits a slight tendency toward conservative outlier detection under nearly noise-free conditions. As demonstrated in the subsequent experiments, this behavior becomes advantageous when genuine outliers are present, where higher detection sensitivity substantially improves breakpoint estimation and overall reconstruction accuracy.

4.2.3 Reconstruction Accuracy

Figure 4 summarizes the reconstruction performance of the three methods using the correlation coefficient, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE).

All methods achieved correlations close to unity, confirming that they successfully recovered the underlying homogeneous series under ideal conditions. RHtestsV4 yielded the smallest reconstruction errors (MAE ≈ 0.05, RMSE ≈ 0.08), followed closely by the proposed BYS-CUSUM framework (MAE ≈ 0.06, RMSE ≈ 0.21). The slightly larger RMSE obtained by BYS-CUSUM is primarily attributable to the single false-positive outlier identified during the reconstruction stage. Climatol exhibited the

largest reconstruction errors (MAE ≈ 0.35, RMSE ≈ 0.46), although its breakpoint detection accuracy remained comparable to the other methods.

Overall, the first benchmark establishes a baseline under ideal conditions, where structural breakpoints are the only source of inhomogeneity. All three methods perform reliably, with only minor differences in accuracy. The following experiments progressively introduce outliers, missing observations, and smaller breakpoint magnitudes to evaluate the robustness of each method under increasingly realistic conditions.

4.3 Example 2: Complex Scenario (Mean Shifts, Outliers, and Missing Data)

The second experiment represents a substantially more challenging and realistic homogenization problem by simultaneously introducing structural breakpoints, high-magnitude outliers, and missing observations. Specifically, the synthetic series contains three imposed mean shifts, five outliers with magnitudes ranging from five to eight standard deviations, and 25 missing observations distributed as ten isolated values and three consecutive five-month gaps. Unlike the baseline experiment, this scenario evaluates the robustness of each method when multiple sources of inhomogeneity coexist and potentially interfere with one another.

4.3.1 Breakpoint Detection and Robustness to Complex Contamination

Table 2 and Fig. 5 demonstrate a clear divergence in the behavior of the three homogenization methods once outliers and missing observations are introduced.

The proposed BYS-CUSUM framework successfully identified all three structural breaks, locating the first and third breakpoints exactly at months 70 and 300, while estimating the second breakpoint at month 174, corresponding to only a six-month deviation despite the presence of nearby missing observations. These results indicate that the Bayesian refinement stage remains stable even when breakpoint candidates are partially obscured by data gaps.

In contrast, Climatol detected only the first two breakpoints and failed to identify the third breakpoint near the end of the series.

RHtestsV4 exhibited a different failure mode. Rather than missing breakpoints, it became susceptible to false detections, introducing an artificial breakpoint at month 5 while simultaneously delaying the detection of the final breakpoint until month 311.

Table 2: Breakpoint detection, magnitude estimation, and outlier performance for Example 2 (True positions: 70, 180, 300).

Method	Detected	Magnitude	Outl.	Notes
True	70, 180, 300	+1.5, -2.0, +1.2	5/5	–
BYS-CUSUM	70, 174, 300	+1.54, -1.90, +1.14	5/5	Precise detection and full outlier recall
RHtest	5, 70, 180, 311	*(Calc.)*	0/5	FP breakpoint at $t = 5$; no outlier routine
Climatol	71, 181	*(Calc.)*	1/5	missed shift at $t = 300$

4.3.2 Outlier Detection and Missing Data Reconstruction

The primary distinction between the proposed framework and the benchmark methods becomes evident in the treatment of anomalous observations. As shown in Fig. 6, BYS-CUSUM successfully detected all five injected outliers and subsequently reconstructed both anomalous and missing observations using the regression model described in Section 3.

By contrast, RHtestsV4 does not include an explicit outlier identification procedure. As a result, the injected outliers remain in the series throughout the breakpoint detection process, increasing the probability of spurious structural changes. This behavior is consistent with the false breakpoint detected at month 5, which occurs immediately after the large outlier introduced at month 2.

Climatol exhibited a different limitation. To ensure stable execution under the high-magnitude anomalies introduced in this example, the outlier tolerance parameter was increased to `dz.max`=9. Although this adjustment prevented execution failure, it substantially reduced the method's sensitivity to anomalous observations. Consequently, only one of the five injected outliers was identified, while the remaining anomalies persisted throughout the homogenization process. In particular, the two untreated outliers surrounding the final breakpoint (months 291 and 311) likely masked the structural change at month 300, explaining why Climatol failed to detect the third breakpoint.

4.3.3 Reconstruction Accuracy

Figure 7 summarizes the reconstruction accuracy achieved by the three methods. Unlike the baseline experiment, substantial differences emerge once outliers and missing observations are introduced.

The proposed BYS-CUSUM framework achieved the smallest reconstruction errors (MAE $\approx$ 0.16, RMSE $\approx$ 0.42), indicating that accurate outlier identification combined with Bayesian breakpoint estimation substantially improves the recovery of the underlying homogeneous series. RHtestsV4 showed a considerable increase in reconstruction error (MAE $\approx$ 0.41, RMSE $\approx$ 1.70). The false breakpoint detected at the beginning of the series together with the delayed detection of the final breakpoint produced persistent reconstruction errors throughout the adjusted series.

Climatol produced the largest reconstruction errors (MAE $\approx$ 0.89, RMSE $\approx$ 1.80). Missing the third breakpoint introduced a systematic bias over the remaining portion of the record, while the limited detection of anomalous observations further degraded the final reconstruction.

Overall, the second benchmark demonstrates that robust performance under simultaneous outliers and missing observations requires reliable discrimination between structural breakpoints and transient anomalies. BYS-CUSUM maintained accurate breakpoint detection while correctly identifying outliers and reconstructing missing observations. By contrast, RHtestsV4 produced false breakpoints, whereas Climatol lost sensitivity after relaxing its outlier rejection threshold to ensure numerical stability.

4.4 Example 3: Low-Magnitude Shifts with High Noise and Outliers

The third benchmark investigates a substantially more challenging situation in which the magnitudes of the imposed breakpoints are reduced while maintaining the same level of contamination introduced in Example 2. The three structural shifts remain located at months 70, 180, and 300, but their magnitudes are reduced to +1.0, -1.5, and +0.7, respectively. Such conditions closely resemble real climatological records, where gradual instrument degradation or subtle observational changes may produce weak discontinuities that are difficult to distinguish from natural variability, particularly in the presence of outliers and missing observations.

Table 3: Breakpoint detection and magnitude estimation for Example 3 (Reduced shift sizes).

Method	Detected	Magnitude	Outl.	Notes
True	70, 180, 300	+1.0, -1.5, +0.7	5/5	–
BYS-CUSUM	63, 178, 295	+1.02, -1.31, +0.50	5/5	High sensitivity
RHtest	70, 182, 298, 303	*(Calc.)*	0/5	FP at $t = 303$
Climatol	None	–	0/5	Failed to detect shifts and outliers

4.4.1 Breakpoint Detection under Weak Structural Signals

Table 3 and Fig. 8 illustrate the behavior of the three methods when the SNR is substantially reduced.

Despite the weaker discontinuities, the proposed BYS-CUSUM framework successfully detected all three structural changes, with estimated breakpoint locations (63, 178, and 295) remaining close to the true positions (70, 180, and 300).

RHtestsV4 correctly identified the first two breakpoints but divided the final structural change into two adjacent breakpoints located at months 298 and 303. This behavior suggests that the algorithm interpreted the local disturbance surrounding the final breakpoint as multiple independent shifts rather than a single structural transition.

Climatol exhibited the largest deterioration in performance, failing to detect any of the three structural breakpoints. This indicates that the method loses sensitivity when breakpoint magnitudes become comparable to the background variability.

4.4.2 Outlier Identification and Its Influence on Breakpoint Detection

Figure 9 further highlights the importance of explicitly separating transient anomalies from permanent structural changes.

The proposed BYS-CUSUM framework successfully identified all five injected outliers, including the observation located at month 299 immediately adjacent to the final breakpoint.

Climatol failed to identify any of the injected outliers. As discussed in Section 4.1, the outlier threshold had to be relaxed to ensure numerical stability, substantially reducing the algorithm's ability to remove anomalous observations. Consequently, the contaminated observations remained in the series throughout the homogenization procedure, further masking the already weak structural changes.

RHtestsV4 does not explicitly identify outliers. Interestingly, the additional breakpoint detected at month 303 occurs immediately after the injected outlier at month 299, suggesting that the algorithm may interpret localized transient anomalies near genuine breakpoints as separate structural changes rather than isolated observations.

4.4.3 Reconstruction Accuracy

Figure 10 summarizes the reconstruction accuracy obtained by the three methods.

The proposed BYS-CUSUM framework consistently achieved the highest correlation with the true homogeneous series together with the lowest MAE and RMSE, indicating the most accurate reconstruction under weak breakpoint conditions.

RHtestsV4 exhibited larger reconstruction errors, primarily because the additional breakpoint detected near the end of the series introduced unnecessary adjustments into the reconstructed record.

Climatol produced the poorest reconstruction performance. Its inability to detect the weak structural changes, together with the presence of untreated outliers discussed in the previous subsection, resulted in persistent reconstruction errors and substantially larger MAE and RMSE values than the proposed framework.

Overall, the third benchmark demonstrates that reducing breakpoint magnitude fundamentally changes the homogenization problem from one of breakpoint localization to one of signal discrimination. Under these conditions, the proposed BYS-CUSUM framework maintains sufficient sensitivity to recover weak structural changes while simultaneously distinguishing them from transient outliers. In contrast, RHtestsV4 tends to over-segment weak signals into multiple breakpoints, whereas Climatol becomes overly conservative and fails to detect either the structural changes or the associated anomalous observations.

4.5 Example 4: Systematic Outliers and Low-Magnitude Shift Detection

The final benchmark represents the most challenging scenario considered in this study. In addition to the weak structural shifts introduced in Example 3, the synthetic series contains "smart outliers", whose magnitudes range between 2.5 and 4 standard deviations of the deseasonalized anomaly series. Unlike extreme outliers, these anomalies closely resemble natural climate variability and are therefore considerably more difficult to distinguish from genuine structural changes. Consequently, this experiment evaluates not only breakpoint detection accuracy but also the ability of each framework to discriminate between subtle transient disturbances and permanent mean shifts.

Table 4: Performance summary for Example 4 (Subtle shifts and smart outliers).

Method	Detected	Magnitude	Outl.	Notes
True	70, 180, 300	+1.0, -1.5, +0.7	5/5	–
BYS-CUSUM	67, 174, 295	+0.84, -1.31, +0.60	4/5	High precision
RHtest	70, 91, 179, 300	*(Calc.)*	0/5	FP at $t = 91$
Climatol	181, 299	*(Calc.)*	1/5	Missed $t = 70$; outl. at 197

4.5.1 Breakpoint Detection in the Presence of Subtle Anomalies

Table 4 and Fig. 11 reveal substantially different responses of the three homogenization methods when structural shifts are accompanied by subtle transient anomalies.

The proposed BYS-CUSUM framework successfully recovered all three structural changes, locating them at months 67, 174, and 295. Although small temporal deviations were observed, all estimated breakpoints remained close to the true locations despite the reduced signal-to-noise ratio and the presence of systematic outliers.

RHtestsV4 exhibited a characteristic over-detection behavior by introducing an additional breakpoint at month 91. Notably, this false breakpoint coincides with two consecutive smart outliers located at months 90 and 91, suggesting that localized transient disturbances were interpreted as a permanent structural change.

Climatol detected only the second and third breakpoints while failing to identify the first breakpoint at month 70. This result indicates that weak structural changes occurring under moderate background variability may remain below the detection sensitivity of the SNHT-based procedure.

4.5.2 Identification of Subtle Outliers

Figure 12 highlights the increasing importance of explicit outlier identification when anomalies become statistically similar to natural climate variability.

The proposed BYS-CUSUM framework correctly detected four of the five injected smart outliers, including the consecutive anomalies at months 90 and 91. Only one subtle outlier located at month 256 remained undetected, most likely because its magnitude was comparable to the local background variability and therefore indistinguishable from ordinary climatic fluctuations.

Climatol detected only one outlier (month 197), leaving the remaining anomalies untreated. Consequently, localized disturbances remained embedded in the series throughout the homogenization process, reducing the method's ability to distinguish weak structural shifts from natural variability.

4.5.3 Reconstruction Accuracy

Figure 13 summarizes the reconstruction performance obtained by the three homogenization methods.

The proposed BYS-CUSUM framework achieved the smallest reconstruction errors, with RMSE and MAE substantially lower than those of both benchmark methods. Although all methods maintained high correlation coefficients owing to the common large-scale climatic signal, the error statistics reveal considerable differences in their

ability to recover the underlying homogeneous series. The reconstruction errors of RHtestsV4 were primarily associated with the additional breakpoint introduced at month 91, whereas the larger errors observed for Climatol resulted from the combined effects of the undetected first breakpoint and the limited identification of subtle outliers. These results indicate that accurate discrimination between transient anomalies and permanent structural changes plays a dominant role in reconstruction quality once breakpoint magnitudes become small.

Overall, the fourth benchmark demonstrates that the principal challenge is no longer detecting breakpoints themselves, but correctly distinguishing weak structural changes from localized transient anomalies. Under these conditions, the proposed BYS-CUSUM framework preserves both breakpoint sensitivity and reconstruction accuracy through robust Bayesian breakpoint verification together with explicit identification and reconstruction of outliers and missing observations. In contrast, RHtestsV4 tends to interpret clustered anomalies as additional structural changes, whereas Climatol exhibits reduced sensitivity to weak breakpoints under complex contamination. These findings highlight the advantage of combining robust probabilistic breakpoint verification with explicit reconstruction of outliers and missing observations within a unified homogenization framework.

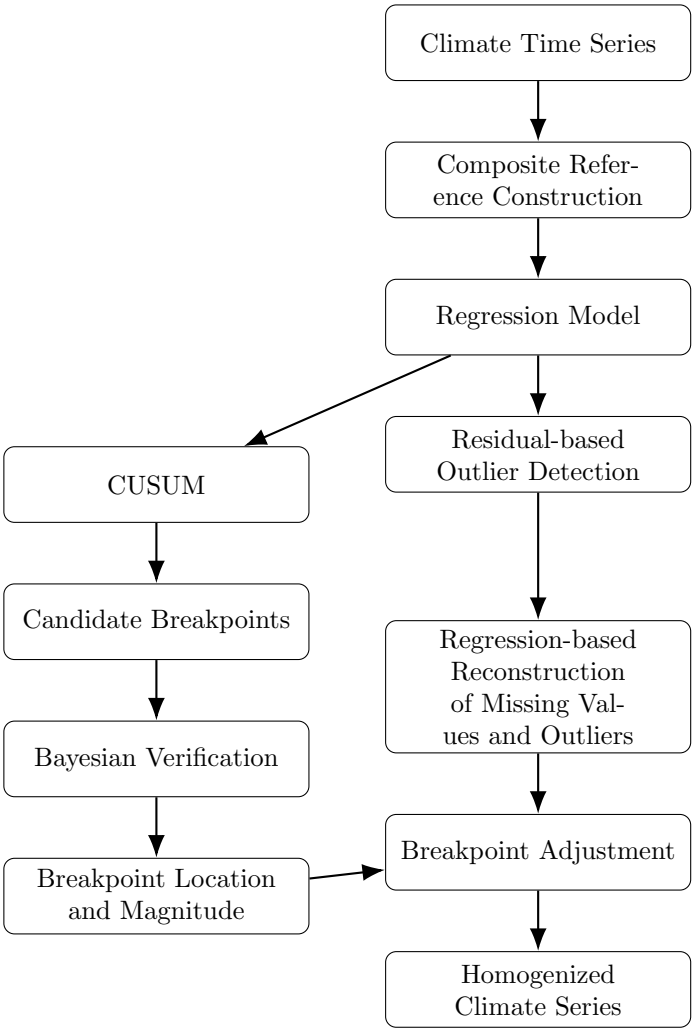


Fig. 1: Workflow of the proposed Bayesian CUSUM (BYS-CUSUM) framework. Regression residuals are first obtained from the regression model and subsequently used in two parallel branches. The left branch identifies candidate breakpoints using the CUSUM transformation followed by Bayesian verification to estimate breakpoint locations and shift magnitudes. The right branch detects transient outliers directly from the regression residuals. After breakpoint verification, detected outliers and missing observations are reconstructed using the regression model, and the validated breakpoint adjustments are finally applied to produce the homogenized climate series.

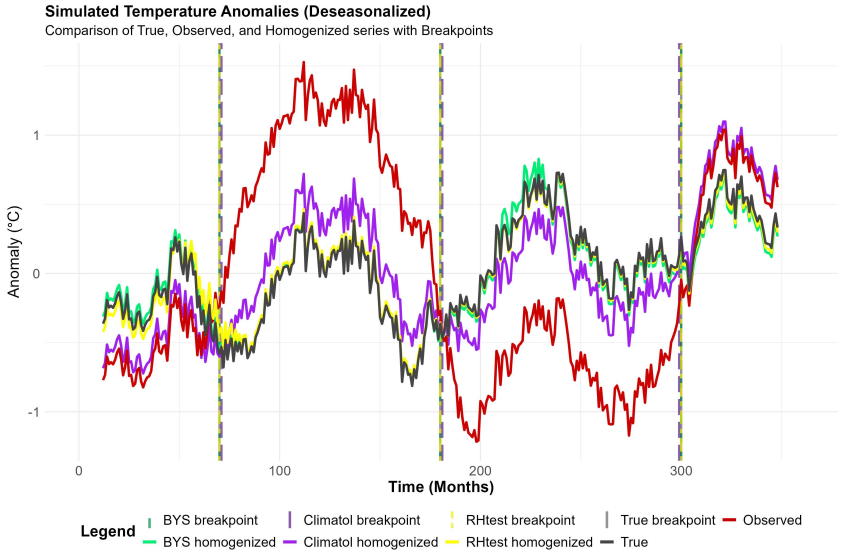


Fig. 2: Visual comparison of homogenized temperature anomalies for Example 1. The reconstructed series from the BYS-CUSUM framework, RHtest, and Climatol all closely follow the True series, demonstrating their high baseline accuracy in the absence of noise and missing data.

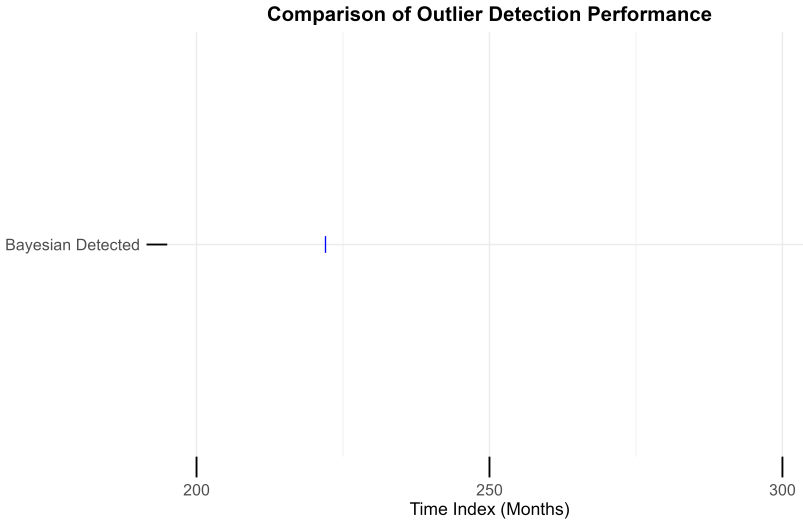


Fig. 3: Outlier detection performance comparison for Example 1. The BYS-CUSUM framework flagged a minor false outlier at $t = 222$, while Climatol correctly remained insensitive.

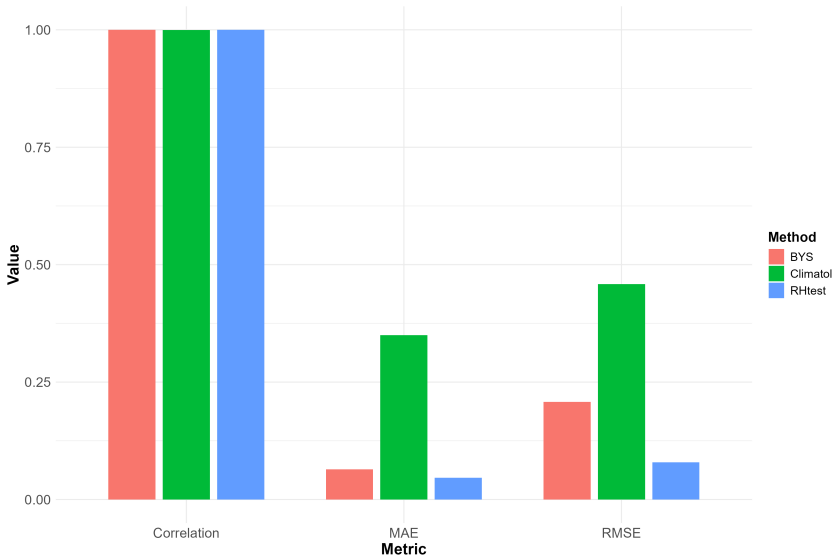


Fig. 4: Benchmark metrics comparison for Example 1: BYS-CUSUM (red), Climatol (green), and RHtestsV4 (blue). All three methods exhibit low error metrics and high correlations, confirming their efficacy in reconstructing undisturbed climate series under ideal experimental conditions.

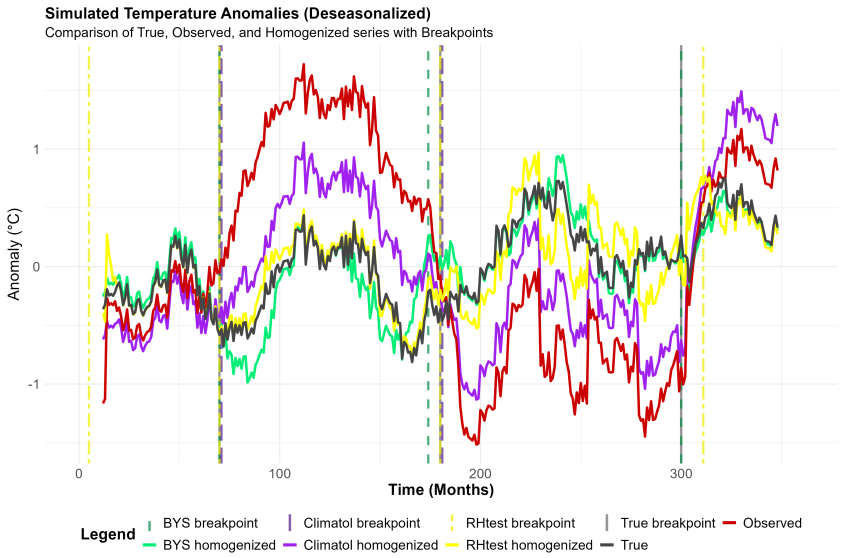


Fig. 5: Visual comparison of homogenized temperature anomaly series for Example 2. The reconstructed series from the proposed BYS-CUSUM framework (green) exhibits near-perfect agreement with the true series (dark grey). In contrast, RHtestsV4 (yellow) introduces early artifacts due to a false breakpoint at index 5 and progressively deviates from the true series because of inaccurate breakpoint placement, while Climatol (purple) shows severe systematic deviations, particularly in the latter part of the series, owing to its failure to detect the final breakpoint at index 300.



Fig. 6: Outlier detection performance comparison for Example 2. The BYS-CUSUM framework successfully flagged all five true outliers, Climatol captured only the major outlier at index 2.

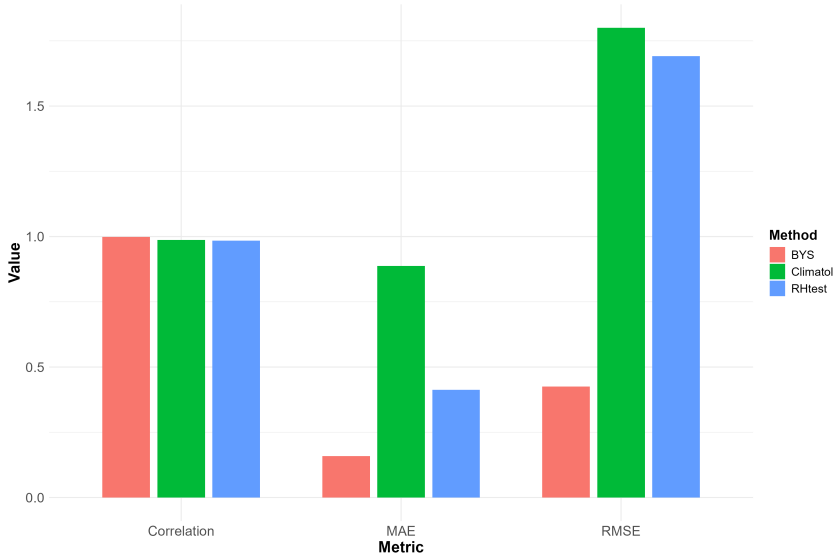


Fig. 7: Quantitative performance metrics for Example 2. The BYS-CUSUM method (red) achieves the highest correlation and the lowest MAE/RMSE across all benchmarks, followed by RHtestsV4 (blue). Climatol (green) exhibits lower accuracy metrics due to the undetected third mean shift.

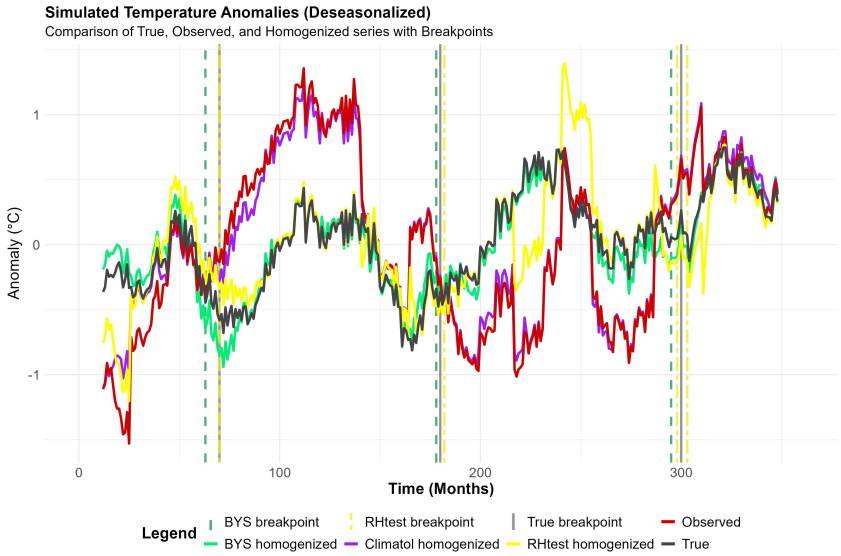


Fig. 8: Comparison of homogenized series for Example 3. The BYS-CUSUM homogenized series (green) shows the best alignment with the True series (dark grey), effectively removing the artifacts of mean shifts and outliers.

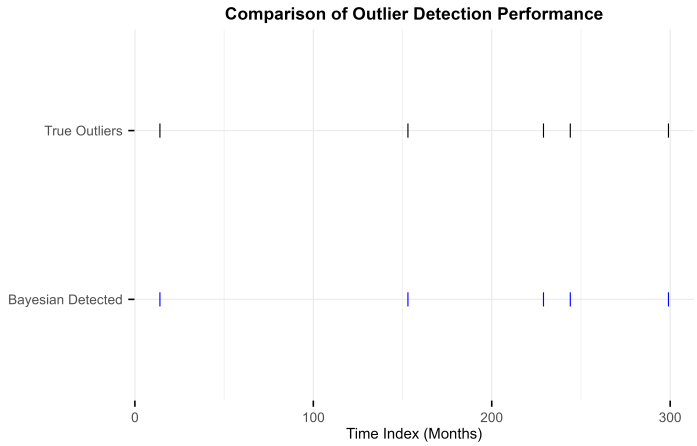


Fig. 9: Outlier detection performance. BYS-CUSUM (blue) matches the True Outliers (black) perfectly.

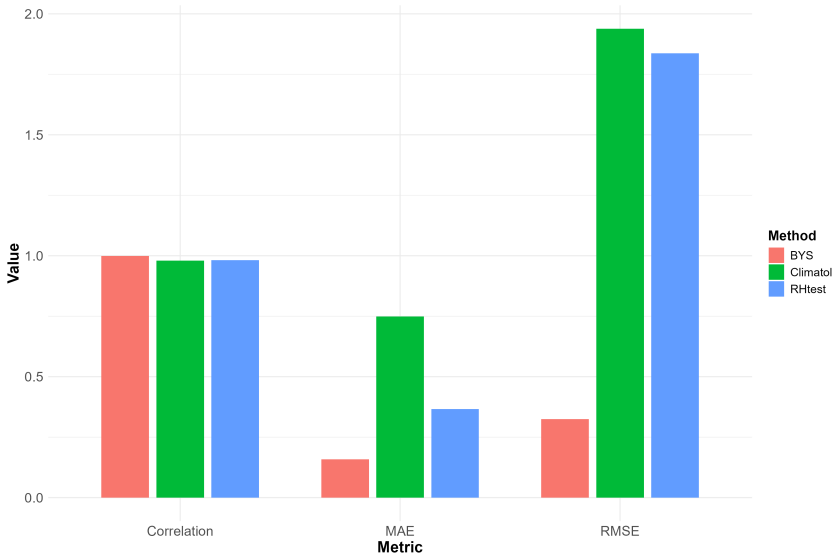


Fig. 10: Error metrics comparison. BYS-CUSUM shows the lowest RMSE and MAE.

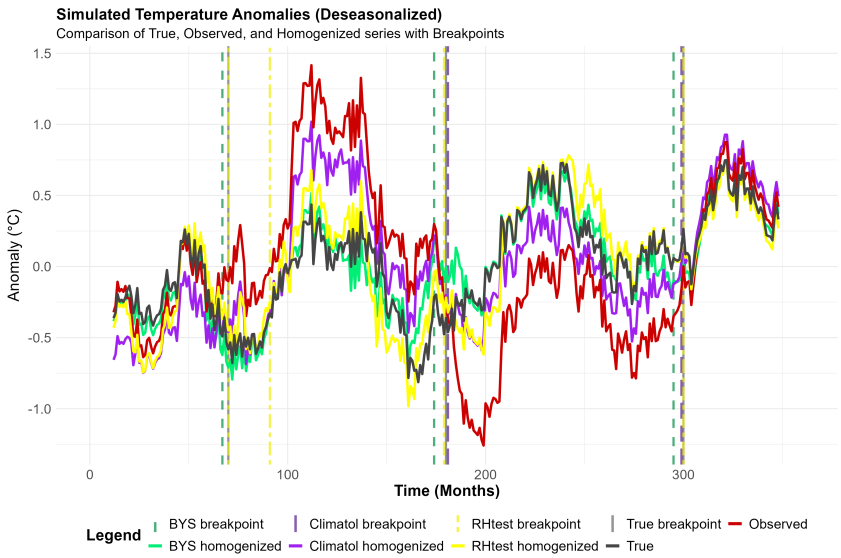


Fig. 11: Comparison of homogenization results for Example 4. The BYS-CUSUM framework successfully avoids the false-positive breakpoint triggered by RHtest at $t = 91$ due to transient outliers, while accurately capturing all true shifts. In contrast, Climatol misses the early subtle shift at $t = 70$.

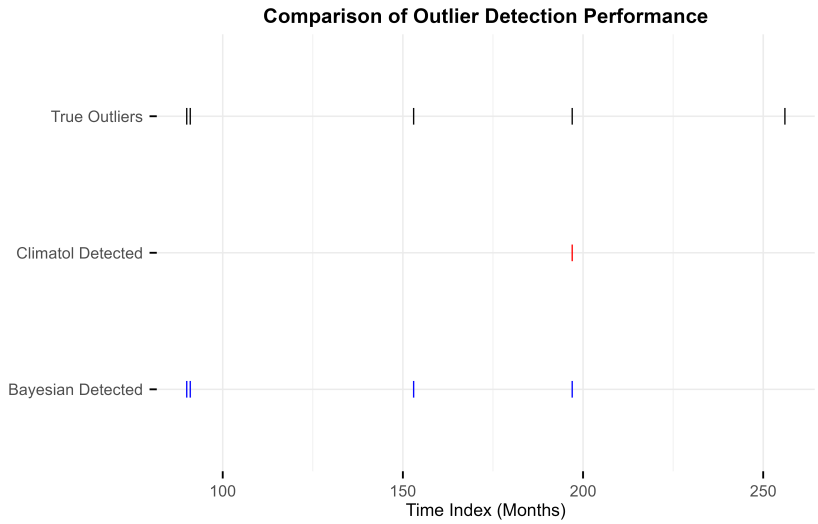


Fig. 12: Outlier detection for Example 4. BYS-CUSUM correctly identifies subtle anomalies at $t = 90, 91, 153, 197$.

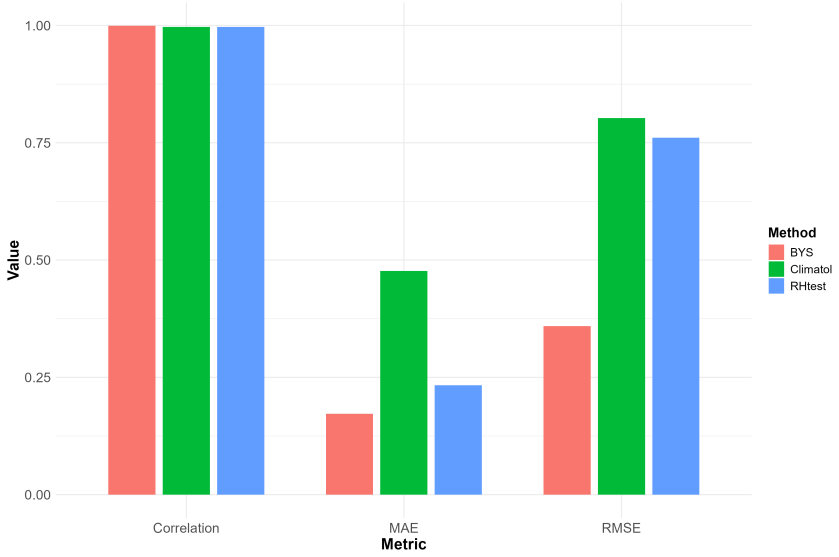


Fig. 13: Quantitative error metrics. Despite the subtle nature of the shifts, BYS-CUSUM achieves the lowest RMSE.

5 Discussion

The experimental results consistently demonstrate that successful climate time series homogenization depends not only on identifying structural breakpoints, but also on correctly distinguishing permanent shifts from transient anomalies before the adjustment process. This distinction becomes increasingly challenging as breakpoint magnitudes decrease and observational series become contaminated with outliers and missing values. Under these realistic conditions, errors associated with transient anomalies may lead to missed discontinuities, false-positive detections, or inaccurate reconstruction of the homogeneous climate record.

The comparative analysis reveals two fundamentally different failure mechanisms in the benchmark methods. RHtestsV4 primarily exhibits an over-segmentation behavior. Since transient anomalies remain embedded in the series during breakpoint estimation, localized clusters of outliers may be interpreted as genuine structural changes. This behavior is evident in the experiments where outliers located close to true breakpoints generated additional changepoints or displaced breakpoint estimates. Such sensitivity indicates that the statistical tests employed by RHtestsV4 become vulnerable when short-lived anomalies locally modify the mean structure of the series.

In contrast, Climatol exhibits an under-detection behavior under contaminated conditions. Although its preliminary outlier filtering improves numerical stability, the method relies strongly on user-defined rejection thresholds. Relaxing the outlier threshold to avoid execution failure inevitably reduces the algorithm's sensitivity to both anomalous observations and weak structural shifts. Consequently, increasing contamination progressively masks subtle breakpoints, while decreasing the threshold may remove excessive neighboring observations and even prevent successful execution. This trade-off illustrates the inherent limitation of sequential preprocessing strategies in which the treatment of anomalous observations directly influences the subsequent homogeneity analysis.

The proposed BYS-CUSUM framework overcomes these limitations through a unified probabilistic framework that combines Bayesian breakpoint verification with explicit reconstruction of outliers and missing observations. A key feature of the proposed model is the adoption of a Student-t likelihood, whose heavy-tailed distribution reduces the influence of extreme observations during breakpoint estimation without requiring their prior removal. Bayesian refinement therefore provides stable estimates of breakpoint locations and magnitudes even under contaminated conditions, while the subsequent reconstruction stage improves the quality of the final homogenized series by replacing detected outliers and missing observations using the regression relationship.

An important consequence of this integrated framework is the substantial improvement in reconstruction accuracy. Across all experimental scenarios, the final imposed breakpoint was intentionally located near the end of the series ($t = 300$ in a 360-month record) to evaluate the robustness of breakpoint detection under boundary conditions. The experimental results indicate that the primary challenge for the benchmark methods was not the boundary location itself, since all methods successfully detected this breakpoint under the noise-free conditions of Example 1. Instead, performance deteriorated when the end-of-series breakpoint coincided with nearby outliers and missing observations or when the breakpoint magnitude became smaller. Under these

conditions, Climatol frequently failed to identify the final breakpoint, whereas RHtestsV4 tended to misplace the breakpoint or split it into multiple nearby detections. In contrast, the Bayesian refinement stage of the proposed framework consistently maintained accurate localization of the final breakpoint across all experiments, demonstrating greater robustness to the combined effects of boundary location, transient anomalies, and reduced signal-to-noise ratio.

Overall, the experimental results indicate that the principal challenge in climate homogenization is not merely detecting changepoints, but reliably separating permanent structural changes from transient observational disturbances under realistic conditions. The proposed BYS-CUSUM framework addresses this challenge through robust Bayesian breakpoint verification, residual-based outlier identification, regression-based reconstruction of outliers and missing observations, and a unified homogenization workflow. Together, these methodological components provide consistently accurate breakpoint localization and superior reconstruction performance across varying levels of noise, missing data, outlier contamination, and breakpoint magnitude, making the proposed framework a reliable alternative for long-term climate variability and climate change studies.

6 Discussion

The experimental results consistently demonstrate that successful climate time series homogenization depends not only on identifying structural breakpoints, but also on correctly distinguishing permanent shifts from transient anomalies before the adjustment process. This distinction becomes increasingly challenging as breakpoint magnitudes decrease and the observational series become contaminated with outliers and missing values. Under realistic observational conditions, transient anomalies may either be falsely identified as structural breakpoints or mask genuine discontinuities, particularly when breakpoint magnitudes are small or occur near the boundaries of the observational record.

The comparative analysis reveals two fundamentally different failure mechanisms in the benchmark methods. RHtestsV4 primarily exhibits an over-segmentation behavior. Since transient anomalies remain embedded in the series during breakpoint estimation, localized clusters of outliers may be interpreted as genuine structural changes. This behavior is evident in the experiments where outliers located close to true breakpoints generated additional changepoints or displaced breakpoint estimates. Such sensitivity indicates that the statistical tests employed by RHtestsV4 are vulnerable when short-lived anomalies locally alter the mean structure of the series.

In contrast, Climatol exhibits an under-detection behavior under contaminated conditions. Although its preliminary outlier filtering improves numerical stability, the method relies strongly on user-defined rejection thresholds. Relaxing the outlier threshold to avoid execution failure inevitably reduces the algorithm's sensitivity to both anomalous observations and weak structural shifts. Consequently, increasing contamination progressively masks subtle breakpoints, while decreasing the threshold may remove excessive neighboring observations and even prevent successful execution. This

trade-off illustrates the inherent difficulty of sequential preprocessing strategies in which outlier removal directly influences the subsequent homogeneity analysis.

The proposed BYS-CUSUM framework overcomes these limitations by combining probabilistic breakpoint verification with a unified regression-based framework for outlier treatment, missing-value reconstruction, and homogenization. Bayesian refinement, together with the Student-t likelihood, provides robust breakpoint estimation under contaminated conditions by reducing the influence of extreme observations during posterior inference. As a result, the proposed framework maintains reliable breakpoint localization and consistently achieves more accurate reconstruction of the homogeneous climate series across a wide range of contamination scenarios.

Another important advantage of the proposed framework is its ability to consistently identify transient outliers across all benchmark scenarios. Unlike RHtestsV4, which does not explicitly distinguish outliers from structural changes, and Climatol, whose outlier detection strongly depends on user-defined thresholds, the residual-based detection procedure successfully identified all injected outliers in every experiment. This capability prevents transient anomalies from remaining in the reconstructed climate record and contributes to the superior reconstruction accuracy observed throughout the experiments.

An important consequence of this integrated framework is the substantial improvement in reconstruction accuracy. Across all experimental scenarios, the final imposed breakpoint was intentionally located near the end of the series ($t = 300$ in a 360-month record) to evaluate the robustness of breakpoint detection under boundary conditions. The experimental results indicate that the primary challenge for the benchmark methods was not the boundary location itself, since all methods successfully detected this breakpoint under the noise-free conditions of Example 1. Instead, performance deteriorated when the end-of-series breakpoint coincided with nearby outliers and missing observations or when the shift magnitude became smaller. Under these conditions, Climatol frequently failed to identify the final breakpoint, whereas RHtestsV4 tended to misplace the breakpoint or split it into multiple nearby detections. In contrast, the Bayesian refinement stage of the proposed framework consistently maintained accurate localization of the final breakpoint across all experiments, demonstrating greater robustness to the combined effects of boundary location, transient anomalies, and reduced signal-to-noise ratio.

Overall, the proposed BYS-CUSUM framework demonstrates robust performance across a wide range of contamination scenarios involving varying levels of noise, missing data, outlier contamination, and breakpoint magnitude. By combining probabilistic breakpoint verification with a unified regression-based framework for outlier treatment, missing-value reconstruction, and homogenization, the proposed approach consistently provides accurate breakpoint localization and improved reconstruction of homogeneous climate series. These characteristics make BYS-CUSUM a reliable alternative for homogenizing climate time series intended for long-term climate variability and climate change analyses.

7 Conclusion

This study presented a Bayesian CUSUM framework for climate time series homogenization that jointly models structural breakpoints, outliers, and missing observations within a unified workflow. Through a comprehensive evaluation using four synthetic benchmark scenarios with progressively increasing levels of contamination, the proposed framework consistently demonstrated greater robustness than Climatol and RHtestsV4, particularly in the presence of outliers, missing observations, and low-magnitude shifts. Across the most challenging scenarios, the proposed method substantially reduced reconstruction errors while providing more reliable breakpoint localization, accurate identification of transient outliers, and consistent estimation of breakpoint adjustment magnitudes.

An important finding of this study is that breakpoint detection becomes substantially more challenging when structural changes coexist with nearby outliers, missing observations, or weak shift magnitudes. Under these conditions, transient anomalies may either mask genuine discontinuities or be incorrectly interpreted as additional structural changes. This behavior was observed in both benchmark methods, with Climatol frequently failing to detect weak breakpoints and RHtestsV4 occasionally introducing false breakpoints in the vicinity of clustered outliers. In contrast, the Bayesian refinement stage of the proposed framework maintained stable breakpoint localization and accurate reconstruction across all experimental scenarios.

Future work will focus on extending the proposed methodology from single-station homogenization to joint network-wide Bayesian homogenization, allowing simultaneous estimation of breakpoints across multiple stations while explicitly accounting for spatial dependence. Another promising direction is to investigate alternative processing strategies, including performing outlier reconstruction prior to Bayesian breakpoint estimation, and to evaluate the applicability of the proposed framework to other climatological variables, particularly precipitation, where stronger spatial variability and non-Gaussian characteristics present additional homogenization challenges.

References

[1] Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P.Y., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L.A., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, M.J., Førland, E.J., Hanssen-Bauer, I., Alexandersson, H., Jones, P.D., Parker, D.E.: Homogeneity adjustments of in situ atmospheric climate data: A review. International Journal of Climatology **18**, 1493–1517 (1998) https://doi.org/10.1002/(SICI)1097-0088(19981115)18:13⟨1493::AID-JOC329⟩3.0.CO;2-T

[2] Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J.B.R., Maycock, T.K., Waterfield, T., Yelekçi, Ö., Yu, R., Zhou, B.: Climate Change 2021: The Physical Science Basis. Cambridge University Press, ??? (2021). https://doi.org/10.1017/9781009157896

[3] Ribeiro, S., Caineta, J., Costa, A.C.: Review and discussion of homogenisation methods for climate data. Physics and Chemistry of the Earth **94**, 167–179 (2016) https://doi.org/10.1016/j.pce.2015.08.007

[4] Domonkos, P.: Relative homogenization of climatic time series. Atmosphere **15**(8), 957 (2024) https://doi.org/10.3390/atmos15080957

[5] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J.-R., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N.: The era5 global reanalysis. Quarterly Journal of the Royal Meteorological Society **146**(730), 1999–2049 (2020) https://doi.org/10.1002/qj.3803

[6] Aguilar, E., Auer, I., Brunet, M., Peterson, T.C., Wieringa, J.: Guidelines on climate metadata and homogenization. Technical Report WMO-TD No. 1186, World Meteorological Organization (2003)

[7] Venema, V.K.C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Müller-Westermeier, G., Lakatos, M., Williams, C.N., Menne, M.J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Brunet, M., Moberg, A., Demarée, G.R., Vose, R.S.: Benchmarking homogenization algorithms for monthly climate data. Climate of the Past **8**, 89–115 (2012) https://doi.org/10.5194/cp-8-89-2012

[8] Mestre, O., Domonkos, P., Picard, F., Auer, I., Robin, S., Lebarbier, É., Böhm, R., Aguilar, E., Guijarro, J.A., Vertacnik, G., Klancar, M., Dubuisson, B., Stepanek, P.: Homer: A homogenization software. Időjárás **117**(1), 47–67 (2013)

[9] Menne, M.J., Williams, C.N.: Homogenization of temperature series via pairwise comparisons. Journal of Climate **22**(7), 1700–1717 (2009) https://doi.org/10.1175/2008JCLI2263.1

[10] Javanshiri, Z., Pakdaman, M., Falamarzi, Y.: Homogenization and trend detection of temperature in iran for the period 1960–2018. Theoretical and Applied Climatology **146**(1–2), 619–637 (2021) https://doi.org/10.1007/s00704-021-03709-5

[11] Rahmdel, M., Sanaeinejad, S.H., Javanshiri, Z., Mohamadian, A.: Exploratory analysis and in-homogeneity study of temperature and rainfall series of meteorological stations in iran (period 1989–2018). Journal of the Earth and Space Physics **47**(2), 387–407 (2021) https://doi.org/10.22059/jesphys.2021.318011.1007286

[12] Javanshiri, Z., Rahmdel, M., Mohamadian, A.: Detection and analysis of spatial–temporal outliers in wind speed and pressure data from iranian meteorological stations for climate change studies. Climate Change Research **2**(7), 33–48 (2021) https://doi.org/10.30488/CCR.2021.314595.1056

[13] Alexandersson, H.: A homogeneity test applied to precipitation data. Journal of Climatology **6**(6), 661–675 (1986) https://doi.org/10.1002/joc.3370060607

[14] Pettitt, A.N.: A non-parametric approach to the change-point problem. Applied Statistics **28**(2), 126–135 (1979) https://doi.org/10.2307/2346729

[15] Wang, X.L., Wen, Q.H., Wu, Y.: Penalized maximal t test for detecting undocumented mean changes in climate data series. Journal of Applied Meteorology and Climatology **47**(1), 13–28 (2008) https://doi.org/10.1175/2008JAMC1741.1

[16] Wang, X.L., Feng, Y.: RHtestsV4 User Manual. Climate Research Division, Environment Canada, (2013). Climate Research Division, Environment Canada

[17] Guijarro, J.A.: Climatol: Climate Tools (Series Homogenization and Derived Products). R package (2018)

[18] Page, E.S.: Continuous inspection schemes. Biometrika **41**(1–2), 100–115 (1954) https://doi.org/10.1093/biomet/41.1-2.100

[19] Basseville, M., Nikiforov, I.V.: Detection of Abrupt Changes: Theory and Application. Prentice Hall, ??? (1993)

[20] Hawkins, D.M., Qiu, P., Kang, C.W.: The changepoint model for statistical process control. Journal of Quality Technology **35**(4), 355–366 (2003) https://doi.org/10.1080/00224065.2003.11980233

[21] Barry, D., Hartigan, J.A.: A bayesian analysis for change point problems. Journal of the American Statistical Association **88**(421), 309–319 (1993) https://doi.org/10.1080/01621459.1993.10594343

[22] Fearnhead, P.: Exact and efficient bayesian inference for multiple changepoint problems. Statistics and Computing **16**, 203–213 (2006) https://doi.org/10.1007/s11222-006-8450-8

[23] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, 4th edn. CRC Press, ??? (2021)

[24] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of Statistical Software **76**(1), 1–32 (2017) https://doi.org/10.18637/jss.v076.i01

[25] Stan Development Team: Stan Modeling Language Users Guide and Reference Manual. Stan Development Team, (2025). Stan Development Team. https://mc-stan.org

[26] Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., Bürkner, P.-C.: Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc. Bayesian Analysis **16**(2), 667–718 (2021) https://doi.org/10.1214/20-BA1221