Addressing Practical Challenges of Stochastic Process Control for Leakage Detection in Water Distribution Networks: A Comparative Analysis

Article *in* Journal of Water Resources Planning and Management · August 2025

DOI: 10.1061/JWRMD5.WRENG-6969

CITATIONS
0

READS
80

2 authors:

Ella Steins
Technische Universität Berlin
7 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Andrea Cominola
Technische Universität Berlin
71 PUBLICATIONS **1,826** CITATIONS
SEE PROFILE

Addressing practical challenges of stochastic process control for leakage detection in water distribution networks: a comparative analysis

Ella Steins* and Andrea Cominola †

Abstract

Various leakage detection algorithms have been proposed in the literature to pursue prompt leakage identification in water distribution networks using sensor data. Some use stochastic process control (SPC) methods, where the leak identification is considered an anomaly detection problem. However, several practical challenges emerge from SPC in leakage detection, including outliers, random fluctuations, as well as non-normally distributed and autocorrelated data. Here, we contribute a comparative analysis of seven advanced SPC techniques based on cumulative sum (CUSUM) for leakage identification in water distribution networks, selected based on an extensive literature review of approximately 100 publications. We test their usability for leakage detection by integrating them into a state-of-the-art data-driven detection method and demonstrating them on the L-Town benchmark network. Ultimately, this study offers actionable recommendations to guide the selection of change point detection methods for leakage detection: our results indicate that transformations are best combined with robust statistics. Nonparametric and adaptive methods are suited to retrieve reliable alarms under heterogeneous conditions. Lastly, incorporating a decorrelation step improves the detection performance if irregular water demand patterns occur in the network.

*Postdoc, Chair of Smart Water Networks, Technische Universität Berlin, Berlin 10623, Germany; Einstein Center Digital Future, Berlin 10117, Germany (corresponding author). ORCID: https://orcid.org/0000-0002-1615-494X. Email: steins@tu-berlin.de

†Assistant Professor, Chair of Smart Water Networks, Technische Universität Berlin, Berlin 10623, Germany; Einstein Center Digital Future, Berlin 10117, Germany. ORCID: https://orcid.org /0000-0002-4031-4704. Email: andrea.cominola@tu-berlin.de

Keywords— Leakage detection, stochastic process control, CUSUM, GWMA-CUSUM, weighted CUSUM, robust statistics, nonparametric & adaptive control charts, autocorrelation, water distribution network

1 Introduction

Water scarcity is expected to dramatically increase by 2050 due to growing demand, increasing pollution of freshwater sources, and the declining availability of water resources (Boretti and Rosa 2019). Along with infrastructure aging, these represent major challenges to water distribution networks (WDNs) management, where the amount of non-revenue water is already estimated to be 126 billion cubic meters per year worldwide (Liemberger and Wyatt 2019). A substantial part of non-revenue water is due to leakages (Bozkurt et al. 2022), which, beyond the direct economic damage due to the revenue loss, also lead to cascading economic costs due to, e.g., increased operational energy consumption (Liemberger and Marin 2006), as well as to potential public health risk through water contamination (LeChevallier et al. 2003; Fox et al. 2016). Addressing these critical issues reinforces the need for proper solutions to detect and mitigate leaks in WDNs, ensuring more sustainable and resilient water supply.

Considerable research has been conducted in the last decades to advance leakage management by developing various leakage identification and localization methods (Puust et al. 2010). Leak identification aims to detect a leakage occurrence promptly, i.e., with the lowest time to detection (TTD), defined as the difference between its detection time and its start time. Leak localization, in turn, deals with determining the leak position in a WDN. Unlike hardware-based methods, software-based methods enable real-time monitoring of changes in flow and pressure regimes by processing observations from a WDN (Wan et al. 2022). They can be categorized into model-based, mixed data-driven and model-based (also known as hybrid), and data-driven methods (Romero-Ben et al. 2023). Model-based methods require a calibrated hydraulic model of the WDN. While these models are well suited to capture the system topology, its hydraulic processes, and operations, their selection and applicability are restricted due to calibration difficulties arising from the complexity of WDNs (Kim et al. 2016) as well as uncertainties in the considered system parameters (Blesa and Pérez 2018). Mixed approaches aim to circumvent these problems by limiting the usage of hydraulic models to offline processes and combining them with data-driven techniques, including machine learning. Lastly, data-driven methods abstain from sophisticated hydraulic models. Further differentiated into unsupervised, semi-supervised, and supervised methods (Wu et al. 2024), they

directly process WDN flow or pressure measurements (unsupervised), employ prediction-classification techniques or reconstruction error analysis (semi-supervised), or develop supervised classification trained with labeled data. Hence, data-driven methods do not face the requirements needed for calibration of model-based approaches, yet supervised and semi-supervised methods still require training data, and their performance typically depends on the availability and quality of such data (Mounce and Machell 2006). Hence, the requirements of leakage detection and identification methods are not mutually exclusive.

In this work, we focus on statistical process control (SPC) techniques, which are commonly employed by both unsupervised and semi-supervised data-driven methods. SPC techniques use statistics to detect the change point of a process, where it is crucial to differentiate random changes from actual change points (Ali et al. 2016), i.e., points in a time series or data sequence at which the underlying statistical properties undergo a significant shift. In an SPC perspective, the leakage detection problem is regarded as an anomaly detection problem, which has also been described as "the problem of finding patterns in data that do not conform to expected behavior" (Chandola et al. 2009). Based on an estimation of the statistical moments, threshold boundaries are constructed from historical data (direct measurements or reconstruction error time series) and used to distinguish between anomalies (i.e., leakage) and *normal* observations. For the purpose of leakage identification, various SPC techniques have been employed, including straightforward thresholding, Shewart control charts and Western Electric Control rules (Bakker, Vreeburg, et al. 2014; Romano et al. 2017; Farah and Shahrour 2017; Wang et al. 2020; Bakker, Jung, et al. 2014; Kim et al. 2016; Wu, Peng, et al. 2023; Loureiro et al. 2016; Romano et al. 2014), univariate exponentially weighted moving average (EWMA), univariate cumulative sum (CUSUM) control charts (Daniel, Pesantez, et al. 2022; Steffelbauer et al. 2022; Eliades and Polycarpou 2012; Anjana et al. 2015; Ahn and Jung 2019; Jung, Kang, et al. 2015; Jung and Lansey 2015; Misiunas et al. 2006), and multivariate control charts and Hotelling's T^2 (Jung, Kang, et al. 2015; Jung and Lansey 2015; Palau et al. 2012). While SPC methods are overall computationally efficient and do not require a comparable amount of training data to supervised methods, they may be built on unrealistic assumptions (Wan et al. 2022).

Three main practical challenges emerge from the existing integration of SPC methods in leakage detection algorithms – first, data fluctuations. In a comparative analysis of SPC methods, Jung and Lansey (2015) compared different univariate and multivariate methods and found the univariate CUSUM and EWMA charts to be less sensitive to outliers in the data than multivariate methods. However, they are more sensitive to small leaks (Jung and Lansey 2015). Additionally, they discussed that the control charts

rely on the assumption that all data is uncorrelated and comes from the same distribution, which might not hold due to daily fluctuations in the flow. To address this limitation, Palau et al. (2012) proposed to divide the day into different periods due to daily fluctuations in the flow and perform a principal component analysis to further reduce variability. Cumulative integrals of shifted pressure data (Kim et al. 2016), moving averages (Bakker, Vreeburg, et al. 2014; Bakker, Jung, et al. 2014), or moving standard deviations (Farah and Shahrour 2017) are suggested to denoise and smooth variations. Additionally, input variables are often normalized to remove fluctuations (Ahn and Jung 2019).

A second common problem resides in setting the threshold that enables discriminating between anomalous and normal observations (Wang et al. 2020; Kim et al. 2016; Wu, Peng, et al. 2023; Eliades and Polycarpou 2012; Nimri et al. 2023; Soldevila, Blesa, et al. 2016). While a lower threshold improves the detection of small leaks, it might also cause false alarm reporting, i.e., an increase in false positive (FP) rate. Some methods mitigate this issue by employing multiple control charts with varying thresholds (Wang et al. 2020; Bakker, Jung, et al. 2014) or choosing thresholds based on computed statistics for small leaks in historical data sets (Eliades and Polycarpou 2012). Further, an outlier detection or leakage verification step before or after the respective leak detection step is often proposed in the literature (Bakker, Vreeburg, et al. 2014; Wang et al. 2020; Soldevila, Boracchi, et al. 2022) to reduce the effect of outliers caused by sensor drifts, errors, or other anomalies not attributed to leaks.

Finally, most SPC methods assume normally distributed data. While this is rarely the practical case, little work in leakage detection addresses this problem. Buchberger and Nadimpalli (2004) propose a Box-Cox Transformation before comparing the trajectories of the observations to the standard normal plot, and Loureiro et al. (2016) introduce a correction factor for heavy-tailed distributions to the Shewart chart.

In this paper, we address the above practical challenges of SPC methods for leakage detection by systematically and comparatively testing advanced SPC methods for data-driven leakage detection. We select seven representative state-of-the-art CUSUM-based methods, which have been developed in SPC literature over the last decades but have not yet been utilized for leakage detection problems. We selected them based on their potential to address the above-mentioned problems, where we explain the selection procedure in Section 3. This paper presents a novel methodological contribution to leak detection by systematically comparing the set of selected advanced SPC techniques against a set of practical limitations identified from extensive literature, thus offering both theoretical insight and directions for practical use. We integrate the selected CUSUM-based methods as the SPC step for leak identification into LILA

(leakage identification and localization algorithm), a recent state-of-the-art pressure-based algorithm for data-driven leakage identification and model-based localization (Daniel, Pesantez, et al. 2022). In its original version, LILA includes the standard CUSUM method (Page 1954) for leak identification. By demonstrating different SPC methods in combination with LILA, we ultimately aim at improving the overall usability of SPC methods for leakage detection and formulate technical recommendations concerning the following crucial capabilities: (i) avoid requiring historical leakage-free data, (ii) handle outliers, fluctuations, and non-normal data, (iii) increase sensitivity to small and non-constant changes (e.g., small transient leaks), and (iv) reduce the need for heuristic hyperparameter setting. In short, the tested methods include techniques such as weighting and an EWMA-CUSUM combination of the data points to increase sensitivity towards incipient leaks, transformations to approximate normal distributions, robust statistics to handle outliers, nonparametric methods that drop assumptions on the underlying data distribution, adaptive features to detect variable ranges of changes, self-starting schemes to avoid the need of historical data, and decorrelation to limit the influence of fluctuations.

The remainder of this paper is structured as follows: first, we introduce the L-Town network used as a case study in this work and LILA in Section 2. The L-Town network is a benchmark problem for testing leakage detection and localization methods, developed as part of the Battle of Leakage Detection and Isolation Methods (BattLeDim) (Vrachimis et al. 2022). The problems of the original CUSUM method are discussed using the example of employing it within LILA for the benchmark problem, as done in Daniel, Pesantez, et al. (2022). We then present the selected advanced CUSUM methods in Section 3. Within the framework of LILA, these methods are tested on the L-Town WDN and the numerical results are presented in Section 4. For each advanced CUSUM method, standard metrics for leakage identification are quantified, i.e., average time to detection (aTTD), precision, recall, and the F_1-score. Based on these metrics, the methods are compared in Section 5, including discussion of their respective use-cases, hyperparameter settings, and transferability. Lastly, Section 6 intends to offer a takeaway for including SPC methods in the future development of leakage detection and localization methods. Tables with the nomenclature (Table 7) and abbreviations (Table 8) used within this paper are reported in the Appendix (Section A.4).

2 Current limitations of the standard CUSUM

In this work, we identify the major challenges related to the usage of CUSUM-based SPC methods for leak detection by considering the demonstrative case study of the L-Town WDN (Vrachimis et al. 2022)

and LILA for leak identification (Daniel, Pesantez, et al. 2022).

The L-Town benchmark WDN, based on an actual WDN in Cyprus, was introduced as part of the BattLeDim competition to enable objective comparative performance evaluation of methods for the detection and localization of leakage events (see all details in Vrachimis et al. 2022). L-Town is characterized by a total length of 42.6 km with 785 nodes, 905 pipes, one pump, three valves, two reservoirs, and one tank. It consists of three district meter areas (DMAs). While DMAs A and B are each connected to a reservoir, a tank that refills during night is used to supply DMA C. A total of 33 pressure sensors, 82 advanced meter reading (AMR) devices within DMA C, and flow meters at the outlet of each reservoir report data through a Supervisory Control and Data Acquisition (SCADA) system every five minutes. We here focus on the data collected in year 2019, which include a total of 19 leakages (8 abrupt and 11 incipient leakages). Figure 1 shows the layout of the L-Town network with the three DMAs and positions of the 19 leakages. Their respective locations (pipe ID), start time and end time are given in Table 1. If a leak remained unfixed throughout the year 2019, the end time is set to 2019-12-31 23:55.

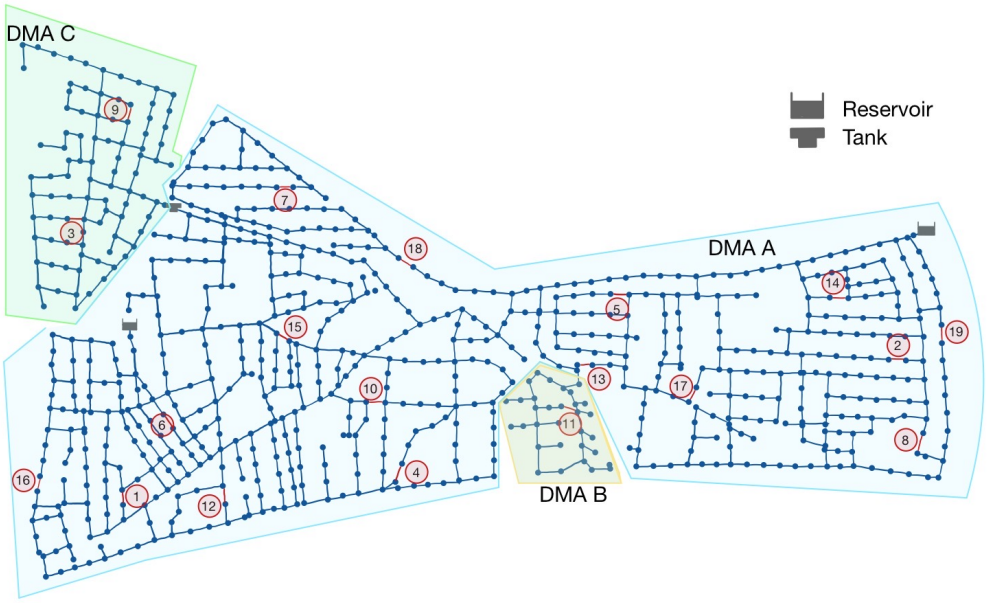

Figure 1: Layout of the L-Town WDN with three DMAs A, B, and C, and the 19 leakages that start during 2019. Figure adapted from Vrachimis et al. (2022).

While different approaches and methods for leak detection and identification have been developed and tested on L-Town as part of the BattLeDIM, data-driven methods are especially valuable for leakage identification and, possibly, localisation if a hydraulic model is not available. Here, we limit the scope of

Table 1: Overview of leaks in BattLeDIM data set

leak ID	pipe ID	start time	end time
1	$p523$	2019-01-15 23:00	2019-02-01 09:50
2	$p827$	2019-01-24 18:30	2019-02-07 09:05
3	$p280$	2019-02-10 13:05	2019-12-31 23:55
4	$p653$	2019-03-03 13:10	2019-05-05 12:10
5	$p710$	2019-03-24 14:15	2019-12-31 23:55
6	$p514$	2019-04-02 20:40	2019-05-23 14:55
7	$p331$	2019-04-20 10:10	2019-12-31 23:55
8	$p193$	2019-05-19 10:40	2019-12-31 23:55
9	$p277$	2019-05-30 21:55	2019-21-31 23:55
10	$p142$	2019-06-12 19:55	2019-07-17 09:25

leak ID	pipe ID	start time	end time
11	$p680$	2019-07-10 08:45	2019-12-31 23:55
12	$p586$	2019-07-26 14:40	2019-09-16 03:20
13	$p721$	2019-08-02 03:00	2019-21-31 23:55
14	$p800$	2019-08-16 14:00	2019-12-31 23:55
15	$p123$	2019-09-13 20:05	2019-21-31 23:55
16	$p455$	2019-10-03 14:00	2019-12-31 23:55
17	$p762$	2019-10-09 10:15	2019-12-31 23:55
18	$p426$	2019-10-25 13:25	2019-12-31 23:55
19	$p827$	2019-11-20 11:55	2019-12-31 23:55

our study on the leakage identification step only, and consider the semi-supervised data-driven method LILA proposed by Daniel, Pesantez, et al. (2022) to demonstrate and discuss SPC techniques for leakage detection. As the leak identification module in LILA is data-driven, it does not require a calibrated hydraulic model of the WDN. The leak identification module of LILA, described in detail in Daniel, Pesantez, et al. (2022), consists of the three steps represented in Figure 2: LILA first estimates pressure values at each time step at all WDN nodes by a linear prediction model, which makes use of the Bernoulli equation to formulate a linear regression between node pairs. This relationship is trained on time series of pressure data referred to normal WDN operations. The published version of LILA does not automate the selection of training periods and sensor selections, which are instead individually chosen for each leak in the published code (Daniel et al. 2021). While this represents a limitation of the published LILA, we consider the same training periods and sensor selection (Daniel et al. 2021) in this work to allow for consistent comparison with the published results. After training of the prediction model, a model reconstruction analysis compares the regression estimates with pressure data. Here, the error between estimated and observed pressure values, i.e., the model reconstruction error (MRE), is computed at each node. At last, LILA performs a change point detection step. It uses the CUSUM control chart to analyze the time series of the MRE (hereafter indicated with $x_1, x_2, \ldots$, where the subscript index refers to the time index, retrieving the starting time of a potential leakage. In this work, we modify this step of LILA (step 3 in Figure 2) by injecting the selected CUSUM-based methods for comparative performance analysis. Hence, the selected methods are tested on the error time series of the 19 leaks in the BattLeDIM data set.

The CUSUM method implemented in the last step of the leak identification module in LILA(Daniel, Pesantez, et al. 2022) was originally introduced by Page in 1954 (Page 1954), and is referred to as the "standard" CUSUM control chart in the following. Let us consider a time series x representing the MRE of

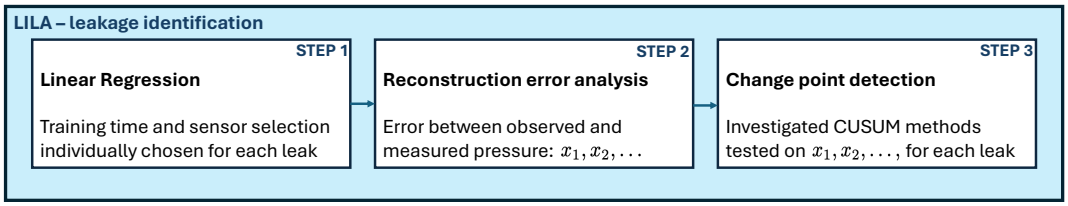

Figure 2: This flowcharts summarizes how the CUSUM methods are tested within LILA. The first two steps of the algorithm, namely the linear regression (step 1) and reconstruction error analysis (step 2), are performed once, resulting in an error time series $x_1, x_2, \ldots$, for each leak. Here, the linear regression is carried out based on leak-wise individual training times and sensor selection, taken from Daniel et al. (2021). In the last step, a change point detection is performed, where the standard CUSUM method is used in the published version of LILA (Daniel, Pesantez, et al. 2022). The selected CUSUM-based methods are integrated in this step by testing them on the error time series $x_1, x_2, \ldots$, of each respective leak in the 2019 BattLeDIM data set.

a selected sensor node. The positive cumulative sum C_i^+ can be calculated at time step (iteration) $i \geq 1$ as follows:

$$C_i^+ = \max\left[0, x_i - (\mu_0 + K) + C_{i-1}^+\right], \tag{1}$$

for given mean μ_0, slack value K, and $C_0^+ = 0$. An alarm is raised if the distance from the target variable exceeds a predefined threshold $H > 0$, i.e., $C_i^+ > H$. In the context of leak detection, only the positive cumulative sum of the distance is computed, as leaks lead to pressure drops and, subsequently, to an increase in the MRE time series. The standard CUSUM method is a repetitious application of Wald's Sequential Probability Test (Wald 1945) with chosen initial score and control bounds (Polunchenko 2016), by testing if the probability density function (pdf) of a series of independent observations $x_1, x_2, \ldots$ has changed from $f_0(x)$ (the in-control distribution) to $f_1(x) \not\equiv f_0(x)$ (the out-of-control distribution). For the test, f_0 and f_1 are assumed to be known and normally distributed, where $f_0 = \mathcal{N}(\mu_0, \sigma_0)$.

In practice, the mean and standard deviation, as well as the hyperparameters needed to set the values of K and H are estimated using an in-control data set. This step is referred to as Phase I, while Phase II refers to monitoring the process to detect changes (Jones-Farmer et al. 2009). Typically, the slack value K and the threshold H are set to be multipliers of the standard deviation of the in-control data, i.e., $K = k \cdot \sigma_0$ and $H = h \cdot \sigma_0$. These formulations require setting hyperparameters k and h. A common approach is setting $k = \frac{\delta}{2}$, where δ is based on the mean shift magnitude $\delta = \mu_1 - \mu_0$ that is to be detected (Moustakides 1986), and either selecting an adequate threshold H based on metrics of the in-control data set or empirical knowledge.

The assumptions required by the standard CUSUM method, however, do not hold true for many

real-world leakage detection cases. More specifically, three limitations may hamper the usage of the standard CUSUM method within data-driven or mixed algorithms such as LILA. The following challenges become particularly evident when using LILA with standard CUSUM on the L-Town WDN benchmark (please note that the test statistics and histograms of the respective leaks mentioned below are shown in Appendix A.3):

1. Known and normally-distributed in-control distribution: the mean and variance of the variable observed for SPC are usually approximated based on an available set of in-control samples. This set might be too small for accurate approximations or might even be not available at all, limiting the informed setting of the slack and threshold values. In Daniel, Pesantez, et al. (2022), the standard CUSUM method computes the mean and variance of the in-control distribution based on the first three days of observations, for instance. However, leak 16 of the benchmark problem starts within 48 hours of the sampled data (Vrachimis et al. 2022). The in-control set needed for the CUSUM methods differentiate from the training set needed for the linear regression in LILA (step 1 in Figure 2), as the in-control set is selected at the beginning of the error time series which is retrieved after the linear regression. Hence, SPC methods that avoid historical data can still be preferable to those that need an in-control set. Further, outliers can affect the performance of the standard CUSUM method, and noise levels vary between the controlled MRE time series. Lastly, the MRE might not be normally distributed at all. The assumption of normally-distributed values limits the applicability of the method to the case of pressure estimates without systematic errors. This limitation is particularly relevant considering that LILA is intended to be used in scenarios in which not all system properties are known. For the benchmark test, for instance, leak 3 and leak 9 lie in a DMA with unknown irregular demands. Consequently, the respective MRE time series do not exhibit normal behavior. In Daniel, Pesantez, et al. (2022), customized CUSUM parameter settings ($k = 0.5$ and $h = 300$) were adopted *ad hoc* for these leaks, as the theshold and slack parameter values used for all other time series ($k = 2$ and $h = 3$) would lead to false alarms in both cases. Additionally, the set of in-control samples was extended to 14 days for approximating the values of mean and variance. This showcases the difficulty of finding suitable global parameters across all MRE time series with the standard CUSUM method. Based on the Kolmogorov-Smirnov test for normality (Massey 1951), leaks 1, 6, 13, and 18 also show non-normal in-control distributions.

2. Known mean change: the standard CUSUM method selects the slack parameter based on a fixed mean change that is to be detected. However, in practice, the magnitude of the mean change may

vary substantially due to different leak sizes. Additionally, incipient leaks do not cause an abrupt change in the mean of the MRE. The h threshold would need to be reduced to shorten the TTD for these leaks. However, a smaller threshold would result in an increase of the overall FP rate, if this threshold was chosen globally for all time series. This further limits the selection of an adequate and generalizable threshold.

3. Autocorrelation: while the standard CUSUM assumes independent observations, some MRE time series are autocorrelated due to the hydraulic of the physical WDN. Again, this is possibly the case, if the forecast model does not account for all actual demands. Based on the Durbin-Watson statistic (Durbin and Watson 1992) of the in-control data, this is indeed the case for four out of 19 leakages in the BattLeDIM benchmark dataset (leaks 3,5,9, and 11). The corresponding test statistics are shown in Appendix A.3.

3 CUSUM-based method selection for comparative analysis

The practical challenges of the standard CUSUM within LILA identified above - i.e., outliers and random variations, non-normal in-control distributions, variable leak size, sensitivity towards small & incipient leaks, and autocorrelation of the data - demonstrate with a specific example the main broader challenges of SPC methods for leakage identification in literature as described in the Introduction: data fluctuations can be either caused by outliers & random variations, and irregular patterns or autocorrelated data might even lead to significantly non-normal distributions. Furthermore, the detection of small & incipient leaks with standrad SPC techniques requires a small threshold; however, finding a suitable threshold is then challenging due to variable & priorly unknown leak sizes.

To address these challenges, we test seven methods. They are summarized in Table 2, along with a short description of their main features and capabilities and literature references. Our selection is based on an extensive literature review of approximately 100 CUSUM (and other SPC) publications. While the full literature review is omitted here as it is not within the primary scope of this paper, and reference is made to other recent comprehensive reviews of SPC methods (Yu and Cheng 2022; Granjon 2013; Saleh et al. 2023; Montgomery 2019), we summarize here our selection procedure and rationale. We primarily limit the approach to univariate CUSUM-based methods, as its challenges are representative for SPC methods in general, and comparison to the standard univariate CUSUM employed in LILA is straightforward. Further, EWMA and CUSUM methods have been previously found superior to other SPC methods for leakage identification (Jung, Kang, et al. 2015), and SPC methods are often used in combination. In particular,

combined EWMA and CUSUM charts are found to be more sensitive towards small shifts than other advanced methods (Lu 2017), and as such a combined chart is employed in this comparison as well (method *gw*). First, we select four SPC methods that address at least one of the identified challenges (methods *t, tr, w, gw* in Table 2). Then, three methods that address multiple challenges are subsequentially tested (methods *adn, ac, corr* in Table 2). The seven methods can be grouped in three main classes, depending on the main challenge they address (non-normal distributions & data fluctuations, sensitivity towards smaller leaks, combinations of multiple challenges). Their rationale and assumptions are described in the following.

Table 2: Summary of the selected CUSUM-based SPC methods for comparative analysis on leak detection.

Acronym	SPC method name	Main features
t	transformed CUSUM (Figueiredo and Gomes 2003)	approximate normality by transforming the data
tr	transformed & robust CUSUM (Figueiredo and Gomes 2003; Nazir et al. 2013)	approximate normality by transforming the data, increase robustness towards outliers by using the trimean
w	weighted CUSUM (Shu et al. 2008)	increase sensitivity by strong weighting of more recent data points, detection of non-constant changes in restricted time window
gw	GWMA-CUSUM (Lu 2017)	increase sensitivity by combining EWMA and CUSUM, more flexible weighting due to design parameter
adn	nonparametric & adaptive CUSUM (Liu et al. 2014)	for any underlying distribution by using sequential ranks, for detecting a range of shifts by recursively estimating the expected mean shift, no in-control data: with self-starting scheme
ac	nonparametric & adaptive CUSUM for arbitrary change (Li 2021)	for any underlying distributiond by using data categorization, for detecting a range of shifts by recursively estimating out-of-control distrbution, no in-control data: self-starting scheme, for arbitrary change by using statistcs for location and scale changes in parallel
corr	nonparametric & adaptive CUSUM for autocorrelated data (Liu et al. 2014; Li and Qiu 2020)	for any underlying distribution by using sequential ranks, for detecting a range of shifts by recursively estimating the expected mean shift, no in-control data: with self-starting scheme, for autocorrrelated data by de-correlation using Cholesky decomposition

Table 3 summarizes the challenges which are addressed by the respective selected CUSUM-based SPC method. The column *increased sensitivity* refers to increased detection sensitivity towards small incipient leaks, either due to weights or consideration of scale changes. Adaptive techniques enable detection of *unknown shift magnitudes*. While robust methods increase the ability to handle *outliers*, fully *non-normal distributions* can be handled by either transformations or non-parametric formulations. Lastly, *autocorrelation* of the data is addressed by a decorrelation step. By use of adaptive and non-parametric statistics, self-starting schemes can be formulated that reduce or even remove the need of in-control data. Similarly, acknowledging both non-normality and autocorrelation reduces the need of heuristic hyperparameter settings due to random fluctuations.

First, to deal with the challenge of non-normal or unknown distributions, *transformations* of the data are generally implemented. While the z-score is commonly used to standarized data, the Box-Cox transformation and Yeo-Johnson can be used to approximate normality. However, the interpretability of

the transformed data might suffer and transformations might be inappropriate to use in case the underlying model exhibits systematic errors. Further, they do not tackle any of the other problems in question. Therefore, transformations are often proposed together with robust statistics with the goal of robustness towards outliers or noise. We here combine the Box-Cox transformation, z-score (Figueiredo and Gomes 2003) (method *t*) and the use of a robust statistic (Nazir et al. 2013) (method *tr*).

Second, *weighted CUSUM methods* are commonly employed to attribute higher weights to recent data, thus improving the sensitivity of CUSUM to gradual changes. Gradual changes may be generated, in our case, by incipient leaks. More recently, mixed GWMA-CUSUM have been shown superior in detecting certain shifts than EWMA or CUSUM methods. Here we test a weighting method designed for fast detection of non-constant mean shifts (method *w*), as well as the GWMA-CUSUM scheme (method *gw*), specifically for their capabilities of detecting incipient leaks. However, both methods fall short in dealing with arbitrary underlying distributions and the setting of the hyperparameters has to be designed for a fixed leak size.

Finally, we test *nonparametric and adaptive methods* to treat both non-normalty and the sensitivity towards small shifts: adaptive methods avoid setting the hyperparameters based on a fixed target mean shift by employing adaptive strategies. Nonparametric (also referred to as distribution-free) methods use nonparametric statistics like sequential ranks and data categorization to detect change points. Hence, they can handle unknown distributions without any assumption of normality. We first test the nonparametric, adaptive method proposed in Liu et al. (2014) (method *adn*). It relies on a self-starting scheme, i.e., it does not require an in-control data set. Additionally, we implement the nonparametric and adaptive CUSUM for arbitrary changes from (Li 2021) (method *ac*). This scheme has similar characteristics to the previous, but additionally tests for scale changes, which are potentially more suitable for detection of incipient leakages. Both methods still build on the assumption of indenpendent observations. The nonparametric and adaptive method *adn* is also combined with a decorrelation technique (Li and Qiu 2020) to treat autocorrelated data (method *corr*).

Table 3: Challenges addressed by each selected CUSUM-based SPC method for comparative analysis on leak detection.

Method acronym	increased sensitivity	unknown shift magnitude	outliers	non-normal distribution	autocorrelated data
t				x	
tr			x	x	
w	x				
gw	x				
adn		x		x	
ac	x	x		x	
corr		x		x	x

4 Numerical results and SPC performance assessment

In the following subsections, all seven selected methods are described in detail, together with their respective performance on the L-Town test case. We here make a note on the hyperparameter setting: the average run length of the in-control distribution, short ARL_0, which refers to the average number of points until a false alarm is given, is commonly used to tune the hyperparametrs of the respective methods and ensure comparability of different control charts. However, using the ARL_0 does not lead to usable results for the leakage detection problem at hand due to violation of the assumptions needed, as the theoretical relation between ARL_0 and FP rate does not hold. An example for the standard CUSUM method is given in Appendix A.2. Yet, it is still possible to find hyperparameter settings which lead to reasonable results. To still ensure comparability, we select the optimal hyperparameters for each method based on a grid search that finds the smallest overall TTD while avoiding FPs. This uses out-of-control data, which is only available if historical leak data is present, corresponding to an empirical hyperparameter setting. As some of the methods are in need of less assumptions, theoretical settings based on in-control metrics become more feasible. This is discussed in Section 5.

The standard CUSUM method, reported in Section 2 and considered here as the baseline method, results in FP for both leaks 3 and 9 of the Benchmark problem (global hyperparameters $h = 3$ and $k = 2$, and an in-control data set of 3 days). Individual TTD for each respective leak are listed in Figure 3 (CUSUM). For each method, the chosen hyperparameter settings related to the presented results in Figure 3 are summarized in Table 9 in the Appendix.

leak ID	CUSUM	method t	method tr	method w	method gw	method adn	method ac	method corr
1	0 days 00:05	0 days 00:00	0 days 00:10	0 days 00:05	0 days 00:05	0 days 17:00	0 days 10:00	0 days 06:35
2	0 days 00:00	0 days 00:00	0 days 00:20	0 days 00:05	0 days 00:05	0 days 17:20	0 days 10:20	0 days 06:45
3	FP	FP	4 days 18:25	FP	4 days 20:30	4 days 23:35	6 days 08:55	5 days 06:45
4	9 days 06:30	21 days 05:15	15 days 19:30	21 days 14:30	16 days 00:15	17 days 10:55	16 days 23:20	15 days 18:45
5	0 days 02:00	0 days 02:05	0 days 02:45	0 days 04:25	0 days 03:35	0 days 14:05	0 days 17:25	1 days 15:15
6	0 days 00:10	0 days 00:30	0 days 00:40	0 days 09:20	0 days 01:00	0 days 19:40	0 days 19:45	0 days 07:10
7	0 days 00:00	0 days 00:00	0 days 00:15	0 days 00:05	0 days 00:15	0 days 22:50	0 days 08:45	0 days 07:00
8	28 days 02:45	FN	49 days 21:45	FN	54 days 13:15	57 days 01:10	49 days 09:20	53 days 19:50
9	FP	FP	FP	FP	17 days 13:45	38 days 11:30	46 days 01:40	59 days 23:40
10	0 days 00:00	0 days 00:00	0 days 00:35	0 days 00:50	0 days 00:30	0 days 17:15	0 days 08:50	0 days 06:45
11	0 days 01:05	0 days 14:05	0 days 08:45	0 days 01:05	1 days 17:15	0 days 18:40	1 days 06:35	0 days 14:00
12	11 days 08:00	14 days 19:10	12 days 17:15	17 days 16:50	11 days 23:00	13 days 18:45	14 days 07:20	11 days 19:55
13	11 days 10:00	30 days 08:25	26 days 05:45	32 days 13:25	20 days 05:55	25 days 22:35	21 days 01:40	19 days 17:05
14	4 days 20:30	5 days 18:15	4 days 20:50	6 days 20:10	4 days 08:00	5 days 05:55	4 days 23:20	4 days 06:10
15	40 days 14:05	51 days 13:20	40 days 14:15	62 days 01:05	50 days 22:15	46 days 02:00	44 days 18:30	43 days 05:40
16	31 days 04:05	40 days 20:30	23 days 10:25	42 days 03:20	31 days 08:15	34 days 04:40	30 days 19:20	30 days 11:30
17	16 days 03:15	19 days 01:40	16 days 04:25	19 days 04:00	16 days 06:00	17 days 01:30	16 days 17:00	16 days 06:55
18	0 days 00:15	0 days 00:45	0 days 00:25	0 days 10:45	0 days 01:00	0 days 16:45	0 days 13:50	0 days 05:28
19	1 days 20:15	15 days 02:30	13 days 22:10	18 days 02:45	12 days 20:30	10 days 17:25	10 days 06:50	9 days 01:20

Figure 3: Time to detection (TTD) for the standard CUSUM, used as a baseline, and the seven selected SPC methods. Each method is assessed on the 19 leaks included in the BattLeDIM dataset for the L-Town WDN (Vrachimis et al. 2022).Cell color is proportional to the TTD (the darker the color, the longer the TTD). Labels FP and FN indicate the presence of false positive (i.e., false alarms) or false negative (i.e., missed leaks) occurrences.

Transformed & robust CUSUM (t & tr)

To deal with non-normal distributed data, transformations have been suggested, e.g. (Figueiredo and Gomes 2003; Hamasha et al. 2022; Peterson 2021), including the Box-Cox and Yeo-Johnson transformation. While the latter can be used on non-positive data, the Box-Cox transformation assumes positive observations, which is achieved by adding a sufficiently large constant. After observing equivalent performance for both transformation methods in preliminary tests, we here focus on the results using the Box-Cox transformation with subsequent z-score standardization, and follow the approach suggested in (Figueiredo and Gomes 2003). The Box-Cox transformation of a variable x is given by

$$x^{BC} = \begin{cases} \log(x+c) & \text{if } \lambda = 0 \\ \frac{(x+c)^{\lambda}-1}{\lambda} & \text{else ,} \end{cases} \tag{2}$$

where x^{BC} is the resulting transformed variable and c is a constant. The variable λ is chosen such that x^{BC} follows a normal distribution. The minimization of the negative log-likelihood function is employed as an optimizer for λ. Accordingly, an in-control data set is needed for the optimization, which is here chosen to be 3 days (except for leak 16, where only the first day is used due to the leak starting on the second day). Hence, the optimal value for λ depends on the in-control data. To increase the robustness

of the estimation of λ, we implement the Bootstrap-estimation approach suggested in Figueiredo and Gomes (2003): the procedure consists of generating 5000 Bootstrap-subsets of size 100, each based on the in-control data, optimizing λ for each subset using the negative log-likelihood, and lastly choosing the mean of all estimates as the final value for λ. More details on the Bootstrap samples can be found in Figueiredo and Gomes (2004). After fixing λ, the Box-Cox transformation is applied to the in-control data, such that its mean μ_{tr} and standard deviation σ_{tr} can be computed. During phase II, the Box-Cox transformation is applied to every new observation x_i. Additionally, the transformed observation x_i^{BC} is then standardized by applying the z-score, i.e.

$$z_i = \frac{x_i^{BC} - \mu_{tr}}{\sigma_{tr}}. \tag{3}$$

As long as no leak occurs, the Box-Cox-transformed and standardized MRE time series is supposed to be standard normally distributed. For $i \leq 1$, the CUSUM statistic for the transformed CUSUM (method t) is then formulated as

$$C_i^+ = \max\left[0, z_i - K + C_{i-1}^+\right], \tag{4}$$

where $C_0^+ = 0$.

As data transformation and standardization do not guarantee robust performance of control charts against outliers, they are commonly employed in combination with robust statistics. These are less affected by outliers than mean and standard deviation. In (Figueiredo and Gomes 2003), the robust statistics total median and total range are used in a Stewart control chart after transformation and standarization of the variable. In order to derive the combined effect of transformation and robust statistics, we here implement the trimean TM for CUSUM, as proposed by (Nazir et al. 2013). The trimean is a weighted average of the mean and two quartiles (Q_1 and Q_3) (Tukey 1977), formulated as

$$\mathrm{TM} = \frac{Q_1 + 2Q_2 + Q_3}{4}. \tag{5}$$

During phase II, the trimean is iteratively computed based on a subset of 5 consecutive samples $\{z_{i-4}, z_{i-3}, z_{i-2}, z_{i-1}, z_i\}$, where the subgroup size corresponds to a tested size in (Nazir et al. 2013). For $i = 4, 9, 14, ...$, we define $j = 1, 2, \ldots$ and

$$C_j^+ = \max\left[0, \mathrm{TM}_j - K + C_{j-1}^+\right], \tag{6}$$

where $C_0^+ = 0$. The slack value $K = k \cdot \sigma_{TM}$ and threshold $H = h \cdot \sigma_{TM}$ depend on the standard deviation of the trimean data, which is estimated using the in-control set. This is summarized as **method tr**.

The original hyperparameter setting of $k = 2$ and $h = 3$ leads to false alarms 3,9,10, and 13 for the transformed method, and to FP for leaks 3, 8, 9, 10, and 13. We present the results for $k = 3$ and $h = 3$ for both statistics in Figure 3 (method *t* and method *tr*). This hyperparameter setting was optimal for both methods. Method *tr* leads to overall comparatively better results than method *t*, as only one FP remains, while all other leaks are detected. Interestingly, the transformed and robust method is able to handle the MRE data of leak 3, which is not the case for the standard CUSUM and method *t*. Hence, the robust statistics seem to have a greater effect than the transformation by itself. This is not surprising: while transformations normalize the data based on the in-control group, they are still sensitive to outliers that appear during phase II (monitoring). The latter are likely for leakage detection problems due to fluctuations in the flow, as well as irregular demands and system operations. If these are not present in the in-control group, their influence cannot be mitigated by transformations, while robust statistics also have an effect during the monitoring phase.

Weighted CUSUM (w)

In many cases, the mean shift is not constant but varies over time. This can occur if residual charts are employed to deal with autocorrelated processes, or if disturbances are compensated by continuous process adjustments (Shu et al. 2008). A possible consequence is the so-called "forecast recovery", which refers to the fast diminishing of the effect of a mean change. Hence, the detection has to happen within a restricted time window. Weighted charts achieve this by attributing higher weight on recent deviations, where a weighting factor w_i is introduced to each increment of the CUSUM chart, i.e., $w_i(x_i - (\mu_0 + k))$.

While the time window for detection is not generally assumed to be limited for leakage detection, incipient leaks lead to non-constant mean shifts. Additionally, the magnitude of the change can be much smaller compared to the case of pipe bursts. Here, we test a weighted CUSUM method (method *w*) introduced by Shu et al. (2008) to increase the performance of the CUSUM chart for incipient leaks. The standard CUSUM method is extended by a weight factor w_i associated to the ith observation of the time series as

$$C_i^+ = \max\left(0, w_i(x_i - (\mu_0 + k)) + C_{i-1}^+\right), \tag{7a}$$

starting with $C_0^+ = 0$, and $w_i = |W_i|$, where

$$W_i = (1 - \alpha)W_{i-1} + \alpha x_i,\ W_0 = 0. \tag{7b}$$

For each time step, W_i is the EWMA-estimator for the mean, with weight hyperparameter $\alpha \in [0, 1]$. By using the EWMA-estimator, more recent data points are given a higher weight, as the weights exponentially decrease for previous samples. The factor $w_i(x_i - (\mu_0 + k)) + C_{i-1}^+)$ in the CUSUM chart then provides a measure of correlation between x_i and the level of mean changes (Shu et al. 2008). Designing this chart includes the setting of three hyperparameters, for which a sequential searching procedure is included in Shu et al. (2008). Seeking a global hyperparameter setting for all 19 processes considered, the setting is derived by minimizing the aTTD while avoiding FPs. An extensive grid search showed that no combination of hyperparameter settings results in the detection of leak 8 while avoiding a FP for leak 11. Further, just as for the standard CUSUM method, using the same hyperparameter settings for leaks 3 and 9 as for all other leakages, leads to FPs in both cases. Hence, we choose a hyperparameter setting which does not detect leak 8, and raises false alarms for leak 3 and 9. The overall TTD for all other leaks is then minimized based on a second grid search, resulting in $H = 3 \cdot \sigma_0$, $K = 2 \cdot \sigma_0$, and $\alpha = 0.75$. The results are listed in Figure 3 (method w).

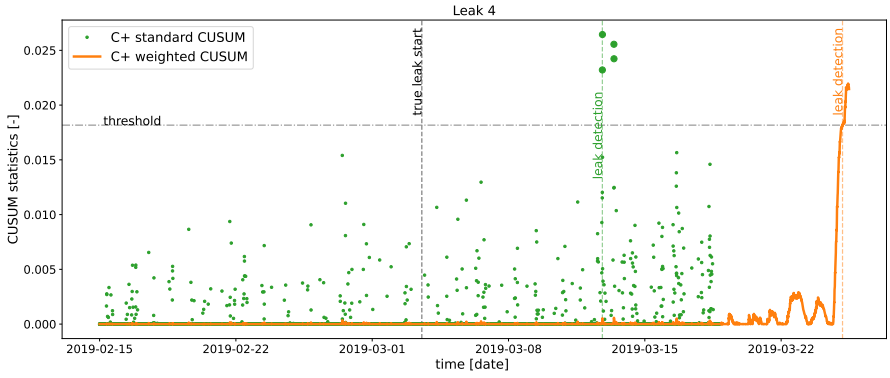

Figure 4: Standard vs. weighted CUSUM: C^+ values are shown both for the standard CUSUM (green dots) and weighted CUSUM (orange line) for Leak 4.

The results show that the weighted method does not solve the problem of non-normal data with high noise (leak 3 and 9). Further, the TTDs is not reduced for the incipient leaks. This is due to the hyperparameter setting. If this was optimized for each respective time series, the corresponding TTDs would be much shorter, in many cases less than a day. However, in order to be useful for the purpose

of leakage detection, an overall hyperparameter setting is desireable. Else, a tuning procedure would be needed on a large set of in-control data for each time series, and due to the violated relation between ARL_0 and the FP rate, the associated setting cannot guarantee a low FP rate. While the capability of the method for small incipient leaks might be underrepresented, this highlights the hyperparameter tuning as one of the crucial downsides. For example, while different value combinations of k and α increase the capability of the method to detect multiple shift sizes, these combinations can only be found through empirical investigation.

On the other side, the method results in more reliable C^+ statistics than the standard CUSUM. This is illustrated for example in leak 4 in Figure 4, which shows the respetive C^+ statistics over time for both the standard CUSUM and w. This type of Figure will be used throughout this work, as it gives insight into the fit of the respective methods for detecting the different leaks and dealing with the different error time series $x_1, x_2, \ldots,$ of the test study. In general, a flat C^+ statistic close to zero before, and a steep and monotonically increasing C^+ statistic after the true leak start show a reliable performance. In controst, C^+-values close to the threshold before the true leak start indicate an increased FP-risk for similar error time series. Further, the TTD is visualized by the horizontal distance between the true leaks start and leak detection. In Figure 4, it can be seen that, while the hyperparameter setting of the weighted methods leads to rather late detection of the leak, it happens to a time for which the C^+ statistic monotonically increases. This leads to a reliable performance regardless of the threshold setting. This is not true for the C^+ statistic of the standard method: Even though an alarm is raised pretty early, this is only due to four values, after which the statistic decreases again. Hence, the performance of the standard method is very sensitive to the selected threshold value. Additionally, the standard C^+ statistic is close to the threshold before the true leak start, decreasing the reliable transferability to new detection problems of the standard CUSUM method.

GWMA-CUSUM (gw)

By proposing the GWMA-CUSUM chart, Lu (2017) combines two approaches to advance the detection of small shifts. The first approach lies in combinations of different charts. One example is the EWMA-CUSUM chart (Abbas et al. 2013), where the EWMA statistic is the input to the CUSUM chart, similar to the weighted CUSUM chart presented here (method w). Additionally, certain shifts can be detected more effectively by introducing a more flexible weighting in the EWMA chart. GWMA charts achieve this

by introducing a design parameter q and adjustment parameter $\tilde{\alpha}$ (Mabude et al. 2021). In Lu (2017), the GWMA statistic Y_i is the input of the CUSUM chart, i.e.

$$Y_i = \sum_{t=1}^{i} \left(q^{(t-1)^{\tilde{\alpha}}} - q^{t^{\tilde{\alpha}}} \right) x_{i-t+1} + q^{i^{\tilde{\alpha}}} \cdot Y_0, \tag{8a}$$

and

$$C_i^+ = \max\left(0, Y_i - (\mu_Y + K) + C_{i-1}^+\right), \tag{8b}$$

where $Y_0 = \mu_Y$ and $C_0^+ = 0$.

First, we define constant slack value $K = k \cdot \sigma_Y$ and threshold $H = h \cdot \sigma_Y$. The hyperparameter setting $\tilde{\alpha} = 0.9$, $q = 0.2$, $k = 1$, and $h = 41$ led to detection of all leaks, but FPs for leak 3 and 9.

In general, the slack and threshold values can be time-dependent based on iterative approximations of mean and standard deviation of the GWMA statistic Y_i. We present the results for an alternative setting of the slack and threshold values: The standard deviation σ_y of the in-control set is substantially higher for some of the processes, including those of leaks 3 and 9, for which a false alarm is raised using $H = h \cdot \sigma_Y$. Therefore, an alternative approach to set the threshold was chosen to amplify the increase of H with increased σ_y, where $H = c_1 \cdot \sigma_y^{c_2}$, and constants $c_1 = 1443.8$ and $c_2 = 1.65$. The results are listed in Figure 3 (method *gw*).

While the GWMA-CUSUM methods allows for a flexible weighting, the hyperparameter setting requires a more intensive tuning. Leaks 3 and 9 could only be correctly detected by the alternative threshold setting, and tuning the constant parameters again requires phase II data. However, more elegant ways to adaptively set the threshold, called adaptive methods, could be a promising way to approach this problem. Additionally, leak 9 should be rather insensitive to outliers, whereas the tuning of the GWMA-CUSUM method is done to be more sensitive towards small shifts. This is not a problem of normality, but it occurs as very differently sized leakages should be detected, while remaining robust towards outliers. Similar to method *w*, method *gw* is not designed for non-normal data with outliers, and this limits the identification of a global hyperparameter setting which allows for an increased sensitivity towards small shifts. For example, increasing q for a fixed $\tilde{\alpha}$ leads to more sensitive statistics, but increases the FP rate across all leaks.

Nonparametric & adaptive CUSUM (adn)

Based on the results of method *gw*, nonparametric and adaptive methods are investigated in the following. Liu et al. (2014) propose a nonparametric & adaptive CUSUM control chart in case both the underlying distribution and the magnitude of shifts is unknown. The chart combines an adaptive approach that estimates the current mean shift $\hat{\delta}_i$ by a modification of the EWMA statistic and the use of standardized sequential ranks R_i^* to circumvent assumptions about the underlying distribution.

Define $R_i^* = \frac{\sum_{t=1}^{i} \mathbb{I}\{x_i \geq x_t\} - (i+1)/2}{((i+1)(i-1))/12}$, then the upper-side CUSUM statistic is given as

$$C_i^+ = \max\left(0, (R_i^* - K_i)/\tilde{h}(K_i) + C_{i-1}^+\right), \tag{9}$$

where the slack value $K_i = \frac{1}{2}\,\hat{\delta}_i$, and $\tilde{h}(K_i)$ is an operating function based on a predefined ARL_0.

The iterative approximation of the current mean shift and use of sequential ranks allow for a self-starting scheme, i.e., no in-control data is needed. Theoretically, threshold H and the operating function $\tilde{h}$, based on ARL_0, can be estimated using an arbitrary synthetic in-control distribution, as non-parametric schemes should exhibit the same ARL_0 properties for any underlying distribution. The current mean shift requires a first estimate of the mean shift, $\hat{\delta}_0^+$, which potentially influences the detection performance.

If no knowledge of shift magnitude is available, $Y_i = \frac{R_i^* + R_{i-1}^*}{2}$, $\hat{\delta}_i = \max(\hat{\delta}_0, Y_i)$ and $\hat{\delta}_0 = 0.7$ are recommended. Here we adopt this setting, and assess the leak detection performance for different ARL_0. The results are shown in Figure 5. No false alarm is raised using this method. Six leakages are not detected for $ARL_0 = 100, 200$, and choosing $ARL_0 = 30, 50$ still results in two undetected leaks. Lowering the threshold H to an empirical value leads to detection of all leaks without raising FPs. All TTD results are shown in Figure 3 (method *adn*). While two leaks (3 and 8) are not detected with an $ARL_0 = 30$ and the respective $H-$value, and are only detected when choosing a low empirical value, the C^+ statistic is very reliable for the chosen method, as no false alarms are raised without hyperparameter tuning based on (potentially unavailable) in-control data. An alarm is raised only after a significant increase of the C^+ statistic after the change point. However, an $ARL_0 = 30$ is lower than typical values. This indicates that the theoretical relation between ARL_0 and FP rate does not hold for the monitored processes. As the proposed CUSUM method assumes independent samples, a possible explanation is the effect of autocorrelation on the true run length. Moreover, outliers and fluctuations of the distribution due to operational variability are common for leakage detection problems.

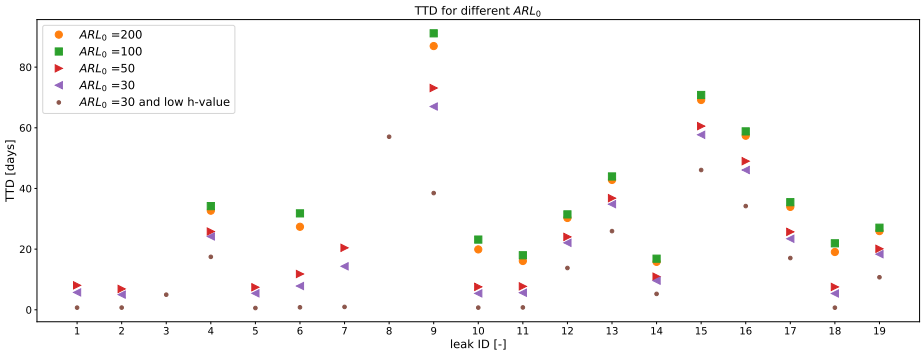

Figure 5: TTD for different in-control average run lengths ARL_0: the represented TTD is obtained using nonparametric and adaptive CUSUM (method adn) for thresholds based on $ARL_0 = 200, 100, 50, 30$ and additionally for a heuristic h-value, for all respective 19 leaks.

Nonparametric & adaptive CUSUM for arbitrary change (ac)

Nonparametric and adaptive methods seem promising as they can detect changes of different magnitude, do not rely on a normality assumption, and do not need calibration data samples. While GWMA-CUSUM searches for changes in the mean, other charts also track scale changes of the distribution, which could be interesting for incipient leaks. Hence, we here test a nonparametric & adaptive method for arbitrary changes (Li 2021) (method *ac*): the method detects increases and decreases in both location and scale parameters. The method includes a built-in post-signal diagnostic function to identify the type of change after an alarm was raised. Instead of assuming underlying in-control and out-of-control distributions, the method employs data categorization based on quantiles of a reference distribution. While the in-control quantiles are sequentially updated, we consider an initial sample of $m_{IC} = 200$ data points (less than a day) to get an initial estimate of the quantiles. The quantiles are used to derive d left-to-right regions for the location change and center-to-outward regions for scale changes, respectively. Based on the regions, a multinomial RV which indicates the categorization information of the current observation. In order to include the ordering information (in time) of the observations, a dependent Bernoulli RV (Z) is computed based on cumulative sums of the multinomial RV which indicates the categorization information. The proposed CUSUM statistic employs the log-likelihood ratio based on Z, i.e., the logarithm of the out-of-control over in-control probabilty of Z. Here, the out-of-control probabilities are derived with a Bayesian estimator, where the prior probability of the out-control-distribution is given by parameters of the Dirchilet distribution (Li 2021). Those can either be obtained by simulation, if prior knowledge of the change point is available, or by simulations using normal distribution.

The regions used for data categorization differentiate the CUSUM statistic for location or scale changes.

Additionally, the upper-side and lower-side statistics use different prior probabilities for the Dirchilet distribution. Hence, a total of four statistics are simultaneously computed, i.e., C_i^{1+} for positive location changes, $C_i^{1-}$for negative location changes, C_i^{2+} for positive scale changes, and C_i^{2-} for negative scale changes. The threshold is set based on ARL_0. However, for $d = 10$ and in-control reference data $m_{IC} = 200$, FPs were rasied for almost all leaks using the theoretical threshold for $ARL_0 = 100$. Hence, an emprical threshold $H = 5000$ was set. An alarm is raised only if the C_1^+ statistic exceeds threshold H. The results are listed in Figure 3 (method *ac*). While the scale changes statistics C_2^+ and C_2^- do not showcase a reliable performance across all leaks, they lead to an earlier detection of some incipient leaks, e.g., leak 8, as can be seen Figure 6.

While monitoring of scale changes seems to be valuable for some leaks, setting of the threshold based on the run length was again not possible for this method. Similar to method *adn*, possible explanations are the presence of autocorrelation, outliers, and fluctuations due to operational variability. Additionally, the prior distribution used to estimate the out-of-control probabilities does effect the performance of the proposed control chart. In particular, it favors the detection of either small or large changes (Li 2021).

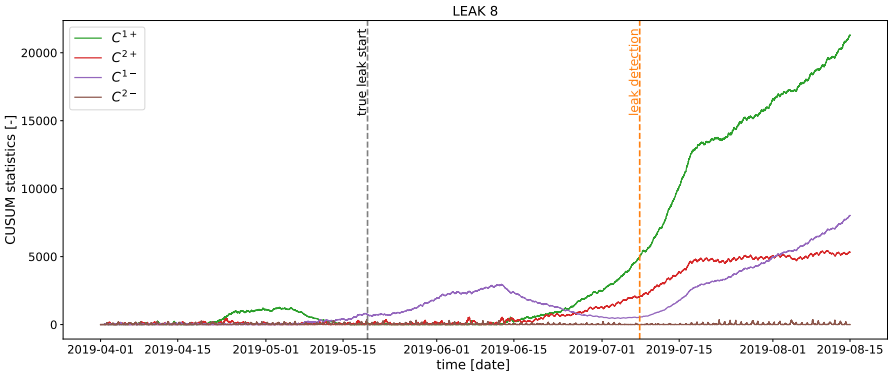

Figure 6: CUSUM Statistics C_1^+, C_2^+, C_1^-, and C_2^- obtained for method ac for leak 8 are shown. The leak is detected on 2019-07-07 when using C_1^+ (vertical orange dotted line), however, the TTD could be much shorter when employing scale statistics.

Nonparametric & adaptive CUSUM for autocorrelated data (corr)

If assumptions of control charts are not met, their performance might be unreliable. In particular, autocorrelation might influence the average run length (George and Box 2000). While applying control charts to residuals of fitted models might mitigate autocorrelation, insufficient models, or fluctuations due to unmodeled operational variability can still lead to autocorrelated MRE time series. Indeed, some of

519 the investigated MRE data can be identified as autocorrelated according to the Durbin-Watson statistic
520 of the in-control data (see Section A.3 for details). In Li and Qiu (2020), a general CUSUM scheme
521 is suggested for autocorrelated data. In a first step, the data is de-correlated based on the Cholesky
522 decomposition (Higham 2009, cf.) under the assumption of stationary covariance. Using a small-to-
523 moderate in-control data set consisting of m_{IC} samples, an initial estimate of mean and covariances. Two
524 observations are assumed to exhibit no covariance if they are more than a maximum number of steps
525 apart, which is given by $b_{max} \in \mathbb{N}^{\geq 1}$. Based on the estimates of mean and covariance, the data is then
526 recursively decorrelated. During phase II monitoring, the estimates are recursively updated, and the
527 current observation is decorrelated with all T_{i-1} previous observations. This number is given by the spring
528 length T_i, i.e. the number of observations between the current time and the last time point at which the
529 C^+ was zero. In short, the Cholesky decomposition recursively produces uncorrelated and standardized
530 samples $x_1^*, x_2^*, ...,$. These samples are then paired with a nonparametric and adaptive method, of which we
531 discussed two approaches, i.e., methods *adn* and *ac* in this work. The latter uses data categorization and
532 an Bayesian estimate of the out-of-control probability which depends on the chosen prior. In comparison,
533 method *adn* requires less assumptions. In the following, the recursive Cholesky decomposition step (Li
534 and Qiu 2020) is combined with the nonparametric and adaptive method *adn* (Liu et al. 2014).

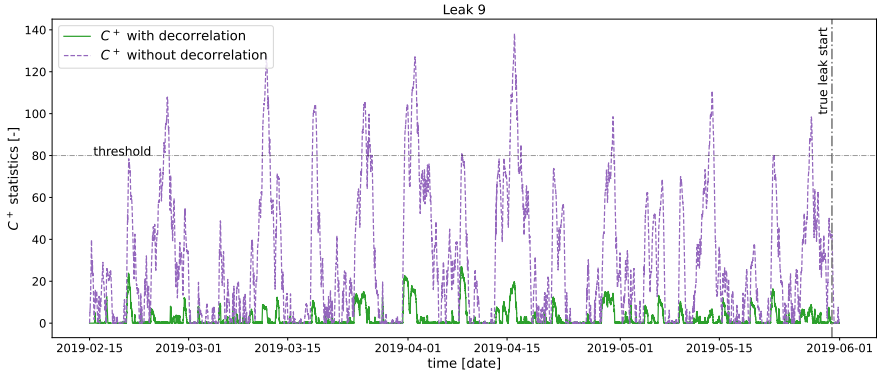

Figure 7: CUSUM Statistic for leak 9: the respective C^+ statistics with (green) and without (purple) decorrelation step is visualized until the true leak 9 starts. The decorrelation step reduces the sensitivity of the statistic due to random fluctuations.

535 The decorrelation does not result in an improved detection for theoretical threshold values based on the
536 ARL_0, possibly as removing the autocorrelation cannot account for all random variations and outliers in
537 the data set. Nonetheless, the decorrelation step decreases fluctuations of the C^+ statistic before the change
538 points. This can be seen in Figure 7, where the respective C^+ statistics with and without decorrelation

step are plotted until the true leak starts. As random fluctuations of the C^+ statistic are significantly smaller when a Cholesky decorrelation is performed, a lower empirical threshold H can be chosen, which in turn reduces the TTD. Nevertheless, the C^+-statistic increases less due to the decorrelation step. This is desireable, if the time series exhibits high fluctuations but not otherwise. The C^+ statistics for all leaks, for which decorrelation was performed, are shown in Figure 8. For all leaks, the C^+-statistisc remain flat before the true leak start due to the decorrelation. After the change point, the statitics still rapidly increase. Nevertheless, this increased might be delayed due to decorrelation, which can be seen for leak 9.

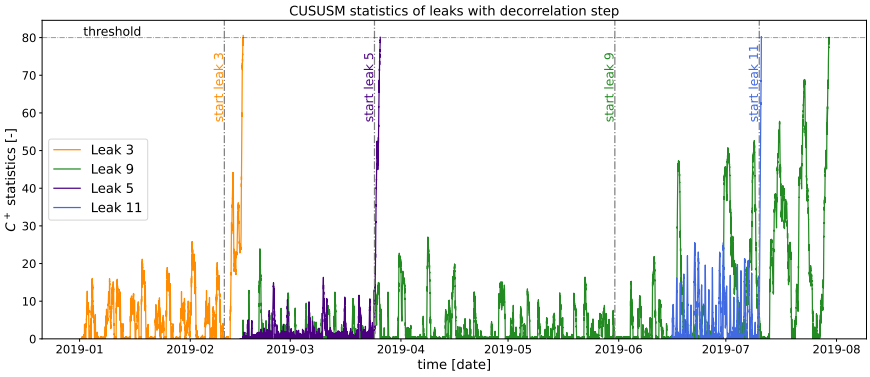

Figure 8: CUSUM statistic of leaks with decorrelation step: the C^+ statistics obtained with method corr are represented for all leaks. An alarm is raised, if the respective statistic exceeds the threshold $H = 80$.

The numerical results for this method are shown in Figure 3 (method *corr*), where a constant threshold $H = 80$ was used across all leaks, and a Cholesky decompostion with $b_{max} = 10$ and $m_{IC} = 200$ was paired with method *adn* for all leaks that exhibit autocorrelation based on the Durbin-Watson statistic. For method *adn*, the same hyperparameters as before are chosen, i.e., $\hat{\delta}_0 = 0.7$ as an initial guess of the mean shift and $ARL_0 = 30$ for the operation function.

5 Discussion and comparative analysis

The seven methods (*t, tr, w, gw, adn, ac, corr*) we selected for comparative analysis represent potential techniques able to address the identified challenges related to SPC for leakage detection, namely unknown & non-normal underlying distributions, robustness towards outliers & random fluctuations, unknown mean change & detection sensitivity towards small or non-constant changes, and autocorrelated data. Here, we evantually compare them based on their leak detection performance, as well as data and training requirements.

First, two techniques are tested to deal with non-normal data, i.e., transformations (method t) and their combination with robust statistics (method tr). While the Box-Cox transformation is one of the few techniques applied to leakage detection (Buchberger and Nadimpalli 2004), results show it does not necessarily mitigate the effect of outliers. In combination with robust statistics, slight detection improvement was made by avoiding a FP for leak 3, however, still no reliable overall detection is achieved across all leaks. Additionally, an in-control data set is needed for selecting an appropiate Box-Cox parameter for the transformation. In practice, no or only a small (and possibly unrepresantative) in-control set might be available. The hyperparameter setting of the slack value and threshold remains challenging, as high variation of the run length can be observed for each leak, along with variable average run length across all processes.

To increase sensitivity towards small (incipient) leaks, we tested a weighted technique (method w) and a combination of CUSUM with GWMA (method gw). However, both methods introduce additional hyperparameters and the subsequent tuning again requires in-control data. Our comparative analysis suggests that these hyperparameters must be tuned individually for each time series to leverage the methods capability for increased detection performance. This in turn may hinder their practical applicability for leakage detection. In order to find a global hyperparameter setting, the threshold setting for method gw was adjusted ad hoc. This led to detection of all leaks, however, transferability to other test cases cannot be guaranteed.

The results of the ad hoc threshold adjustment motivates the use of adaptive techniques, which embed iterative threshold adaptations instead of fixed control limits. A naive approach is the use of multiple CUSUM statistics, which has previously been used for leakage detection (Bakker, Jung, et al. 2014). However, each statistic still requires a fixed control limit, which corresponds to an assumption of a specified constant mean shift magnitude δ. If this constant cannot be properly set due to missing prior information about the mean shift size, more advanced options make use of recursive estimation of the mean shift or out-of-control distribution based on iteratively available phase II process data. At the same time, nonparametric methods avoid the normality assumptions, either by use of sequential ranks or data categorization. Method adn combines sequential ranks and expected mean shift estimation, while method ac employs data categorization and out-of control-probabilities.

For both methods adn and ac, respective global hyperparameter settings could be found that lead to detection of all leaks with average TTD (aTTD) of 14.6 and 14 days. However, both indirectly incorporate

information about a belief of mean shift size, which might impact their practical use cases: in case of *adn*, the adaptive threshold setting requires setting an estimated initial mean shift $\hat{\delta}_0$. Smaller values for $\hat{\delta}_0$ increase detection capabilities of small leaks and vice versa. In case of scant knowledge, the recommended setting $\hat{\delta}_0 = 0.7$ performed well for all leak sizes. In contrast, method *ac* utilizes a Bayesian estimate of the out-of-control probability. Even though the recommended prior in case of scant knowledge produced good results for method *ac*, Bayesian estimates are typically heavily influenced by the chosen prior. An assessment of the overall effect of the prior across all time series is not straightforward. Even though Bayesian CUSUM methods (Javed et al. 2024; Heard and Turcotte 2017; Ali 2020; Bourazas et al. 2023) are not part of this comparison, initial tests of a Bayesian CUSUM method, employing a prior distribution based on in-control data, have shown that the required size of the in-control set needed to form a suitable prior differs substantially for the different leaks of the test case. Nonetheless, assuming availability of a large in-control set, Bayesian CUSUM methods potentially provide an elegant framework to incorporate information of the specific in-control distribution into the CUSUM control chart.

While both *adn* and *ac* adopt a self-starting scheme, $m_{IC} = 200$ initial data points are used as a small in-control group for *ac* to derive an initial estimate of the in-control quantiles. Unlike the other methods, arbitrary change can be detected with method *ac*, referring to change in the location as well as the scale parameter of the distribution. Some results indicate that tracking the scale parameter increases detection performance for small incipient leaks, such as leak 8.

In general, nonparametric methods such as *adn* and *ac* require less hyperparameter tuning. However, the theoretical thresholds are still based on the average in-control run length. For example, the theoretical threshold for method *adn* is derived based on a desired ARL_0 using an arbitrary in-control distribution. Using ARL_0 might lead to practical problems for the leakage detection problem (see, for more detail, Appendix A.2). Indeed, the theoretical threshold for method *adn* led to no detection of some leaks within the simulated time period, while no false alarms were raised. Since false alarms are avoided using the theoretical threshold, one practical solution is using it initially and update it based on emerging historical phase II data, as previously suggested for leakage detection (Eliades and Polycarpou 2012). Indeed, for both methods *adn* and *ac*, global empirical thresholds could be chosen based on the leakage time series data, such that all leakages are detected without raising false alarms.

Nonparametric & adaptive techniques effectively address simultaneously the problem of variable leak

sizes, no or little in-control data, and non-normal distributions. However, methods *adn* and *ac* still assume independence of data points. LILA employs the CUSUM to MRE time series, which is a common method to mitigate autocorrelation. Yet, this approach can have limited reliability in leakage detection and high model errors when unknown water demands occurr in some WDN nodes. For the reasons mentioned above, the implementation of method *adn* is easier for the leakage detection problem. Therefore, it was combined with a decorrelation approach based on the Cholesky-decomposition for all MRE that exhibit autocorrelation, presented as method *corr*. To test the presence of autocorrelation and get an initial estimate of the moments needed for the decorrelation, a small in-control group of $m_{IC} = 200$ samples needs to be employed. In spite of that, the additional decorrelation step resulted in decreased C^+ statistics before the change point. This effect is particularly interesting because it affects those time series whose CUSUM statistic of other methods was subject to large random fluctuations, so that either a false alarm was raised or higher global control limits had to be selected. One example is Leak 9, where results suggest that the decorrelation step increases the reliability of the C^+ statistic, even though it cannot remove all random fluctuations and outliers. As the decorrelation step can be assumed to decrease the initial effect of incipient leaks, the use of the Durbin-Watson statistic to test for autocorrelation is a useful easy-to-implement statistcal test to decide whether the decorrelation step is necessary.

We ultimately complete the above analysis based on TTD with a compilation of various performance metrics for the selected CUSUM-based SPC methods. The results for each method are reported in Table 4, where recall, precision, and the F_1-score, which are standard classification metrics (Lever 2016), are calculated. The definition of the metrics is reported in Section A.1.

Table 4 shows the aTTD and metric results for each method, where the best values are highlighted. The empirical hyperparameter setting limits a complete comparison based on these metrics. For example, while the ad hoc threshold setting of *gw* leads to a detection of all leaks, its transferability to online leakage detection without knowledge of the phase II data is limited. Still, the lower recall, precision, and F_1-scores for the standard method, *t, tr*, and *w* indicate a lower reliability of these methods, compared to more advanced GWMA and nonparametric and adaptive methods. In contrast, methods *adn, ac, corr* result in more reliable performance. The mitigation of random fluctuations of the C^+ statistics through decorrelation increases the reliability of method *corr*. The decorrelation step results in an $aTTD$ of 14.4 days. In fact, decorrelation can lead to a delayed increase of the C^+ statistic. Therefore, it should only be used if high random fluctuations are expected – in practice, this is the case if, e.g., irregular demands are

present. For the L-town case study irregular demands occur in DMA C, and concerns leak 3 and 9 of the 2019 benchmark set. Indeed, the decorrelation step resulted in improved C^+ statistics for the two leakages, as has been shown in Figures 7 and 8. All in all, the use of methods *adn* and *corr* with an initial theoretical threshold that is updated based on historical data, promises the greatest reliability for all identified problems of leakage detection.

The selected CUSUM-based methods of this comparison are tested on the MRE time series of the published LILA code, which, as mentioned in Section 2, relied on manually-selected training times and sensor combinations for each leak in the published code (Daniel, Pesantez, et al. 2022). We have conducted further tests, where the linear regression was carried out using different systematic training times and a variety of sensor combinations. Preliminary results not reported here suggest that *adn*, *ac*, and *corr* are performing robust towards these changes. This further underlines their usability for (semi-supervised) leakage detection methods in real-case scenarios, where training times cannot be chosen manually, and the best sensor combination for detecting a specific leak is unknown a priori.

Table 4: Performance metrics for the selected CUSUM-based SPC methods for comparative analysis.

Method acronym	Method	aTTD [days]	recall [-]	precision [-]	F_1 [-]
	standard CUSUM	**9.1**	**1**	0.89	0.94
t	transformed CUSUM (Figueiredo and Gomes 2003)	12.5	0.94	0.89	0.91
tr	transformed & robust CUSUM (Figueiredo and Gomes 2003; Nazir et al. 2013)	11.6	**1**	0.95	0.97
w	weighted CUSUM (Shu et al. 2008)	13.8	0.94	0.89	0.91
gw	GWMA-CUSUM (Lu 2017)	12.8	**1**	**1**	**1**
adn	nonparametric & adaptive CUSUM (Liu et al. 2014)	14.6	**1**	**1**	**1**
ac	nonparametric & adaptive CUSUM for arbitrary change (Li 2021)	14.0	**1**	**1**	**1**
corr	nonparametric & adaptive CUSUM for autocorrelated data (Liu et al. 2014; Li and Qiu 2020)	14.4	**1**	**1**	**1**

6 Conclusion and recommendations for SPC in leak detection

This work is concerned with the problem of promptly and reliably detecting leakages in water distribution networks based on sensor data. While model-based detection methods are generally accurate, they require a calibrated hydraulic model of the water distribution network, which is not often available. Several data-driven methods have thus been proposed in the literature to abstain from sophisticated hydraulic modeling, many of them relying on statistical process control techniques to detect change points on sensor

data time series, which signify the occurrence of a leak. However, several practical challenges emerge from existing integration of SPC methods in leakage detection, including non-normal distributions, outliers, random fluctuations, and autocorrelated data. Additionally, different change/leak magnitudes and little in-control data for calibration are common for leakage detection. To address these challenges, here we contribute a systematic and comparative analysis of advanced SPC techniques based on cumulative sum (CUSUM) charts for leakage identification in water distribution networks. Starting from an extensive literature review of approximately 100 SPC publications, we select advanced state-of-the-art CUSUM-based methods, integrate them as the SPC step for leak identification into a state-of-the-art the data-driven leak identification method "LILA" (Daniel, Pesantez, et al. 2022), and test their detection performance on the L-Town benchmark WDN released as part of the BattLeDIM (Vrachimis et al. 2022). Based on the results of our comparative analysis, we ultimately formulate recommendations on potential, requirements, and limitations for practical usage of the respective methods in the context of leak detection under different use cases. These recommendations are summarized in Table 5.

Three main outcomes emerge from our analysis. First, increasing detection performance towards small incipient leaks has been highlighted as an important topic for future research (Wan et al. 2022; Wu et al. 2024). In this work, the use of a weighted CUSUM and GWMA-CUSUM method highlights their potential for improved leak detection. However, the FP rate might increase due to sensitivity of the methods towards random changes. Furthermore, both methods introduce additional hyperparameters in comparison with standard or transformation-based CUSUM, thus hampering ease of training and transferability. A possible starting point for methods that combine techniques for robustness and weighting could be the distribution-free mixed GWMA-CUSUM chart (Mabude et al. 2021). Method *ac* in this work includes tracking of scale changes of the distribution, which also showed sensitivity towards incipient leaks.

Second, Wan et al. (2022) also raise the question how to handle spurious outliers. We show that the robust statistics used for method *tr* achieve superior results to the standard CUSUM method, but are not sufficient for all random variations. Furthermore, these methods require an in-control dataset to set the transformation parameters and hyperparameter setting remains challenging. In contrast, the nonparametric and adaptive methods (*adn* and *ac*) show improved performance as they account for different change magnitudes, and adapt to non-normal distributions, guaranteeing better transferability.

If all of the identified SPC problems for leakage detection occur, the most robust results are achieved with nonparatric and adaptive CUSUM for autocorrelated data (method *corr*). It combines nonparametric,

adaptive CUSUM control charts with optional decorrelation. Though not all random variations can be expected to be removed, this lead to decreased fluctuations of the C^+ statistic before the true change point. This results in substantially increased reliability and transferability to new detection problems.

While control charts might require specifying multiple hyperparameters, method *corr* only requires setting the control-limit. Starting from a theoretical threshold based on the average in-control run length, a practical solution involves its adaption based on (emerging) historical data. Further investigations may include techniques used for multivariate data (Woodall and Ncube 1985), and dynamic probability control limits (Steiner et al. 2000; Zhang and Woodall 2015). The latter are developed to mitigate the effect of variable ARL_0 due to rare events. Multivariate data, on the other hand, though not investigated in this work, is commonly monitored by multiple univariate control schemes. If a global threshold is applied, the problem becomes similar to a global threshold for the respective MRE time series in this work.

This work eventually advances water loss management by addressing both theoretical and practical aspects. Theoretically, it fills a gap in the literature as, to our knowledge, no comparative analysis of change point detection methods for leak detection has been conducted. Practically, our findings offer actionable recommendations, guiding the selection of appropriate change point detection methods based on real-world considerations such as data characteristics, performance expectations, and specific boundary conditions.

Table 5: Use-cases, requirements, and limitations of selected CUSUM-based SCP methods for practical usage in leak detection in WDNs.

Method acronym	Use-cases	Requirements	Limitations
t	approximate normality for non-normal distributions	in need of sufficiently large in-control set, hyperparameter tuning for each time series needed	not robust towards outliers, fixed shift size to be detected, not for autocorrelated data
tr	approximate normality, & mitigate influence of outliers	in need of sufficiently large in-control set, hyperparameter tuning for each time series needed	fixed shift size to be detected, not for autocorrelated data
w	detection within restricted time window: increased sensitivity towards small shifts	in need of sufficiently large in-control set, hyperparameter tuning for each time series needed	not robust towards outliers or non-normality, fixed shift size to be detected, not for non-normal & autocorrelated data
gw	increased sensitivity towards small leaks, flexible weighting of data points	in need of sufficiently large in-control set, hyperparameter tuning for each time series needed	not robust towards outliers or non-normality, fixed shift size to be detected, not for non-normal & autocorrelated data
adn	detection of range of shift changes of unknown distribution	with no or little IC-data	not for autocorrelated data, for multiple processes empirical threshold might perform better
ac	detection of range of shift & scale changes of unknown distribution	with no or little IC-data	not for autocorrelated data, for multiple processes empirical threshold might perform better, in-control information might improve Bayesian estimate
corr	detection of range of shifts of unknown distribution for autocorrelated data	with little IC-data	for multiple processes empirical threshold might perform better, decorrelation step increases TTD of transient leaks: should only be used if needed

A Appendix

A.1 Metrics

In the following, the three performance metrics used in this work are formulated (Lever 2016).

Recall is given as the proportion of known positives that are predicted correctly. Precision measures the proportion of true positives (TP) to FP, and the F_1-score is the harmonic mean of Recall and Precision, where $0 \leq F_1 \leq$, and a larger F_1-score indicates better classification. These metrics are formulated as

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{10a}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{10b}$$

and

$$F_1 = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}. \tag{10c}$$

A.2 Note on in-control average run length ARL_0

The in-control average run length ARL_0, defined as the average number of iterations before a false alarm is given based on the in-control data. It is a common metric to compare different control charts, and can be used to chosse an appropiate threshold parameter for a chosen slack value. A common choice is $ARL_0 = 370$ which corresponds to a FP risk $r = \frac{1}{ARL_0} \approx 0.27\%$.

Aside from analytical approximations, there are three main approaches to approximate the ARL_0, i.e. integral equation, Monte Carlo simulation and Markov Chain. Latter, initially proposed by Brook and Evans (Brook and Evans 1972), estimates ARL_0 by discretization of the probability distribution of the C^+ statistic, where the transition probabilities are derived based on an in-control set. The integral equation method (Crowder 1987) derives ARL_0 as a solution of an integral equation, which is solved numerically in most cases. Lastly, the Monte Carlo approach repeatedly generates sequences of charting statistics based on an in-control data set until the control limit is passed to determine the average of the recorded run lengths (Lim and Lee 2024).

For the leakage detection problem at hand, a universal hyperparameter setting across all MRE time series is desirable, as in-control data is not evaluated for each newly occuring leak. However, the values for the ARL_0 associated to a specific hyperparameter setting varies between the processes, making a universal selection of hyperparameters for a constant ARL_0 across all time series infeasable. This is shown in Figure 9, where, using the standard CUSUM method, the ARL_0 is computed for $k = 2$ and various values of the threshold paramter h for leak 1 and 3. We choose the Markov Chain method (Knoth 2021) and two Monte Carlo approaches (Lim and Lee 2024) to estimate the ARL_0. The first Monte Carlo method, labeled MCD, relies on Monte Carlo dropout. This approach drops the run length RL_i computed at simulation iteration i, if no alarm was raised within the generated sequence. However, this is only approximately true, if the drop out rate is not too high. The truncated Monte Carlo approach MCT instead records $RL_i = m_{IC}$ in case no alarm is given, where m_{IC} is the length of the sequence. As no alarms are raise, ARL_0 based on the truncated Monte Carlo approach approaches m_{IC}, and simulatnouesly, the MCD estimate becomes less reliable, as few simulation iterations lead to a recirded run length. In Figure 9, boxplots of the 5000 recoreded run lengths, i.e. the set $\{RL_i\}_{i=1}^{5}000$, for $k = 2$, $m_{IC} = 865$, and variuos h-values are shown. It can be seen, that the number of the raised alarms decreases significantly as larger values for h are chosen, and thus reducing the estimation accuracy. At the same time, the Markov Chain approach also leads to unreliable estimates of ARL_0 in this case, as no transition probability for some states $[0, H]$ of the C^+ statistic can be estimated based on the in-control data.

If the threshold h would be chosen based on $ARL_0 = 370$ for $k = 2$, the three methods (MCD, MCT, and Markov Chain) still produce reliable results (see Figure 9). However, for the standrad CUSUM methods, this leads to false alarms for all leakages. Thus, the FP risk is significantly higher than 0.27%. At the same time, the hyperparameter setting of $k = 2$ and $h = 4$ leads to estimation problems of the ARL_0 as shown above. These robustness issues for practical issues, as well as negative effects the skewness of the run-length distribution are discussed in (Graham et al. 2014). Effects of estimation errors are investigated in (Jones and Steiner 2012). Additionally, there are multiple reasons why the required ARL_0 exceeds the theoretical value, including autocorrelation, non-normality and outliers. Some of the investigated control charts mitigate the influence of these characeristics, as discussed in detail in Section 5. To still showcase the "best" possible performance of each method for this practical problem, an empirical hyperparmeter setting based on TTD and FP was employed in this work. To realise this method in practice, historical

data would be needed to tune the hyperparameters. We discuss this as well in Section 5.

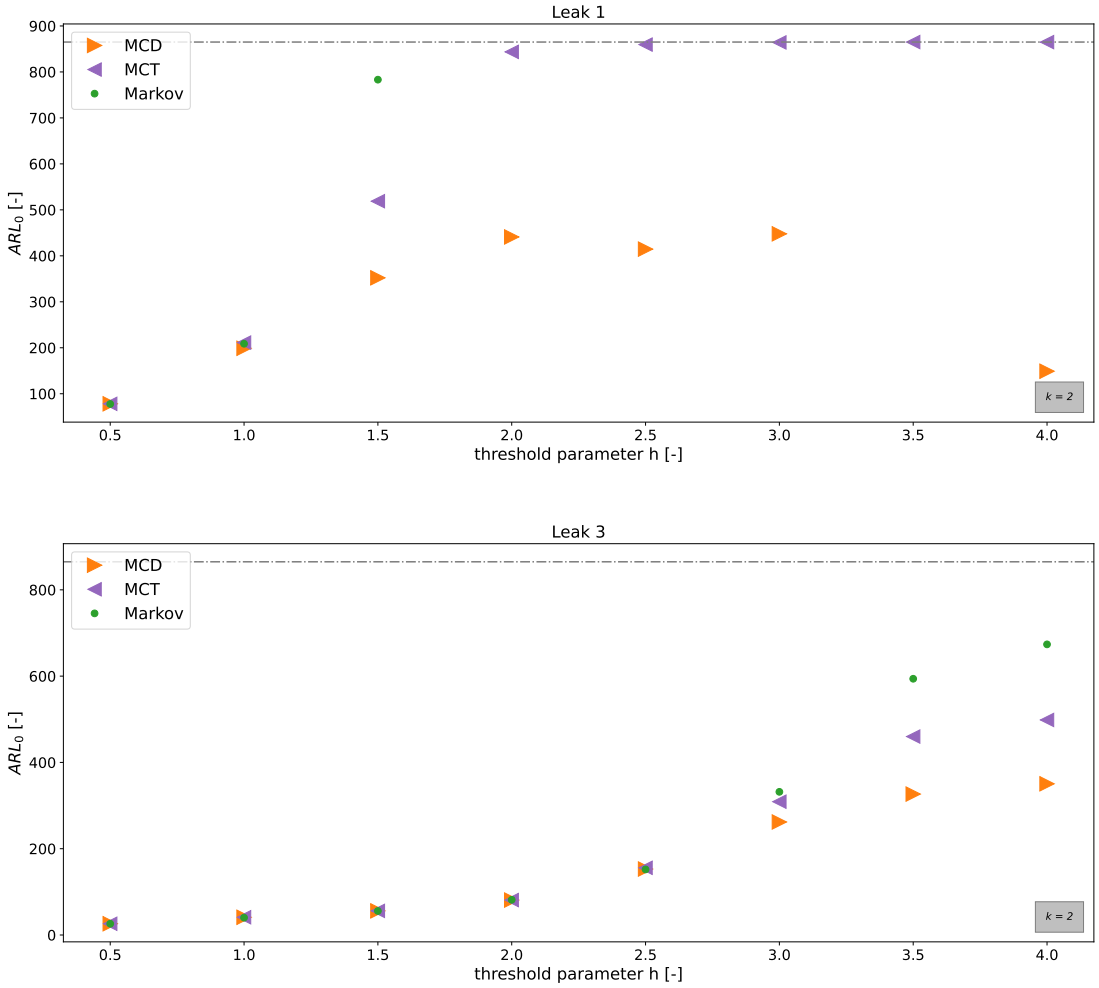

Figure 9: Estimation ARL_0: estimates of ARL_0 using MCD, MCT, and Markov Chain are shown for leak 1 (upper Figure) and leak 3 (lower Figure) are shown. While similar estimation results of the three methods indicate reliable estimation, the respective ARL_0 does not result in avoidance of false alarms: For example, leak 1 requires $h \geq 2.5$ in order to not raise a FP, and leak 3 raises FP across all plotted hyperparameter settings and corresponding ARL_0 estimates.

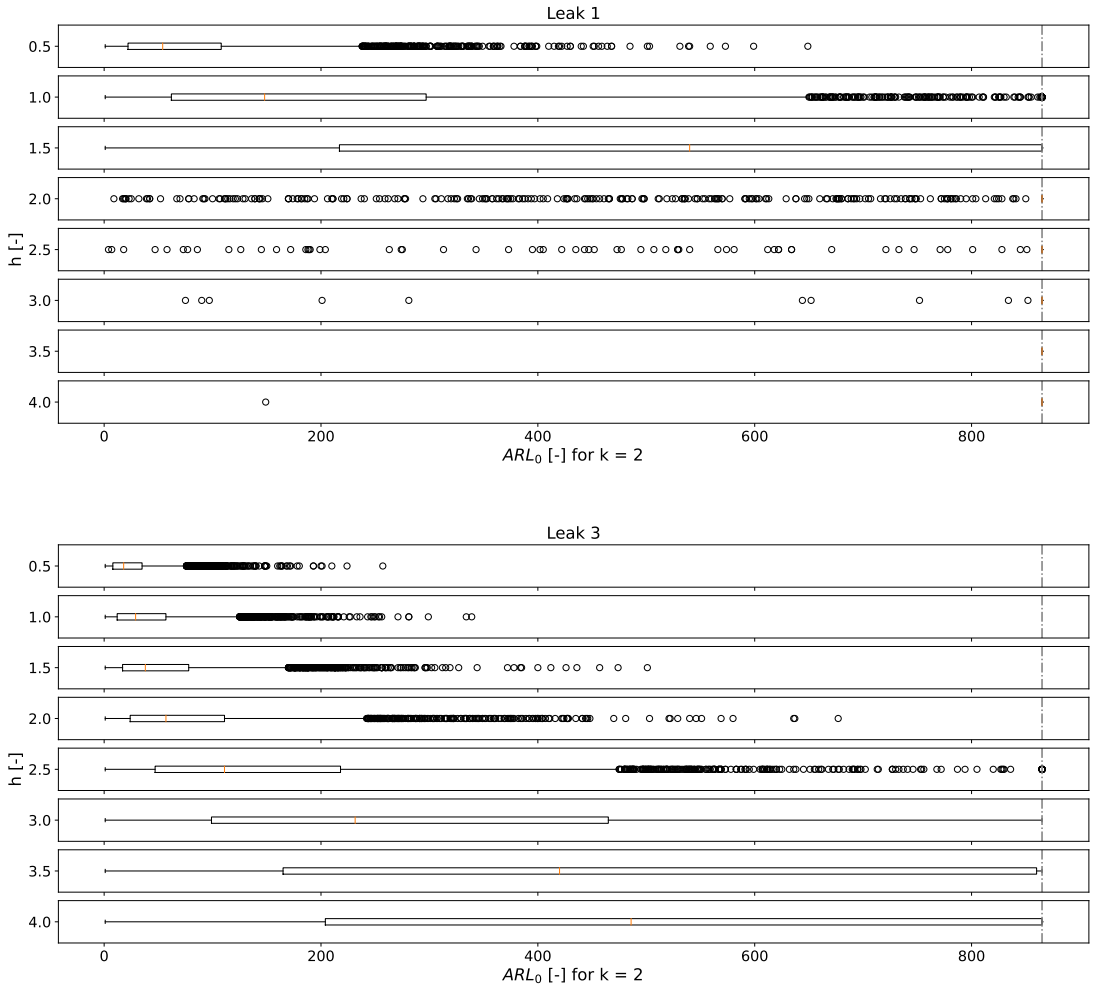

Figure 10: Boxplot of recorded run lengths, i.e. the set $\{RL_i\}_{i=1}^{5000}$ for MC approaches: for different h-values and $k = 2$, this Figure shows the recorded run lengths for leak 1 (upper Figure) and leak 3 (lower Figure). As h increases, the number of usable recoreded run-length decreases, which in turn increases the estimation error of MCD and MCT estimates.

A.3 Statistical tests

In order to test for the presence of autocorrelation, and non-normal in-control distributions, the Durbin-Watson (Durbin and Watson 1992) and Kolmogorov-Smirnov (Massey 1951) tests are employed. For each time series, the first three days (864 data points) are used as in-control data with exception of leak 16, where the true leak starts on the second day, and thus only the first day (288 data points) is used.

Table 6 presents the results for the test-statistic t_{DW} of the Durbin-Watson test, and the p-value of the Kolmogorov-Smirnov test. For former test, the time series is assumed to be uncorrelated if $1.5 \leq t \leq 2$. Therefore, leaks 3,5,9, and 11 are assumed to exhibit autocorrelation. Similarily, for p-values $p < 0.5$, the in-control distribution is assumed to be non-normal. Hence, leaks 1,3,6,9,13, and 18 are assumed to be non-normal. Indeed, the histograms of the in-control data of leaks 3,9, and 18 are clearly non-normal, while leaks 1,6, and 13 seem to be slighly non-normal (see Figure 11).

Table 6: Statistical tests results

Leak ID	type	Durbin-Watson test statistic t_{DW}	Kolmogorov-Smirnov p-value p
1	abrupt	2.07	**0.01**
2	abrupt	1.86	0.9
3	abrupt	**0.18**	**0.0**
4	incipient	1.98	0.36
5	abrupt	**1.31**	0.27
6	abrupt	2.03	**0.04**
7	abrupt	1.56	0.12
8	incipient	2.09	0.2
9	incipient	**0.16**	**0.0**
10	abrupt	1.98	0.38
11	abrupt	**0.91**	0.34
12	incipient	1.73	0.42
13	incipient	1.61	**0.04**
14	incipient	1.98	0.2
15	incipient	1.97	0.15
16	incipient	1.92	0.82
17	incipient	1.87	0.47
18	incipient	1.97	**0.0**
19	incipient	1.97	0.09

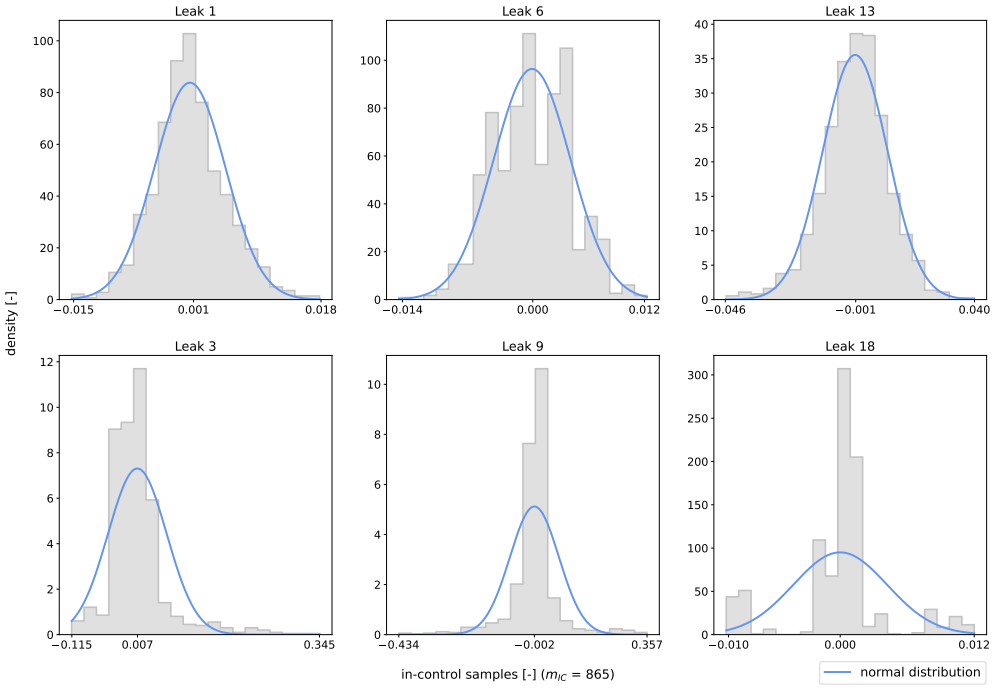

Figure 11: Histograms of in-control distributions: this Figure shows the histograms of the in-control distributions of leaks 1, 6, and 13 (upper Figure), and leaks 3, 9, and 18 (lower Figure). Compared to the normal distributions plotted, these distributions are non-normal based on the Kolmogorov-Smirnov test.

A.4 Nomenclature and Abbreviations

Table 7: Nomenclature

Symbol	Description
b_{max}	number of steps after which samples are assumed to be uncorrelated
c	Box-Cox constant
c_1, c_2	constants for alterantive threshold
C^+	positive CUSUM
$C_1^+, C_1^-, C_2^+, C_2^-$	positive CUSUM (shift), negative CUSUM (shift), positive CUSUM (scale), negative CUSUM (scale)
d	number of regions in which initial samples are divided
F_1	F_1-score
h	threshold hyperparameter
H	threshold
k	slack value hyperparameter
K	slack value
m_{IC}	in-control data set size
p	p-value
precision	precision
q	design hyperparameter
Q_1, Q_2, Q_3	-first, second, and third quartile
r	false positive risk
recall	recall
R	standarized sequential rank
t_{DW}	Durbin-Watson test statistic
T	spring length
TM	trimean
w	weighting factor
W	EWMA-estimator of mean
$x_1, x_2, \ldots$	monitored samples with time index
$x_1^*, x_2^*, \ldots$	uncorrelized and standardized samples with time index
$x_1^{BC}, x_2^{BC}, \ldots$	Box-Cox transformed samples with time index
$Y_1, Y_2, \ldots$	GWMA statistic with time index
$z_1, z_2, \ldots$	z-score standardized and Box-Cox transformed samples with time index
Z	dependent Bernoulli random variable
α	weight hyperparameter
$\tilde{\alpha}$	adjustment hyperparameter
δ	true constant mean shift
$\hat{\delta}$	approximation of mean shift
$\hat{\delta}_0$	initial approximation of mean shift (at time index 0)
μ_0, σ_0	mean and standard deviation of in-control set of $x_1, x_2, \ldots, x_{m_{IC}}$
μ_{tr}, σ_{tr}	mean and standard deviation of Box-Cox transformed in-control set $x_1^{BC}, x_2^{BC}, \ldots, x_{m_{IC}}^{BC}$
μ_Y, σ_Y	mean and standard deviation of in-control set of GWMA statistic $Y_1, Y_2, \ldots, Y_{m_{IC}}$
λ	Box-Cox variable

Table 8: Abbreviations

Abbreviation	Meaning
ac	method acyronym: nonparametric & adaptive CUSUM for arbitrary change
adn	method acronym: nonparametric & adaptive CUSUM
aTTD	average time to detection
ARL_0	in-control average run length
BattLeDIM	Battle of Leakage Detection and Isolation Methods
corr	method acyronym: nonparametric & adaptive CUSUM for autocorrelated data
CUSUM	cumulative sum
DMA	district meter area
EWMA	exponentially weighted moving average
FN	false negative
FP	false positive
gw	method acronym: GWMA-CUSUM
GWMA	generally weighted moving average
IC	in-control
LILA	leakage identification and localization algorithm
MCD	drop-out Monte Carlo
MCT	truncated Monte Carlo
MRE	model reconstruction error
RL	run length
SCADA	Supervisory Control and Data Aquisition
SPC	stochastic process control
t	method acronym: transformed CUSUM
tr	method acronym: transformed & robust CUSUM
TP	true positive
TTD	time to detection
w	method acronym: weighted CUSUM
WDN	water distribution network

A.5 Summary of hyperparameter selection

Table 9: Summary of hyperparameter selection.

Method acronym	Method	Hyperparameters
	standard CUSUM	$h = 3, k = 2$
t	transformed CUSUM (Figueiredo and Gomes 2003)	$h = 3, k = 3$
tr	transformed & robust CUSUM (Figueiredo and Gomes 2003; Nazir et al. 2013)	$h = 3, k = 3$
w	weighted CUSUM (Shu et al. 2008)	$h = 3, k = 2, \alpha = 0.75$
gw	GWMA-CUSUM (Lu 2017)	$c_1 = 1443.8, c_2 = 1.65, \bar{\alpha} = 0.9, q = 0.2, k = 1, h = 41$
adn	nonparametric & adaptive CUSUM (Liu et al. 2014)	$\hat{\delta}_0 = 0.7, ARL_0 = 30,$ empirical threshold $H = 200$
ac	nonparametric & adaptive CUSUM for arbitrary change (Li 2021)	$m_{IC} = 200, d = 10,$ empirical threshold $H = 5000$
corr	nonparametric & adaptive CUSUM for autocorrelated data (Liu et al. 2014; Li and Qiu 2020)	$b_{\max} = 10, ARL_0 = 30, m_{IC} = 200, \hat{\delta}_0 = 0.7,$ empirical threshold $H = 80$

Data Availability Statement

Some or all data, models, or code generated or used during the study are available in online repositories in accordance with funder data retention policies (Daniel et al. 2021; Vrachimis et al. 2022). Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgement

This work is supported by the iOLE project, which receives funding from the Federal Ministry of Education and Research (BMBF) within the funding measure "Digital GreenTech—Environmental Engineering meets Digitalisation" as part of the "Research for Sustainability (FONA) Strategy" (funding code: 02WDG1689A).

References

Abbas, N., Riaz, M., and Does, R. (2013). "Mixed exponentially weighted moving average–cumulative sum charts for process monitoring". In: *Quality and Reliability Engineering International* 29.3, pp. 345–356.

Ahn, J. and Jung, D. (2019). "Hybrid statistical process control method for water distribution pipe burst detection". In: *Journal of Water Resources Planning and Management* 145.9, p. 06019008.

Ali, S. (2020). "A predictive Bayesian approach to EWMA and CUSUM charts for time-between-events monitoring". In: *Journal of Statistical Computation and Simulation* 90.16, pp. 3025–3050.

Ali, S., Pievatolo, A., and Göb, R. (2016). “An overview of control charts for high-quality processes”. In: *Quality and reliability engineering international* 32.7, pp. 2171–2189.

Anjana, G. et al. (2015). “A particle filter based leak detection technique for water distribution systems”. In: *Procedia Engineering* 119, pp. 28–34.

Bakker, M., Jung, D., et al. (2014). “Detecting pipe bursts using Heuristic and CUSUM methods”. In: *Procedia Engineering* 70, pp. 85–92.

Bakker, M., Vreeburg, J., et al. (2014). “Heuristic burst detection method using flow and pressure measurements”. In: *Journal of Hydroinformatics* 16.5, pp. 1194–1209.

Blesa, J. and Pérez, R. (2018). “Modelling uncertainty for leak localization in water networks”. In: *IFAC-PapersOnLine* 51.24, pp. 730–735.

Boretti, A. and Rosa, L. (2019). “Reassessing the projections of the world water development report”. In: *NPJ Clean Water* 2.1, p. 15.

Bourazas, K., Sobas, F., and Tsiamyrtzis, P. (2023). “Predictive ratio CUSUM (PRC): A Bayesian approach in online change point detection of short runs”. In: *Journal of Quality Technology* 55.4, pp. 391–403.

Bozkurt, C., Firat, M., and Ateş, A. (2022). “Development of a new comprehensive framework for the evaluation of leak management components and practices”. In: *AQUA—Water Infrastructure, Ecosystems and Society* 71.5, pp. 642–663.

Brook, D. and Evans, D. (1972). “An approach to the probability distribution of CUSUM run length”. In: *Biometrika* 59.3, pp. 539–549.

Buchberger, S. and Nadimpalli, G. (2004). “Leak estimation in water distribution systems by statistical analysis of flow readings”. In: *Journal of water resources planning and management* 130.4, pp. 321–329.

Chandola, V., Banerjee, A., and Kumar, V. (2009). “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3, pp. 1–58.

Crowder, S. (1987). “A simple method for studying run-length distributions of exponentially weighted moving average charts”. In: *Technometrics* 29.4, pp. 401–407.

Daniel, I., Letzgus, S., and Cominola, A. (2021). *A high-resolution pressure-driven leakage identification and localization algorithm [Code]*. `https://github.com/SWN-group-at-TU-Berlin/LILA`, accessed on 10 April 2025.

Daniel, I., Pesantez, J., et al. (2022). “A sequential pressure-based algorithm for data-driven leakage identification and model-based localization in water distribution networks”. In: *Journal of Water Resources Planning and Management* 148.6, p. 04022025.

Durbin, J. and Watson, G. (1992). “Testing for serial correlation in least squares regression. II”. In: *Breakthroughs in Statistics: Methodology and Distribution*. Springer, pp. 260–266.

Eliades, D. and Polycarpou, M. (2012). “Leakage fault detection in district metered areas of water distribution systems”. In: *Journal of Hydroinformatics* 14.4, pp. 992–1005.

Farah, E. and Shahrour, I. (2017). “Leakage detection using smart water system: Combination of water balance and automated minimum night flow”. In: *Water Resources Management* 31, pp. 4821–4833.

Figueiredo, F. and Gomes, M. (2003). “Box-Cox transformations versus data modelling—robust charts in Statistical Process Control”. In: *Notas e Comunicaçoes CEAUL* 13, p. 2003.

— (2004). “The total median in statistical quality control”. In: *Applied stochastic models in business and industry* 20.4, pp. 339–353.

Fox, S. et al. (2016). “Experimental quantification of contaminant ingress into a buried leaking pipe during transient events”. In: *Journal of Hydraulic Engineering* 142.1, p. 04015036.

George, A. and Box, E. (2000). “Influence of the sampling interval, decision limit and autocorrelation on the average run length in CUSUM charts”. In: *Journal of Applied Statistics* 27.2, pp. 177–183.

Graham, M., Chakraborti, S., and Mukherjee, A. (2014). “Design and implementation of CUSUM exceedance control charts for unknown location”. In: *International Journal of Production Research* 52.18, pp. 5546–5564.

Granjon, P. (2013). “The CUSUM algorithm-a small review”. In.

Hamasha, M., Ali, H., and Ahmed, A. (2022). “Ultra-fine transformation of data for normality”. In: *Heliyon* 8.5.

Heard, N. and Turcotte, M. (2017). “Adaptive sequential Monte Carlo for multiple changepoint analysis”. In: *Journal of Computational and Graphical Statistics* 26.2, pp. 414–423.

Higham, N. (2009). “Cholesky factorization”. In: *Wiley interdisciplinary reviews: computational statistics* 1.2, pp. 251–254.

Javed, A. et al. (2024). “Designing Bayesian paradigm-based CUSUM scheme for monitoring shape parameter of the Inverse Gaussian distribution”. In: *Computers & Industrial Engineering* 192, p. 110235.

Jones, M. and Steiner, S. (2012). “Assessing the effect of estimation error on risk-adjusted CUSUM chart performance”. In: *International journal for quality in health care* 24.2, pp. 176–181.

Jones-Farmer, L., Jordan, V., and Champ, C. (2009). “Distribution-free phase I control charts for subgroup location”. In: *Journal of Quality Technology* 41.3, pp. 304–316.

Jung, D., Kang, D., et al. (2015). “Improving the rapidity of responses to pipe burst in water distribution systems: A comparison of statistical process control methods”. In: *Journal of Hydroinformatics* 17.2, pp. 307–328.

Jung, D. and Lansey, K. (2015). “Water distribution system burst detection using a nonlinear Kalman filter”. In: *Journal of Water Resources Planning and Management* 141.5, p. 04014070.

Kim, Y. et al. (2016). “Robust leak detection and its localization using interval estimation for water distribution network”. In: *Computers & Chemical Engineering* 92, pp. 1–17.

Knoth, S. (2021). “Steady-state average run length(s): Methodology, formulas, and numerics”. In: *Sequential Analysis* 40.3, pp. 405–426.

LeChevallier, M. et al. (2003). “The potential for health risks from intrusion of contaminants into the distribution system from pressure transients”. In: *Journal of water and health* 1.1, pp. 3–14.

Lever, J. (2016). “Classification evaluation: It is important to understand both what a classification metric expresses and what it hides”. In: *Nature methods* 13.8, pp. 603–605.

Li, J. (2021). “Nonparametric adaptive CUSUM chart for detecting arbitrary distributional changes”. In: *Journal of Quality Technology* 53.2, pp. 154–172.

Li, W. and Qiu, P. (2020). “A general charting scheme for monitoring serially correlated data with short-memory dependence and nonparametric distributions”. In: *IISE Transactions* 52.1, pp. 61–74.

Liemberger, R. and Marin, P. (2006). “The challenge of reducing non-revenue water in developing countries–how the private sector can help: A look at performance-based service contracting”. In.

Liemberger, R. and Wyatt, A. (2019). “Quantifying the global non-revenue water problem”. In: *Water Supply* 19.3, pp. 831–837.

Lim, J. and Lee, S. (2024). “Efficient ARL estimation for general control charts using censored run lengths”. In: *Quality Engineering*, pp. 1–10.

Liu, L., Tsung, F., and Zhang, J. (2014). “Adaptive nonparametric CUSUM scheme for detecting unknown shifts in location”. In: *International Journal of Production Research* 52.6, pp. 1592–1606.

Loureiro, D. et al. (2016). “Water distribution systems flow monitoring and anomalous event detection: A practical approach”. In: *Urban Water Journal* 13.3, pp. 242–252.

Lu, S. (2017). “Novel design of composite generally weighted moving average and cumulative sum charts”. In: *Quality and Reliability Engineering International* 33.8, pp. 2397–2408.

Mabude, K. et al. (2021). “Generally weighted moving average monitoring schemes: Overview and perspectives”. In: *Quality and Reliability Engineering International* 37.2, pp. 409–432.

Massey, F. (1951). “The Kolmogorov-Smirnov test for goodness of fit”. In: *Journal of the American statistical Association* 46.253, pp. 68–78.

Misiunas, D. et al. (2006). “Failure monitoring in water distribution networks”. In: *Water science and technology* 53.4-5, pp. 503–511.

Montgomery, D. (2019). *Introduction to statistical quality control*. John wiley & sons.

Mounce, S. and Machell, J. (2006). “Burst detection using hydraulic data from water distribution systems with artificial neural networks”. In: *Urban Water Journal* 3.1, pp. 21–31.

Moustakides, G. (1986). “Optimal stopping times for detecting changes in distributions”. In: *the Annals of Statistics* 14.4, pp. 1379–1387.

Nazir, H. et al. (2013). “Robust CUSUM control charting”. In: *Quality Engineering* 25.3, pp. 211–224.

Nimri, W. et al. (2023). “Data-driven approaches and model-based methods for detecting and locating leaks in water distribution systems: a literature review”. In: *Neural Computing and Applications* 35.16, pp. 11611–11623.

Page, E. (1954). “Continuous inspection schemes”. In: *Biometrika* 41.1/2, pp. 100–115.

Palau, C., Arregui, F., and Carlos, M. (2012). “Burst detection in water networks using principal component analysis”. In: *Journal of Water Resources Planning and Management* 138.1, pp. 47–54.

Peterson, R. (2021). “Finding Optimal Normalizing Transformations via best Normalize.” In: *R Journal* 13.1.

Polunchenko, A. (2016). “A note on efficient performance evaluation of the Cumulative Sum chart and the Sequential Probability Ratio Test”. In: *Applied Stochastic Models in Business and Industry* 32.5, pp. 565–573.

Puust, R. et al. (2010). “A review of methods for leakage management in pipe networks”. In: *Urban Water Journal* 7.1, pp. 25–45.

Romano, M., Kapelan, Z., and Savic, D. (2014). “Automated detection of pipe bursts and other events in water distribution systems”. In: *Journal of Water Resources Planning and Management* 4, pp. 457–467.

Romano, M., Woodward, K., and Kapelan, Z. (2017). “Statistical process control based system for approximate location of pipe bursts and leaks in water distribution systems”. In: *Procedia Engineering* 186, pp. 236–243.

Romero-Ben, L. et al. (2023). “Leak detection and localization in water distribution networks: Review and perspective”. In: *Annual Reviews in Control*.

Saleh, N. et al. (2023). “A review and critique of auxiliary information-based process monitoring methods”. In: *Quality Technology & Quantitative Management* 20.1, pp. 1–20.

Shu, L., Jiang, W., and Tsui, K. (2008). “A weighted CUSUM chart for detecting patterned mean shifts”. In: *Journal of Quality Technology* 40.2, pp. 194–213.

Soldevila, A., Blesa, J., et al. (2016). “Leak localization in water distribution networks using a mixed model-based/data-driven approach”. In: *Control Engineering Practice* 55, pp. 162–173.

Soldevila, A., Boracchi, G., et al. (2022). “Leak detection and localization in water distribution networks by combining expert knowledge and data-driven models”. In: *Neural Computing and Applications* 34.6, pp. 4759–4779.

Steffelbauer, D. et al. (2022). “Pressure-leak duality for leak detection and localization in water distribution systems”. In: *Journal of Water Resources Planning and Management* 148.3, p. 04021106.

Steiner, S. et al. (2000). “Monitoring surgical performance using risk-adjusted cumulative sum charts”. In: *Biostatistics* 1.4, pp. 441–452.

Tukey, J. (1977). “Exploratory data analysis”. In: *Reading/Addison-Wesley*.

Vrachimis, S. et al. (2022). “Battle of the leakage detection and isolation methods”. In: *Journal of Water Resources Planning and Management* 148.12, p. 04022068.

Wald, A. (1945). “Sequential Tests of Statistical Hypotheses”. In: *The Annals of Mathematical Statistics* 16.2, pp. 117–186.

Wan, X. et al. (2022). “Literature review of data analytics for leak detection in water distribution networks: A focus on pressure and flow smart sensors”. In: *Journal of Water Resources Planning and Management* 148.10, p. 03122002.

Wang, X. et al. (2020). “Burst detection in district metering areas using deep learning method”. In: *Journal of Water Resources Planning and Management* 146.6, p. 04020031.

Woodall, W. and Ncube, M. (1985). “Multivariate CUSUM quality-control procedures”. In: *Technometrics* 27.3, pp. 285–292.

Wu, X., Peng, S., et al. (2023). “Leakage Detection in Water Distribution Networks Based on Multi-Feature Extraction from High-Frequency Pressure Data”. In: *Water* 15.6, p. 1187.

Wu, Y., Liu, S., and Kapelan, Z. (2024). “Addressing data limitations in leakage detection of water distribution systems: Data creation, data requirement reduction, and knowledge transfer”. In: *Water Research*, p. 122471.

Yu, X. and Cheng, Y. (2022). “A comprehensive review and comparison of CUSUM and change-point-analysis methods to detect test speededness”. In: *Multivariate Behavioral Research* 57.1, pp. 112–133.

Zhang, X. and Woodall, W. (2015). “Dynamic probability control limits for risk-adjusted Bernoulli CUSUM charts”. In: *Statistics in medicine* 34.25, pp. 3336–3348.