AQUA
Water Infrastructure, Ecosystems and Society

IWA PUBLISHING

 AQUA — Water Infrastructure, Ecosystems and Society Vol 71 No 1, 31 doi: 10.2166/aqua.2021.091

Data-driven and model-based framework for smart water grid anomaly detection and localization

Z. Y. Wu[a,*], A. Chew[b], X. Meng[b], J. Cai[b], J. Pok[b], R. Kalfarisi[b], K. C. Lai[c], S. F. Hew[c] and J. J. Wong[c]

[a] Bentley Systems, Incorporated, Watertown, CT, USA
[b] Bentley Systems Singapore Pte Ltd., Singapore
[c] Water Supply (Network) Department, Public Utilities Board (PUB), Singapore
*Corresponding author. E-mail: zheng.wu@bentley.com

AC, 0000-0002-4555-6462; XM, 0000-0001-9489-3986; JP, 0000-0002-7296-3872; RK, 0000-0003-2090-9898

ABSTRACT

With increasing adoption of advanced meter infrastructure, smart sensors together with SCADA (Supervisory Control and Data Acquisition) systems, it is imperative to develop novel data analytics and couple the results with hydraulic modeling to improve the quality and efficiency of water services. One important task is to timely detect and localize anomaly events, which may include, but not be limited to, pipe bursts and unauthorized water usages. In this paper, a comprehensive solution framework has been developed for anomaly detection and localization by formulating and integrating data-driven analytics with hydraulic model calibration. Data analysis for anomaly detection proceeds in multiple steps including the following: (1) data pre-processing to eliminate and correct erroneous data records, (2) outlier detection by statistical process control methods and deep machine learning, and (3) system anomaly classification by correlation analysis of multiple sensor events. Classified system anomaly events are subsequently localized via hydraulic model calibration. The integrated solution framework is developed as a user-friendly and effective software tool, tested, and validated on the selected target areas in Singapore.

Key words: anomaly detection, anomaly localization, data analytics, hydraulic model, smart water grid

HIGHLIGHTS

- Comprehensive solution framework for anomaly detection and localization.
- Integration of data-driven analytics with hydraulic model calibration.
- Data analysis for anomaly detection.
- Classification of system anomaly events.
- Testing and validating the solution on the selected target areas in Singapore.

INTRODUCTION

Over the last decade, with the continual advancement of emerging technologies, especially cost-effective sensors and ubiquitous internet connectivity, more and more smart meters, sensors, and data loggers are deployed for monitoring water distribution systems. A large amount of data, often referred to as big data, are collected for enabling smart water network operation and management (PUB 2016). One important system operation task is to detect the abnormal conditions or so-called anomaly events that are captured by and embedded in flow and pressure time-series data.

Traditionally, only flows are recorded at the inlet of a system or district metered areas (DMAs). The average flow rate between 2 AM and 4 AM, so-called minimum night flow (MNF) hours, is used for evaluating if any new pipe burst occurs (Alegre *et al.* 2000). This is because there is little water usage and pressure is usually high at MNF hours; hence, pipe bursts are more likely to be caused by high pressures than low pressures. If the most recent MNF is significantly greater than that of previous days, it is very likely that new pipe bursts have occurred in a DMA, and pre-emptive actions are required to localize the event. The MNF method is simple and easy to implement; however, it prevents engineers from leveraging on the entire dataset to detect the anomaly events occurring at non-MNF hours.

In general, anomaly detection is the identification of rare items, events, or observations, which invoke suspicions by differing significantly from the benchmark data (flow and pressure). Over last decades, much research works have been conducted for developing anomaly detection methods (Mounce & Boxall 2011). In general, the following three types of methods – including prediction–classification (PC) approaches, clustering algorithms (CA), and statistical process control (SPC) methods – have been developed and applied to anomaly detections in water distribution systems.

The PC approach is to construct a prediction model using historical data. Once established, the prediction model is used to forecast the flow and/or pressures for the normal conditions; if the prediction is out of predefined bounds, the anomaly is deemed to be detected or occur. Mounce *et al.* (2002) pioneered the research in applying artificial neural network (ANN) for burst detection. Since then, the research in applying ANN for the anomaly detection has been improved and applied to some real-system case studies (Mounce *et al.* 2011; Romano *et al.* 2014a, 2014b). In addition to conventional ANN models, other prediction methods have also been applied to pipe burst detection. A polynomial function with an expectation–maximization algorithm (Ye & Fenner 2014) has been investigated for fitting historical flow data for automatic burst detection in the UK water systems. A Bayesian demand forecasting approach (Huttona & Kapelan 2015) was also applied to pipe burst detection. The effectiveness of the PC approach depends on the accurate prediction, which is challenging to achieve for real systems. To address the challenge, an evolutionary deep learning framework with data assimilation feature has been developed for effective prediction and outlier detection (Li *et al.* 2019). In general, the PC methods work well but require for relatively much more historical data than other approaches to establish or train an effective prediction model.

The CA approach is proposed to avoid the need for the prediction model. It is to construct a classifier using the data features carefully selected and processed or automatically learned from a historical dataset without anomaly events. After established, the classifier is applied to new data for anomaly detection. Mounce *et al.* (2011) applied support vector machine (SVM) to detect and classify flow events. The approach was further improved by integrating the pattern matching with an ANN (Mounce *et al.* 2014) and an artificial immune network as the classifier (Tao *et al.* 2014) for pipe burst detection. Most recently, Wu *et al.* (2018) applied the CA (Rodriguez & Laio 2014) for burst detection using flow time-series collected at the inlet and the outlet of a DMA. This approach seems to work well for large pipe bursts but may produce many false alarms for quasi-bursts, so that a sophisticated alarm rule must be developed as authors indicated.

The SPC method is the approach widely applied for quality assurance (QA) and quality control (QC) in manufacture industry. It was defined as the use of statistical techniques to control a process by the control chart that helps to detect an unusual or anomaly event such as a very high or low observation compared with 'normal' process performance. Jung & Lansey (2015) applied SPC together with nonlinear Kalman filter (NKF) for pipe burst detection. The NKF was employed for estimating system states (pressures and flows), while SPC methods were applied to detect outliers or pipe bursts with many assumptions that can be difficult to meet. Loureiro *et al.* (2016) and Meeus & Marshallsay (2017) implemented a SPC approach for pipe burst detection but only flow data were used, and no generic SPC-based method was developed for anomaly detection using both flow and pressure monitoring data. In addition, a SPC method is limited to the stationary time-series data. Unfortunately, flows and pressures vary periodically, e.g., daily, weekly, and monthly; thus, the nonstationary time-series must be decomposed to meet the stationarity as required by SPC methods.

In addition to detecting when an anomaly event or leakage occurs in a system, it is essential to localize whereabouts the anomaly. Early study in localizing a leakage is by transient analysis. Brunone & Ferrante (2001) has reported on the laboratory results to confirm the reliability and validity of the inverse transient analysis for a long pipeline. Since then, many transient-based approaches have been developed for leakage detection. Nixon *et al.* (2006) have carefully studied the range of validity of the inverse transient analysis method and concluded that its applicability is limited to the instantaneous small amplitude disturbances within simple reservoir–pipe–valve-type configurations or reservoir–pipe–reservoir systems. However, it is a well-known fact that leakage is pressure-dependent demand, which is usually represented as an emitter in hydraulic models. Wu *et al.* (2010) developed the pressure-dependent leak detection (PDLD) method, which emulates leakage hotspots as emitter flows at model junction nodes in a network. A given maximum number of leakage hotspots or nodes are optimized as part of model calibration to minimize the discrepancy between the modeled and observed values (flows and pressures). The authors verified the method's capability for a district water system having 1,122 pipes, 841 junction nodes, and a single variable head reservoir. More than a dozen of historical leaks has been successfully localized using the PDLD method with the monitoring data.

Some other studies of model-based leakage localization have also been reported in the literature. Applying the same model formulation as Wu *et al.* (2010), Sophocleous *et al.* (2019) applied the method to localize the simulated leak in a real network

with the search space reduction by excluding those nodes that are not connected to pipes. Ponce *et al.* (2014) proposed a model-based approach for leak localization using pressure sensitivity matrix. Five different ways were formulated for isolating the leaks and tested on two academic small networks (Hanoi and Nova) with the simulated leaks. Perez *et al.* (2014) presented a slightly different approach from Ponce *et al.* (2014) for leak localization by generating a sensitivity-to-leak matrix and demonstrated the application of their method using the same synthetic dataset. To consider the uncertainty of model parameters, such as demand uncertainty, sensor noise, and leak sizes, Soldevila *et al.* (2017) reported a method for leak localization based on the Bayesian classifier, which was also tested on the same academic dataset as Ponce *et al.* (2014). More recently, Vrachimis *et al.* (2021) presented leakage detection and localization via model-invalidation, by which the uncertainties of the model parameters including pipe roughness and nodal demand are emulated by the prescribed upper and lower bounds. A 'health' or none-leak condition can be established by the model simulations using the parameters within the bounds. The approach was tested on the simple Hanoi network with the simulated dataset. Up to date, the published research on this topic has been mostly limited to using the simulated or synthetic data for the detection and localization of anomaly event or leakage hotspots. To address the challenges, this paper proposed an integrated solution approach for detecting and localizing anomaly events in real-world large-scale water distribution systems.

METHODS

An integrated solution framework has been developed for anomaly detection and localization. Figure 1 illustrates the proposed solution procedure. It proceeds in two phases of anomaly detection and anomaly localization. The first phase starts with data pre-processing to correct erroneous data records of historical flow and pressure time-series records. The corrected time-series is then decomposed to ensure data stationarity. The outliers can subsequently be detected using the data-driven analytics, e.g., SPC methods. The detected outliers are classified into anomaly events, which are localized by employing hydraulic model calibration in the second phase. The detailed methodology for event detection and localization is elaborated as follows.

Event detection

Anomaly event detection is conducted by analyzing the monitoring data (flows and pressures) by so-called data-driven modeling. Two types of data-driven models, including (1) deep machine learning with data assimilation and (2) statistical control methods, have been developed and tested for the Public Utilities Board (PUB) water systems. Each of the data-driven models is employed for detecting the possible outliers of time-series profiles. In this paper, we focus on statistical control methods such as *X-bar*, exponential weighted moving average (*EWMA*), and cumulative sum (*CUSUM*), which have been integrated

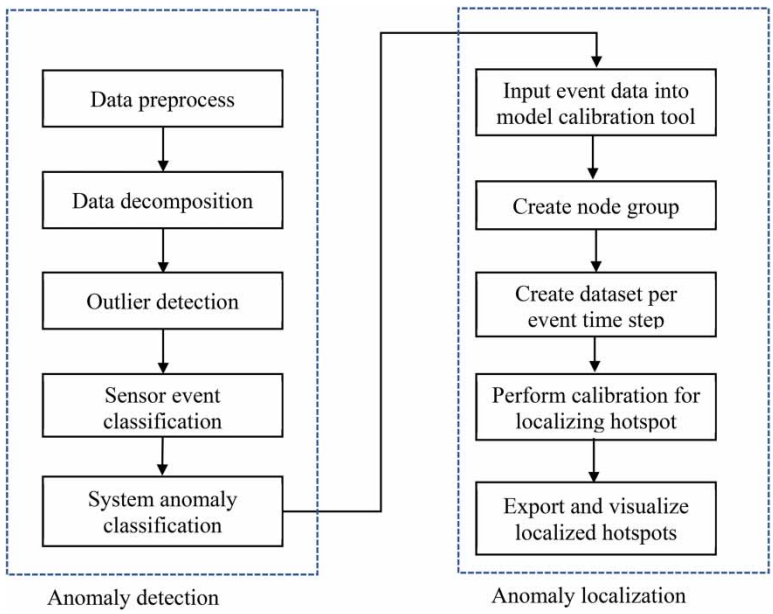

Figure 1 | Integrated solution framework for anomaly detection and localization.

into the Water Event Detection Tool (Wu & He 2021). It facilitates the analysis of time-series flow and pressure data independently via data-driven statistical algorithms by adopting the following systematic workflow.

Data pre-processing

By its nature, raw data collected by sensors exhibit many issues, including but not limited to missing data records and various sensor failures. Therefore, it is important and necessary to pre-process the raw data before any anomaly detection method is applied. Four types of data errors including the following:

- *Missing time steps*: The missing data records will be filled with the values of same time steps from the last period if the number of missing data records is less than the user-specified maximum number of missing time steps; otherwise, the data records of the whole operation cycle, e.g., 24 h, will be simply removed from the dataset.
- *Duplicated time steps*: If there are >1 data records with the same timestamp, then they are duplicated for the time step. If their values are the same, one data record is kept, and the rest is simply deleted. If their values are different, an average value of the data records is used to replace them.
- *Irregular time steps*: Time-series data are usually recorded in a fixed time interval. For instance, if the dataset is in time step of 5 min between two consecutive data records, the data recorded out of this time step are considered as the irregular time-step data. The data record at the expected time is filled by the interpolation of the data records at adjacent times.
- *Sensor failure time steps*: A failed sensor may cause extremely big or small values of data records, which can be eliminated by the prescribed upper and lower thresholds. If the data records remain constant consecutively for more than the specified number of time steps, it is also considered as sensor failure. The identified sensor failure data records will be removed if the number of the error data records is greater than the specified maximum number of data records; otherwise, the error data records will be refilled with the data records of the same time steps over the most recent period.

Decomposition

To effectively apply data analytical methods such as SPC methods for outlier detection, it is essential to ensure that time-series data are stationary, that is, the mean and variance of the dataset do not change over time. Since a typical flow or pressure time-series data contain daily, weekly, and monthly seasonality, it is identified as nonstationary time-series profile. Therefore, flow and pressure time-series are required to be first decomposed to eliminate the influence of seasonality. An improved version of the Seasonal-Trend decomposition procedure based on Loess (STL) (Cleveland *et al.* 1990) has been leveraged to decompose raw flow and pressure time-series.

Outlier detection

After seasonality component is removed from time-series data (flows and pressures), outlier detection can be conducted by applying multiple statistical methods (*X-bar*, *EWMA*, and *CUSUM*) based on the user's choice.

- *X-bar*: *X-bar* chart is a type of SPC methods used to monitor variable attribute. It assumes that the attribute to be monitored is adequately modeled by a normally distributed random variable. The upper control limit (*UCL*) and lower control limit (*LCL*) are defined as below:

$$UCL = \mu + 3\sigma \tag{1}$$

$$LCL = \mu - 3\sigma \tag{2}$$

where μ is the sample mean and σ is the standard deviation.

- *CUSUM*: *CUSUM* control chart is a sequential analysis technique typically used for monitoring a small shift in sample data. It involves the calculation of a *CUSUM*. When the value of sum exceeds a certain threshold value, a change in the value has been identified. *CUSUM* will calculate high-side cumulative sum (*SH*) and low-side cumulative sum (*SL*) of a dataset by the following definitions:

$$SH(0) = SL(0) = 0 \tag{3}$$

$$SH(i) = \mathrm{Max}[0,\ SH(i-1) + X_i - \mu - 0.5\sigma] \tag{4}$$

$$SL(i) = \mathrm{Min}[0,\ SL(i-1) + X_i - \mu + 0.5\sigma] \tag{5}$$

- *EWMA*: *EWMA* charts are generally used for detecting small shifts in the process mean. It weights samples in geometrically decreasing order, so that the most recent samples are weighted most highly while the most distant samples contribute very little. It monitors the *EWMA* value instead of the original sample value.

Anomaly event detection. An anomaly event can be an unauthorized water usage or a pipe burst, which always causes an increased system inflow and pressure drops in some portion of a water distribution network. Thus, a true system anomaly is detected and classified with two important characteristics of high-flow (HF) sensor anomaly (HF outliers) and low-pressure (LP) sensor anomaly (LP outliers) within the same time-window. In general, after the sensor events are identified as elaborated early, system anomaly events can then be classified via correlative analysis of all sensor events within the same time-window. The more sensor events identified within the same time-window, the greater confidence a system anomaly event can be classified. A system anomaly event identified with significantly high confidence must subsequently be localized to enable field crew to pinpoint the predicted anomaly event.

Event localization

Anomaly localization is conducted by applying the PDLD method (Wu *et al.* 2010) via hydraulic model calibration with the pressure and flow monitoring data collected for the detected anomaly event. The PDLD method is formulated to search for the likely anomaly hotspots that are emulated as emitters at nodes where unauthorized water usages or leaks are modeled as pressure-dependent demands in addition to real consumptions. The emitter locations and coefficients at possible anomaly nodes are the decision variables to be optimized such that the difference, i.e., the cost function, between the simulated and field observed pressures/flows is minimized. Identifying and quantifying leakage emitters are effectively part of the model calibration effort to represent additional pressure-dependent demands (e.g., leaks) in a hydraulic model. The PDLD method is integrated with the optimization-based model calibration tool (Wu *et al.* 2002; Wu 2009; Bentley 2018), which can be executed repetitively for the same anomaly event. Multiple runs provide good indication of the most likely anomaly hotspot areas occurring within a water distribution network.

Although a completely calibrated model is not required for anomaly hotspot localization by the PDLD method to obtain good results, it is imperative to ensure the following aspects: (1) the hydraulic model is constructed with adequately accurate nodal demands that reflect real water consumption; (2) the elevations are accurate at locations (nodes) where pressure data are recorded; and (3) boundary conditions, such as service reservoirs, tanks, and pumps, are also accurately represented in the model. The essential steps for applying the optimization-based anomaly localization are as follows.

- *System evaluation*: Prior to embarking on anomaly localization, it is important to first evaluate the existing condition using a hydraulic model that comprehensively represents the system layout, connectivity, base demands, and demand patterns (actual consumption recording) along with operational controls. The evaluation may include an extended hydraulic simulation to assess the existing system by comparing the observed data with the corresponding simulated data for inflow and pressure parameters.
- *Import event data*: Whenever an anomaly event is detected with high confidence, the monitoring data, including the inflow into the water system and the pressures recorded throughout the system, can be imported into the calibration tool. The data are prepared and classified into different datasets, each representing a snapshot of system observations at a given time step of the detected event. For instance, a 2-h event, with 15-min time step, will result in a total of eight snapshots of datasets that can be prepared for the anomaly localization optimization modeling.
- *Anomaly event localization*: Anomaly location modeling proceeds with setting up and performing the model calibration runs, each of which can be dedicated to search for the maximum number of anomaly nodes among a user-specified group of nodes. It is recommended that engineers use system-specific expert knowledge to make selection of the candidate anomaly nodes for effective modeling, although one node group is adequate for many models (those at the scale of a DMA, for example). Each node in a group is treated as a possible anomaly node during the optimization and the most likely anomaly nodes are to be identified by the novel procedure. Optimization runs can be executed with one or multiple snapshots of the datasets for the detected event. The nodes that are repetitively identified by multiple runs are deemed to be good indicators of anomaly locations.
- *Results visualization*: Localized anomaly hotspots are represented as nodal emitters with positive emitter coefficients. They are the attributes of the nodes that can be color-coded to visualize the hotspots of the localized anomaly events in a water distribution network.

RESULTS AND DISCUSSION

The above-proposed solution approach has been applied to two study areas in Singapore.

Anomaly detection example

One-year data from 1 June 2016 to 30 June 2017 have been provided for Area I (total pipe length of 328 km) originating from one service reservoir flow sensor and 11 pressure loggers in the network. The leaks are recorded in three categories of different leak sizes from large to small, including Type I leaks, Type II leaks, and Type III leaks. As shown in Figure 2, detection rates of 84, 78, and 80% have been achieved with an *X-bar* method for three categories of leaks by applying the proposed anomaly detection solution.

Detection performance

Evaluation of the results derived from using the different data-driven statistical algorithms of our proposed Water Event Detection Tool is performed by computing two key metrics, namely (a) accuracy-hit value (*AHV*), as defined in Equation (6); (b) precision (*P*), as defined in Equation (7); and (c) false positive rate (*FPR*) given by Equation (8) and true positive rate (*TPR*) defined by Equation (9). In Equation (6), it is the ratio between the total number of correctly detected system anomaly events, noted as *Npr* (i.e., true positives, *TP*) within a predefined time-window (*T*), as measured in days (i.e., considering only dates), and the number of actual complaints events reported (noted as *Nre*) by PUB. On the other hand, the precision metric in Equation (7) represents a more rigorous evaluation of the proposed model's predictive capability by considering the number of false system anomaly events (i.e., false positives, *FP*) predicted by the model.

$$AHV = \frac{Npr}{Nre} \tag{6}$$

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$FPR = \frac{FP}{TP + FP} \tag{8}$$

$$TPR = \frac{TP}{TP + FP} \tag{9}$$

As our first approach, we consider the lead- and lagged-time of 3 days and 1 day, respectively, in this study. For example, if the reported/complaint event is on 1 May 2020, the allowable time range for detecting the anomaly events can be any date times between 29 April 2020 and 3 May 2020 as illustrated in Figure 3. Therefore, *TP* events (considering dates and timings)

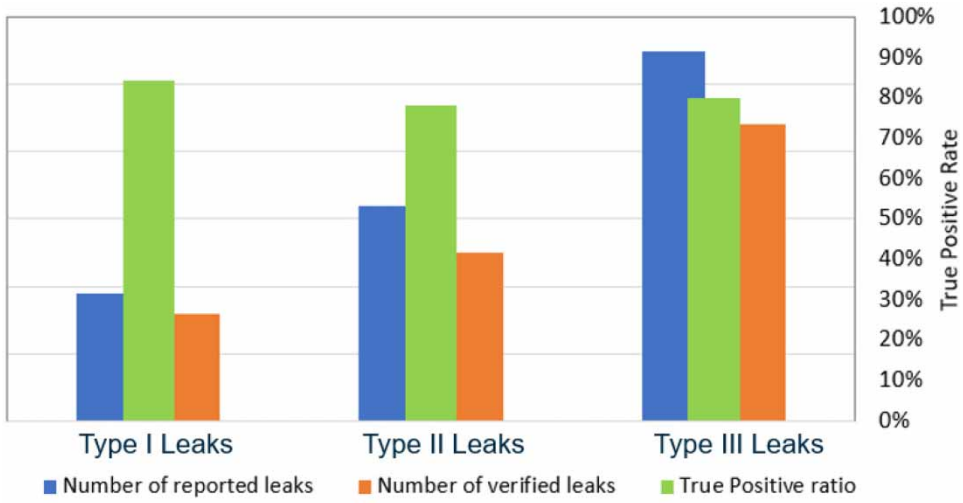

Figure 2 | Summary of detected leaks for Area I.

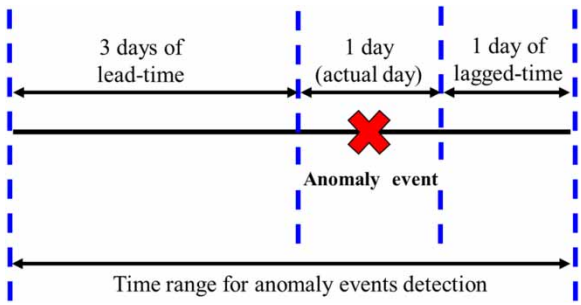

Figure 3 | Illustration of time range for predicted anomaly system events from our proposed Water Event Detection Tool.

are predicted events which lie within the allowable time range as defined, while *FP* events are those which lie outside of the allowable time range. Table 1 summarizes the computed values for *AHV* and *P* from the use of three different statistical methods for high- flow and low-pressure (HF–LP) conditions in the studied water system in Singapore's context. Overall, both *EWMA* and *X-bar* methods can effectively detect the reported anomaly events in the system within the defined time range (see Figure 3) by achieving reasonably high *TP* rate or precision scores of at least 70% and above, while also ensuring >95% of the actual reported events have been correctly detected.

Anomaly localization example

The system contains 3,368 pipes, one service reservoir, and 672 valves. The system flow is monitored in 5-minutes' time step along with eight pressure monitoring stations where data are collected every 15 min. Three-month monitoring data from 1 September 2017 to 30 November 2017 have been provided for the study area. Likewise, the raw time-series data were first processed to manage the above-mentioned types of data errors. Table 2 summarizes the error data records pertaining to the study area. As shown in the same table, sensor failure was the most common type of error data record, which was identified by specifying the upper- and lower-value limits, and the maximum number of consecutive time steps when the recorded value is the same. For instance, the sensor failure data records for the flow time-series were found with the lower and upper limit values of 0 and 50 MGD, respectively, and two consecutive time steps (equivalent to 10 min) of the same flow values.

For error data records which last for a long period, e.g., 4 h, the data records for the whole day are disregarded for further analysis. Otherwise, the error data records are refilled using the values of the same time steps from the most recent period

Table 1 | Summary of computed *AHV* and *P* values for HF–LP conditions

Methods	*X-bar*	*EWMA*	*CUSUM*
TP rate	0.710	0.728	0.778
FP rate	0.290	0.272	0.222
Precision (*P*)	0.710	0.728	0.778
AHV	0.972	0.981	0.208

Table 2 | Summary of data pre-processing for case II

Data types	Inflow	Pressure						
Sensor names	F10	Stn11	Stn12	Stn13	Stn14	Stn15	Stn16	Stn17
Duplicated time steps (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Irregular time steps (%)	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Missing time steps (%)	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sensor failure time steps (%)	9.96	0.25	0.24	0.26	0.26	0.38	0.23	0.41
Total error data records (%)	49.89	16.71	16.71	16.71	16.71	16.71	16.71	16.71

such as previous day. The pre-processed time-series are then decomposed into three components including trend, seasonal, and remainder. It is the remainder time-series component which can be stationary and analyzed via the *X-bar* method to detect the possible outliers in the study area. Figures 4 and 5 illustrate the outliers that are detected with three-sigma of the bounds for flow and pressure data over 4 days, respectively. As elaborated earlier, the outliers that consecutively last for a user-specified time, e.g., 30 min, for each sensor are further classified into sensor events. Those detected events are finally classified as system anomaly events by the criteria of both HF (increased flows) and LP (decreased pressures) sensor events observed within the same time-window.

Using the proposed approach, two anomaly events have been detected as shown in Figure 6. The first anomaly event was detected on 23 September 2017. Using the monitoring data for this event, the anomaly event hotspots were localized as shown in Figure 7. The second event on 26 September 2017 was obvious and easy to detect due to the huge spike in the system inflow, as shown in Figure 6. It was a reported event of the 700-mm watermain break, which caused significant interruption of water service for the area. Coincidently, the reported 700-mm watermain break is at the same location of the hotspots as identified for the first event. Although the first event was not reported because it did not cause a visible watermain break with water flowing in the street, it indeed indicates the hidden burst that should have been timely mitigated before it became a disruptive watermain break. Therefore, the results of the first anomaly event demonstrate that the reported pipe burst on 27 September 2017 was successfully detected and localized on 23 September 2017, 3 days before the big watermain break occurred.

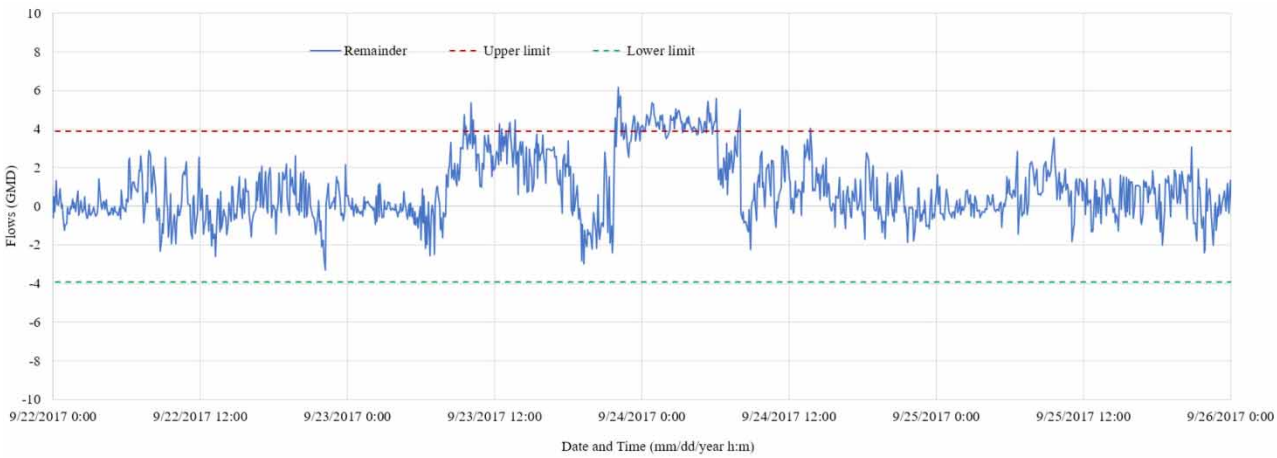

Figure 4 | Illustration of flow outliers detected with decomposed flow time-series.

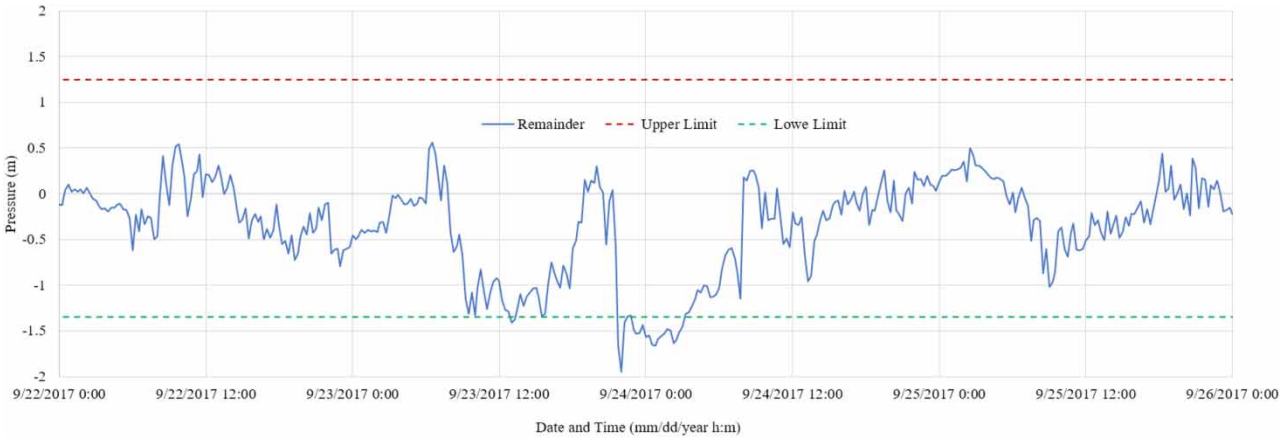

Figure 5 | Illustration of flow outliers detected with decomposed pressure time-series.

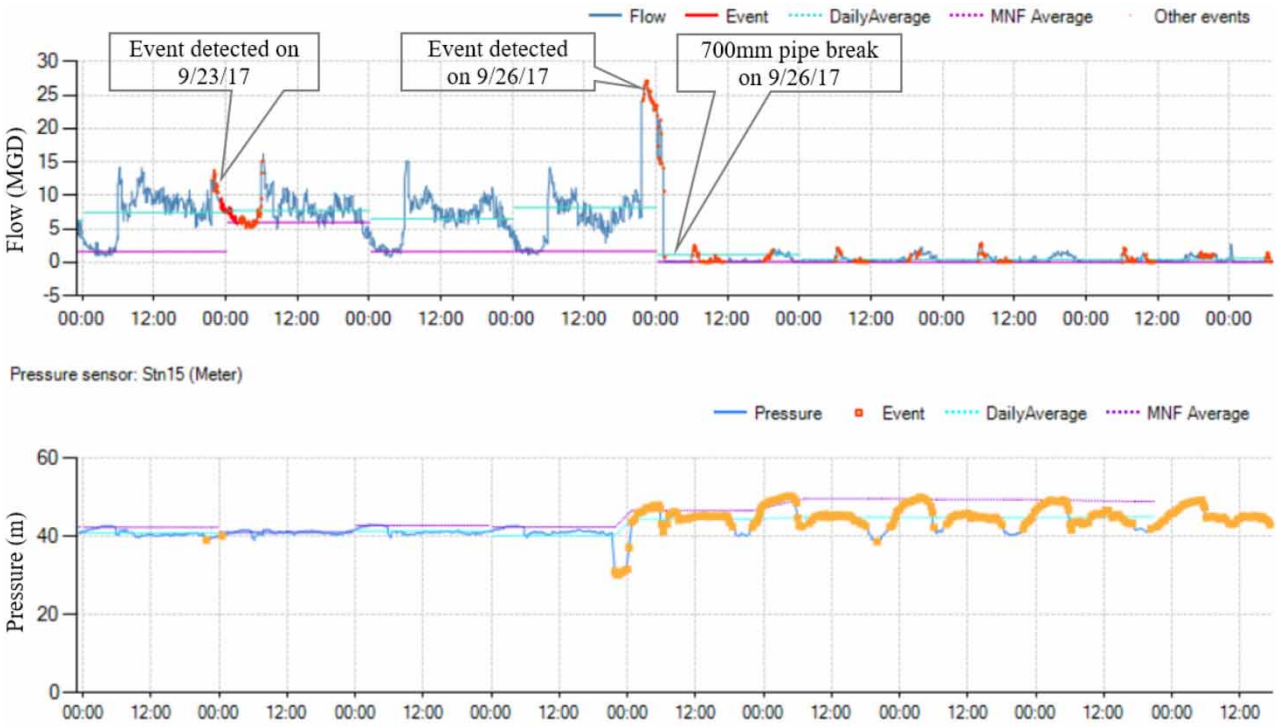

Figure 6 | Anomaly detection results for Case II.

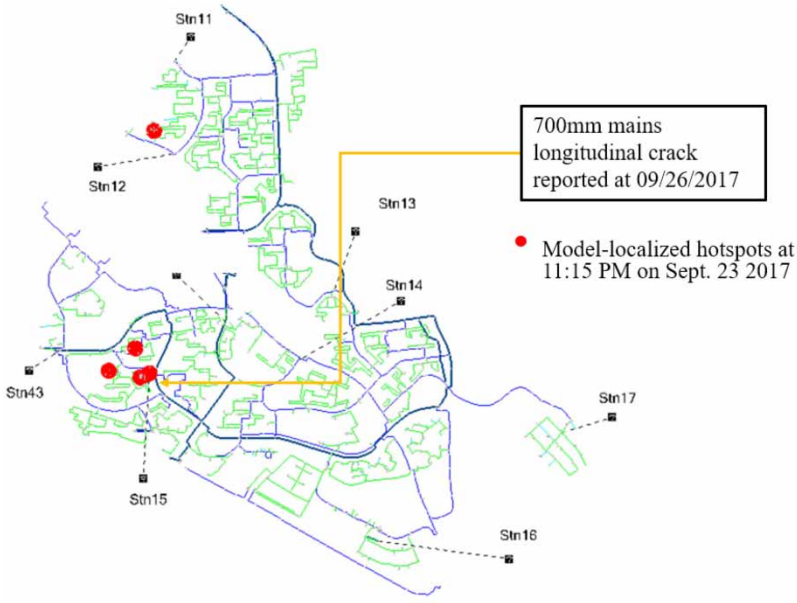

Figure 7 | Anomaly localization and field verification for Case II.

The case studies have proved that the developed methods are effective at detecting and localizing anomaly events and that the integrated solution has been successfully validated with good outcomes for real-world study areas. Thus, it is believed that the developed methods and the integrated tools are effective for practical applications of anomaly detection and localization in urban water systems.

DISCUSSIONS

With the applications of the developed framework in two areas of a large water system, a great deal of lessons has been learned. First of all, before applying any data-driven model or analytics, it is ultimately important to pre-process raw time-series data to eliminate all possible data errors and/or bad data. The time-series data derived from the case studies tested in this research seem to indicate that the methods for handling the data errors are adequate, but it is certainly desirable to test the approach on more use cases with various time-series data, so that a comprehensive and robust tool can be developed for processing various data errors in practice. Second, SPC methods are effective at outlier detection. It is easy to apply the methods, which work well with short time-series data too such as 3-month dataset for the second case study. But it is essential to decompose the time-series to ensure the stationarity requirement. Third, not all outliers are anomaly events, especially those isolated ones at single-time steps. The detected outliers must be analyzed for each sensor and correlated with all the sensors to classify a genuine system anomaly event. Hundreds of outliers have been detected for the first case study, by applying the simple heuristic rule, such as the minimum duration of 30 min for the outliers to be continuously detected for one sensor (flow or pressure). The outliers are categorized into several anomaly events per sensor. Correlating the anomaly events of all flow and pressure sensors, namely HF anomaly, and LP anomaly within the same time-window, only a few key anomaly events are subsequently identified and verified for the focused system. Finally, the localization of anomaly events via hydraulic model calibration has proved to be useful for improving the efficiency of pinpointing the exact anomaly in the field. It is noticed that some hotspots are consistently localized and worthwhile being verified in the field. This is because the LDPD approach (Wu *et al.* 2010) can localize not only the anomaly events that occur as detected in the current time-series data but also the hotspots that might have lasted for a long time.

CONCLUSION

As elaborated and demonstrated in this paper, the proposed approach is proved to be effective at analyzing the monitoring data for flow and pressure to detect anomaly and calibrating the hydraulic model for localizing anomaly hotspots. The proposed approach has potential to enhance the returns for PUB on the present smart water grid investment. In particular, the solution approach can facilitate the near real-time analysis of near real-time sensor data and improves the detection and localization of the anomaly events.

ACKNOWLDGEMENT

This research is supported by the Singapore National Research Foundation under its Competitive Research Programme (CRP) (Water) and administered by PUB (PUB-1804-0087), Singapore's national water agency.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Alegre, H., Hirner, W., Baptista, J. M. & Parena, R. 2000 *Performance Indicators for Water Supply Services. IWA Manual of Best Practice*. IWA Publishing, London.

Bentley Systems, Incorporated 2018 *WaterGEMS CONNECT Edition User Manual*. Bentley Systems, Incorporated, Exton, PA, USA.

Brunone, B. & Ferrante, M. 2001 Detecting leaks in pressurized pipes by means of transients. *Journal of Hydraulic Research* **39** (5), 539–547.

Cleveland, R. B., Cleveland, W. S., Mcrae, J. E. & Terpenning, I. 1990 STL: a seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* **6** (1), 3–73.

Huttona, C. J. & Kapelan, Z. 2015 Real-time burst detection in water distribution systems using a Bayesian demand forecasting methodology. *Procedia Engineering* **119** (2015), 13–18.

Jung, D. & Lansey, K. 2015 Water distribution system burst detection using a nonlinear Kalman filter. *ASCE Journal of Water Resources Planning and Management* **141** (5), 04014070.

Li, Q., Wu, Z. Y. & Rahman, A. 2019 Evolutionary deep learning with extended Kalman filter for effective prediction modeling and efficient data assimilation. *ASCE Journal of Computing in Civil Engineering* **33** (3), 04019014.

Loureiro, D., Amado, C., Martins, A., Vitorino, D., Mamade, A. & Coelho, S. T. 2016 Water distribution systems flow monitoring and anomalous event detection: a practical approach. *Urban Water Journal* **13** (3), 242–252.

Meeus, S. & Marshallsay, D. 2017 'A real time system for detecting events in water networks.' CCWI 2017, the 5th–7th Sept. 2017, Sheffield, UK.

Mounce, S. R. & Boxall, J. 2011 Online monitoring and detection. In: *Water Loss Reduction* (Wu, Z. Y., ed.). Bentley Institute Press, Exton, PA, USA. ISBN: 978-1-934493-08-3.

Mounce, S., Day, A., Wood, A., Khan, A. D., Widdop, P. & Machell, J. 2002 A neural network approach to burst detection. *Water Science and Technology* **45** (4–5), 237–246.

Mounce, S. R., Mounce, R. B. & Boxall, J. B. 2011 Novelty detection for time series data analysis in water distribution systems using support vector machines. *Journal of Hydroinformatics* **13** (4), 672–685.

Mounce, S. R., Mounce, R. B., Jackson, T., Austin, J. & Boxall, J. B. 2014 Pattern matching and associative artificial neural networks for water distribution system time series data analysis. *Journal of Hydroinformatics* **16**, 617–632.

Nixon, W., Ghidaoui, M. S. & Kolyshkin, A. A. 2006 Range of validity of the transient damping leakage detection method. *Journal of Hydraulic Engineering* **132** (9), 944–957.

Perez, R., Sanz, G., Puig, V., Quevedo, J., Escorfet, M. A. C., Nejjari, F., Meseguer, J., Cembrano, G., Tur, J. M. M. & Sarrate, R. 2014 'Leak Localization in Water Networks: A Model-Based Methodology Using Pressure Sensors Applied to a Real Network in Barcelona.' IEEE Control Systems Magazine.

Ponce, M. V. C., Castanon, L. E. G. & Cayuela, V. P. 2014 Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities. *Journal of Hydroinformatics* **16** (3), 649–670.

Public Utilities Board (PUB) Singapore 2016 Managing the water distribution network with a smart water grid. *Smart Water* **1**, 4. doi:10.1186/s40713-016-0004-4.

Rodriguez, A. & Laio, A. 2014 Clustering by fast search and find of density peaks. *Science* **344** (6191), 1492–1496.

Romano, M., Kapelan, Z. & Savic, D. A. 2014a Automated detection of pipe bursts and other events in water distribution systems. *ASCE Journal of Water Resources Planning and Management* **140** (4), 457–467.

Romano, M., Kaplena, Z. & Savic, D. A. 2014b Evolutionary algorithm and expectation maximization strategies for improved detection of pipe bursts and other events in water distribution systems. *ASCE Journal of Water Resources Planning and Management* **140** (5), 572–584.

Soldevila, A., Fernandez-Canti, R. M., Blesa, J., Tornil-Sin, S. & Puig, V. 2017 Leak localization in water distribution networks using Bayesian classifiers. *Journal of Process Control* **55**, 1–9.

Sophocleous, S., Savic, D. & Kapelan, Z. 2019 Leak localization in a real water distribution network based on search-space reduction. *ASCE Journal of Water Resources Planning and Management* **145** (7), 1943–5452.0001079.

Tao, T., Huang, H. D., Li, F. & Xin, K. L. 2014 Burst detection using an artificial immune network in water-distribution systems. *ASCE Journal of Water Resources Planning and Management* **140** (10), 04014027.

Vrachimis, S. G., Timotheou, S., Eliades, D. G. & Polycarpou, M. M. 2021 Leakage detection and localization in water distribution systems - a model invalidation approach. *Control Engineering Practice* **110**, 104755.

Wu, Z. Y. 2009 A unified approach for leakage detection and extended period model calibration of water distribution systems. *Urban Water Journal* **6** (1), 53–67.

Wu, Z. Y. & He, Y. 2021 Time series data decomposition-based anomaly detection and evaluation framework for smart water system management. *ASCE Journal of Water Resources Planning and Management*. doi:10.1061/(ASCE)WR.1943-5452.0001433.

Wu, Z. Y., Walski, T., Mankowski, R., Cook, J., Tryby, M. & Herrin, G. 2002 Calibrating water distribution model via genetic algorithms. In: *Proceedings of the AWWA IMTech Conference*, April 16–19, Kansas City, MI.

Wu, Z. Y., Sage, P. & Turtle, D. 2010 Pressure dependent leakage detection approach and its application to district water systems. *ASCE Journal of Water Resources Planning and Management* **136** (1), 116–128.

Wu, Y., Liu, S., Smith, K. & Wang, X. 2018 Using correlation between data from multiple monitoring sensors to detect bursts in water distribution systems. *ASCE Journal of Water Resources Planning and Management* **144** (2), 1943–5452.

Ye, G. & Fenner, A. 2014 Weighted least square with expectation-maximization algorithm for burst detection in U.K. water distribution systems. *ASCE Journal of Water Resources Planning and Management* **140** (4), 417–424.

First received 20 July 2021; accepted in revised form 19 November 2021. Available online 16 December 2021