A semi-automated approach to validation and error diagnostics of water network data

Supporting Information (SI)

Jonas Kjeld Kirstein[1], Klavs Høgh, Martin Rygaard and Morten Borup

The aim of the SI is to offer additional insight into selected methods and results in the paper '*A semi-automated approach to validation and error diagnostics of water network data'*.

We refer to the corresponding section of the main paper at the beginning of each section of the SI.

Contents: Supporting information

[1] E-mail: jkir@env.dtu.dk; Technical University of Denmark, Department of Enviromental Engineering, Bygningstorvet, Building 115, 2800 Kgs. Lyngby, Denmark.

A. Anomaly sources

Section 2.1 of the main paper

Table A-1 lists major anomaly sources that contribute to errors and irregularities captured in the raw meter data sets. In addition, it is stated which of the applied tests in the anomaly testing were applied and designed to identify mainly a certain type of anomaly.

Table A-1. Potential anomaly sources that contribute to errors and irregularities and the tests applied to detect their main occurrence in raw meter data. Anomaly sources based on actual events in the network are not considered.

Anomaly source	Anomaly background	Main tests
 **Operator**	Wrong time settings (incl. 'drifting' clocks) are an effect of erroneous synchronization or manual manipulation of the internal clock, incorrect adjustment for daylight saving time, etc. These settings can bias data applications such as DMA balances and reduce the overall value of smart online control of the water network.	I, VII
 **Sensor**	Many data applications depend on complete data streams. Errors within transmission and storage procedures, power outages, etc. can lead to illegitimate and missing data periods and highlight systematic problems of data acquisition.	II, VI
 **Transmission**  **Storage/application**	The miscalibration and loss in sensitivity of sensors, mechanical errors, and observations above or below cut-off values/deadbands decrease the overall reliability of the collected data streams. Among other things, these anomalies typically highlight meter malfunctions.	III–V

B. Anomaly testing - additional information

Section 2.2 of the main paper

TEST IV) A description of the flagging procedure in the rate of change test (test IV) is summarized in the following:

1) Identify all $u \in t_n$ where $|\Delta x|/\Delta t \geq \theta$.

2) Split u into v sets of consecutive numbers.

3) Flag all first entries within v in $\mathbf{M}_4$.

4) Return to step 1 and leave the flagged values out of the analysed data set. Stop, if no further u are found.

C. Data set description

Section 3 and 4 of the main paper

C.1 Time series examples

To provide insight on the analysed data series from the three utilities, we grouped measurements from selected meter data sets into a week's 168 hours. Then, to illustrate the variation in demand/flow and pressure, the percentile distribution was computed for each hour. The outcome for flow and pressure time series is shown in Figure C-1 and C-2, respectively.

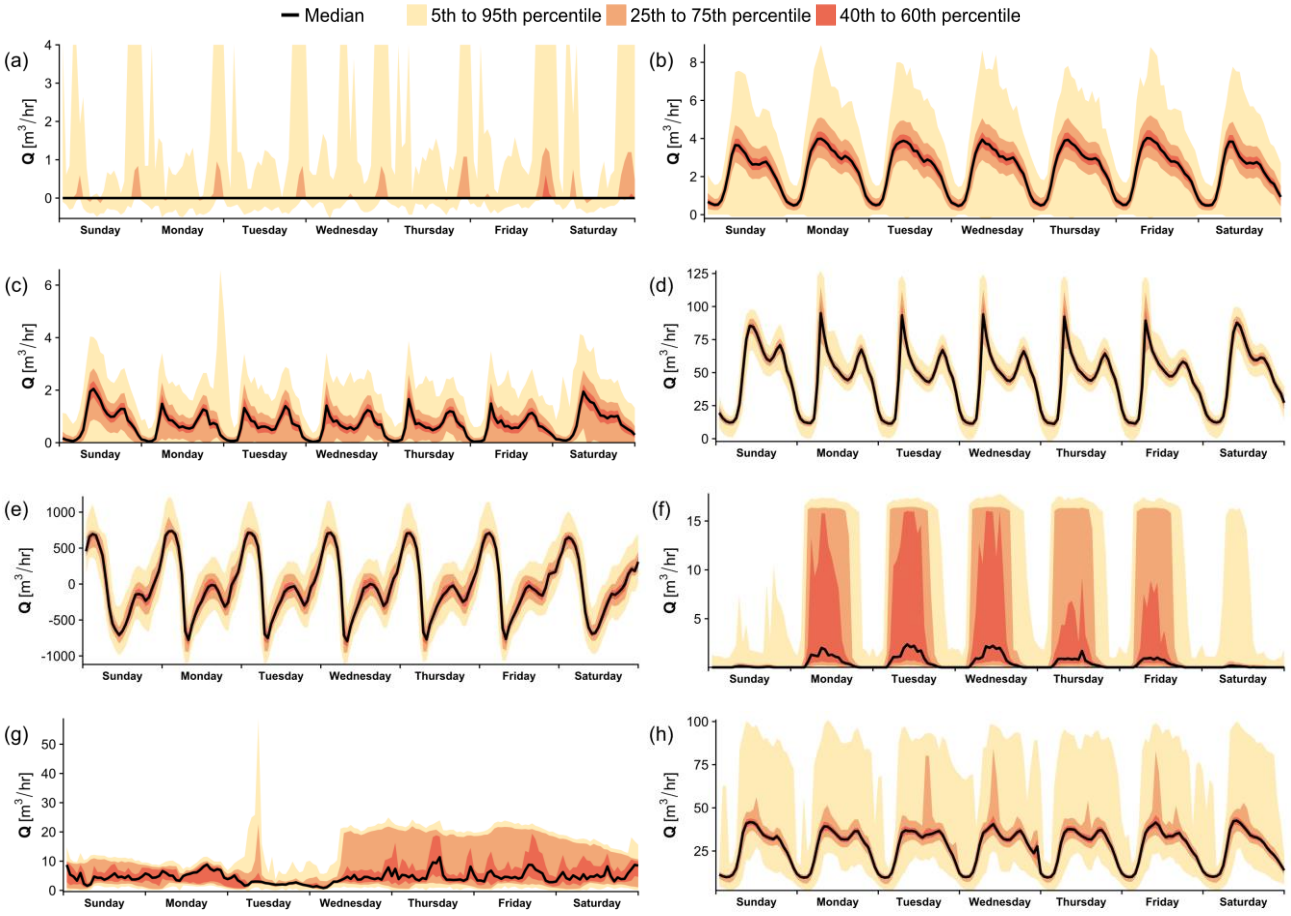


Figure C-1. Example of eight flow time series applied in the case study. The percentile distribution is based on having grouped all measurements into a week's 168 hours.

Figure C-1 shows that there is a great variety in the analysed flow time series. For example, the median flow and lower percentiles in Figure C-1a reflect that no flow is usually measured at all; however, there is a tendency towards higher flow rates during the hours before midnight, as shown by the 25th to 75th percentiles. Other time

series show a distinct daily pattern, as in Figure C-1c; this is the result of changing flow rates between summer and winter months. Figure C-1d shows an observation point with minor seasonal variations. Whereas these time series have a residential pattern, Figure C-1f shows high deviations in flow only during working days. As with flow, in selected pressure series a distinct daily pattern is visible, pressure being higher during the night than over the course of the day (e.g. Figure C-2d and C-2g). The pressure series in Figure C-2b and Figure C-2h indicate a higher share of anomalies, as a large proportion of measurements is around or even below 0 bar.

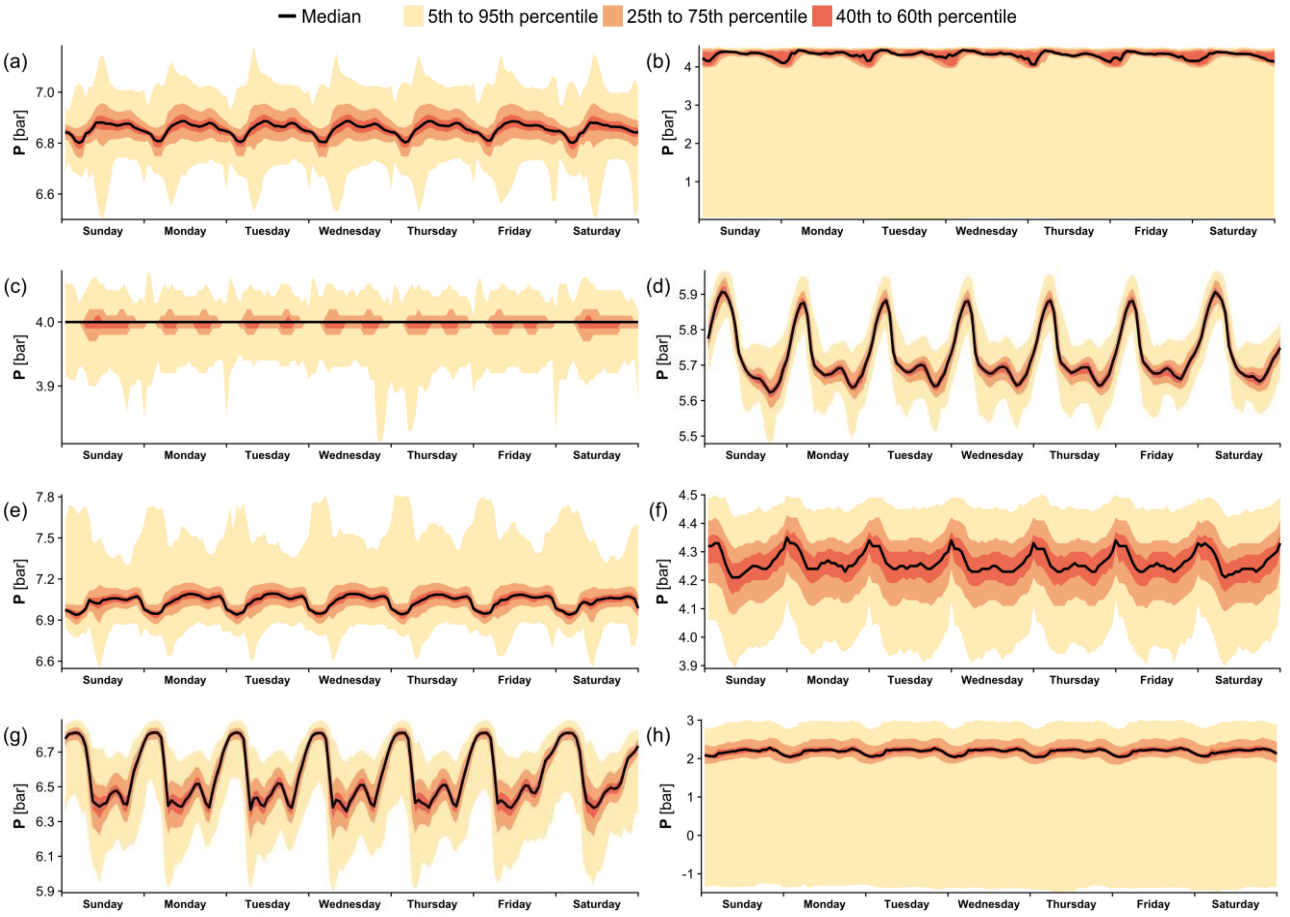


Figure C-2. Example of eight pressure time series applied in the case study. The percentile distribution is based on having grouped all measurements into a week's 168 hours.

C.2 Comparison of raw and validated data

All validated data in this section is based on the same parameter selection as summarized in Table 2. Table C-1 summarizes some key features of the applied data sets before and after data validation. It can be seen that the largest data sets were collected by utility C. Most of the data sets of utilities A and C are collected in 1-min timestamp intervals, whereas the highest share of data from utility B is stored at 5- or 15-min intervals.

Our data validation found the highest number of invalid data points (29.88%) in flow measurements of utility B.

Table C-1. Total number of data points in flow (Q) and pressure (P) meter data sets from utility A-C before and after data validation including timestamp interval distribution.

Utility	Type	Figure		Total data points	Timestamp interval distribution [%] 1 [min]	5 [min]	10 [min]	15 [min]	60 [min]	Other [min]
Raw data										
A	Q	Figure C-2(a)		18,150,889	91.15	0.37	6.11	0.46	0.93	0.99
	P			15,336,944	92.45	<0.0	6.06	0.41	0.22	0.85
B	Q	Figure C-3(a)		2,842,734	<0.0	53.85	<0.0	46.12	<0.0	0.03
	P			2,848,399	<0.0	53.78	<0.0	46.19	<0.0	0.03
C	Q	Figure C-4(a)		71,249,575	97.80	0.07	0.02	0.29	<0.0	1.82
	P			86,626,723	96.50	0.03	<0.0	0.02	<0.0	3.45
Validated data (excluding anomalies flagged by test I-VI)										
A	Q	Figure C-2(b)	(-8.09%)	16,682,587	91.40	0.30	6.60	0.49	0.97	0.23
	P		(-4.85%)	14,592,458	93.38	<0.0	5.96	0.34	0.19	0.12
B	Q	Figure C-3(b)	(-29.88%)	1,993,408	<0.0	67.46	0.02	30.93	0.03	1.57
	P		(-7.05%)	2,647,703	<0.0	54.14	0.06	45.46	<0.0	0.34
C	Q	Figure C-4(b)	(-4.52%)	68,030,580	99.21	0.04	0.01	0.01	<0.0	0.72
	P		(-8.75%)	79,043,261	97.83	0.11	0.04	0.02	<0.0	2.00

Figure C-3 exemplifies the possible distribution of raw data collected in selected flow and pressure meters for each of the three utilities before and after application of data validation (tests I–VI). Each of the six histograms represents a unique distribution pattern and it is difficult to identify clear similarities in the underlying distributions; however, some features are shared between selected meters, such as a lower bound of zero in the flow meters of utilities A and C. Moreover, all raw data histograms, except the pressure meter in utility C, indicate that a large proportion of type I and II anomalies have been stored in the data. For example, in the shown (raw data) pressure meter of utility B, more than 10,000 data points have a pressure value below 0 bar and a smaller proportion includes values around 30 bar. Observations within this range are highly unlikely, considering the utility's pressure meters are located on distribution mains inside the city. After data validation, a large number of anomalies have been removed from the histograms. In terms of pressure data, all 0-bar data was deemed invalid in utility A. In utility B, a reduced amount of infeasible pressure measurements is still kept in the data, demanding a stricter choice of parameters. In the case of utility C, no values were flagged. Similarly, the number of 'clear' anomalies decreased in the three flow meters.

The histograms with validated data show the problems behind a global parameter selection of the different tests when applied to utilities with varying network setups. As in the case of pressure measurements, a simple test of min/max measurements above or below a certain threshold could have removed the remaining infeasible measurements. Thus, including specific system knowledge could increase the value of the validation method. In some cases, this knowledge is not available and a stricter choice of parameters could be a solution, although this could also increase the detection of false positives. In the case of the shown pressure meter in utility C, the selection of parameters has not negatively affected the anomaly detection visually, i.e. there was no flagging of values that appear correct.

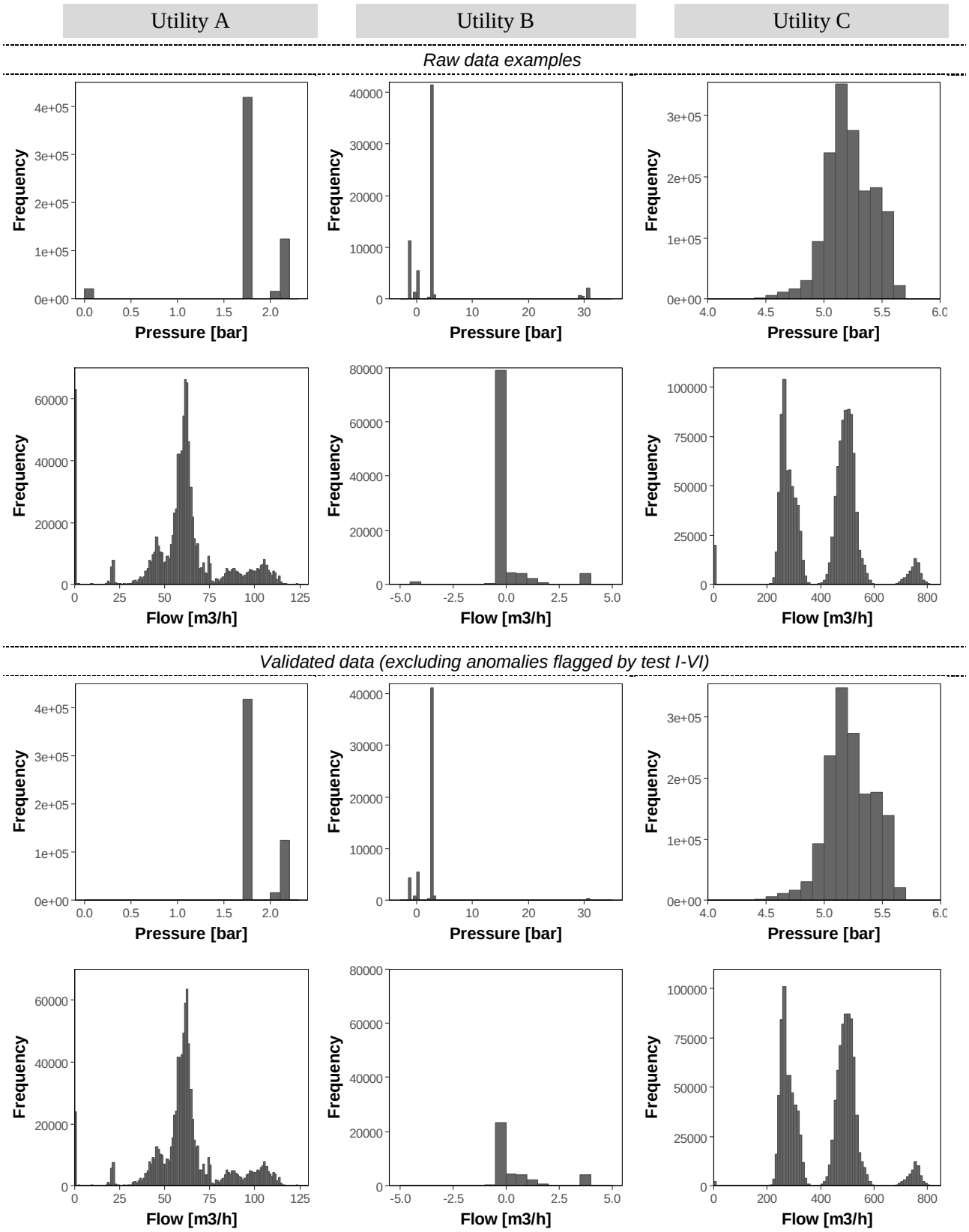


Figure C-3. Flow and pressure data histograms from six selected meters in utility A-C before and after data validation. Owing to the selected histogram bin size and margins, a minor share of data is not shown at maximum 0.06% (in this case raw pressure data in utility C).

Figure C-4 to Figure C-6 show the raw and validated data sets (tests I–VI) of utilities A–C distributed on daily, weekly and monthly time scales. Even though around 8% of the flow data (Table C-1) was flagged as anomalous in utility A (Figure C-4a) the data validation has no clear effect on the distribution of the utilities' flow data (Figure C-4b). Also, a daily pattern and some seasonality are visible in terms of higher summer consumption. The raw pressure data show a clear deviation between the 50th and 75th percentiles in November and December, and in the 23th hour compared to the remaining time steps. This can be explained by the fact that parts of the data collection system stopped working properly at the end of October 2016 and first continued functioning at the beginning of 2017. Moreover, in utility A, often at 23.00 hours each day, invalid (non-numeric) data was collected. Having applied the data validation tests, the 75th percentile decreased notably in February, in October and in the first hour of the day. This can partly be explained by around 25% of the pressure data being metered around 4 bar and the remaining 75% around 2 bar. Even slight deviations in the number of flagged values can move the 75th percentile. In February, for example, multiple timestamp duplicates (test I) were flagged in the high pressure meters, moving the 75th percentile towards a lower pressure. Among other things, a higher number of flatline anomalies (test V) had the same effect in the first hour.

The raw data of utility B is displayed in Figure C-5a. With regards to flow data, low flow hours (20.00 to 04.00 hours), certain days (Saturday and Sunday) and certain months (December–May) are particularly prone to having a median flow of around 0 m^3/h. Having applied the data validation procedure (Figure C-5b), the median flow increased on all time scales. This is mainly because of the high percentage of flatline anomalies (test V) flagged in utility B that occur during low flow periods. The raw and validated flow data display a daily flow pattern as seen in utility A. However, the median has increased significantly in the validated data sets, indicating a good performance of the selected test parameters. Also, a very high pressure measured in April, May and June (Figure C-5a) was no longer visible after data validation (Figure C-5b). Possible explanations include anomalies flagged by tests III and IV. Finally, the data validation has not flagged a certain percentage of negative pressure values. This can be explained by the fact that the majority of pressure values in selected meter data sets, erroneously, are negative. As in utility A, there is no general difference in the observed pressure on different time scales.

In the case of utility C, there are only minor differences between the raw data set (Figure C-6a) and the validated data set (Figure C-6b). This can partly be explained by a more proactive approach in utility C to repair and solve issues within the data acquisition in a reasonable amount of time, reducing the overall number of anomalies in the raw data sets. As in utility B, bidirectional flows are also captured in the data sets, as can be seen in the hourly flow plot. The greatest difference between raw and validated data sets can be seen in the 2.5^{th} pressure percentile on a monthly scale. In the validated data set, the lower hinge is more constant around 50 mWC. This stabilization was most likely caused by anomalies flagged by test V. A daily flow and pressure pattern is visible in the raw and validated data sets, but cannot be seen on a larger time scale.

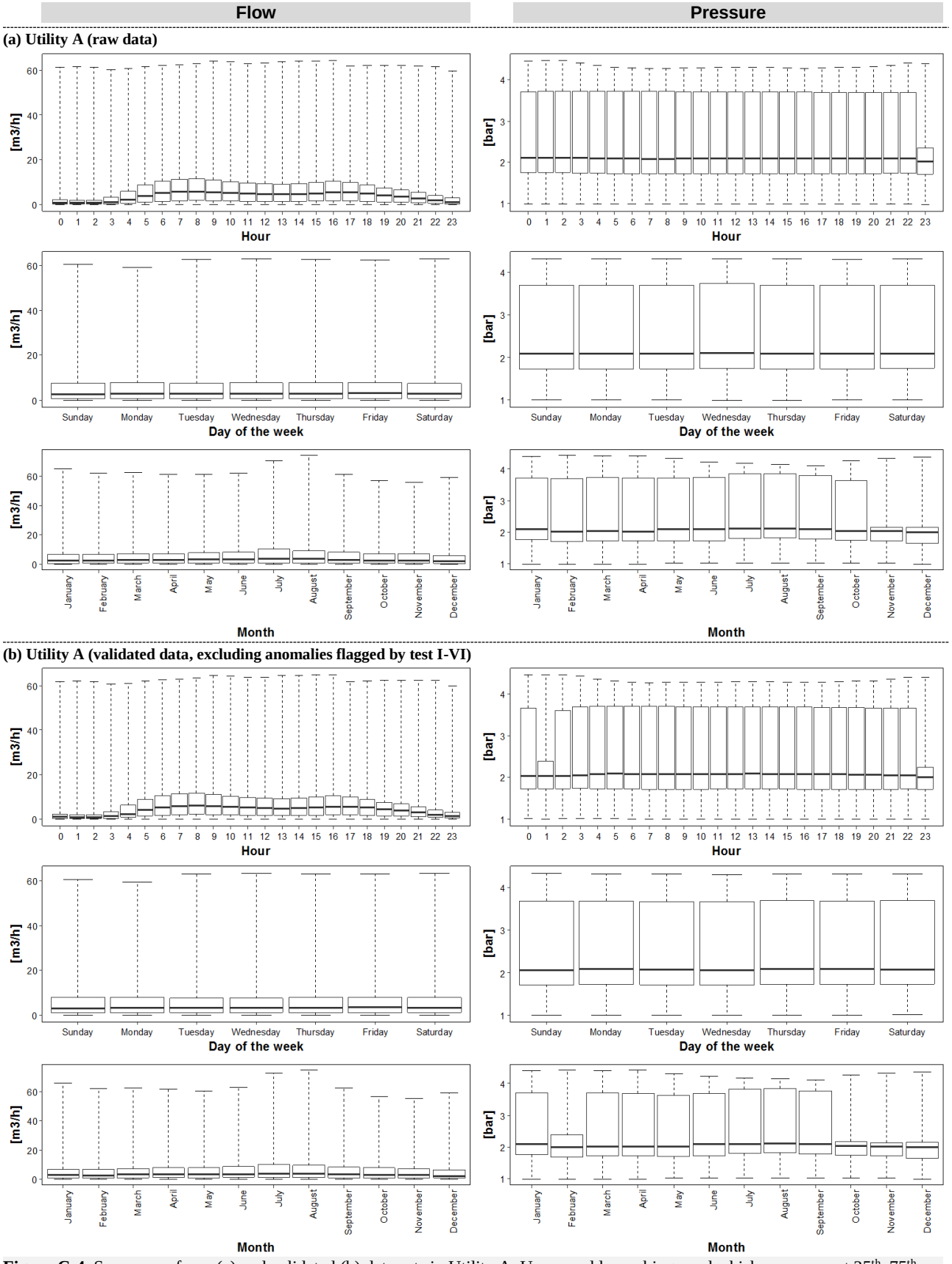


Figure C-4. Summary of raw (a) and validated (b) data sets in Utility A. Upper and lower hinge and whiskers represent 25th, 75th, 2.5th and 97.5th percentiles. The line across the box displays the median. Except non-numeric values, no outliers have been removed from the raw data analysis. Values below the 2.5th and above the 97.5th percentiles are not displayed.

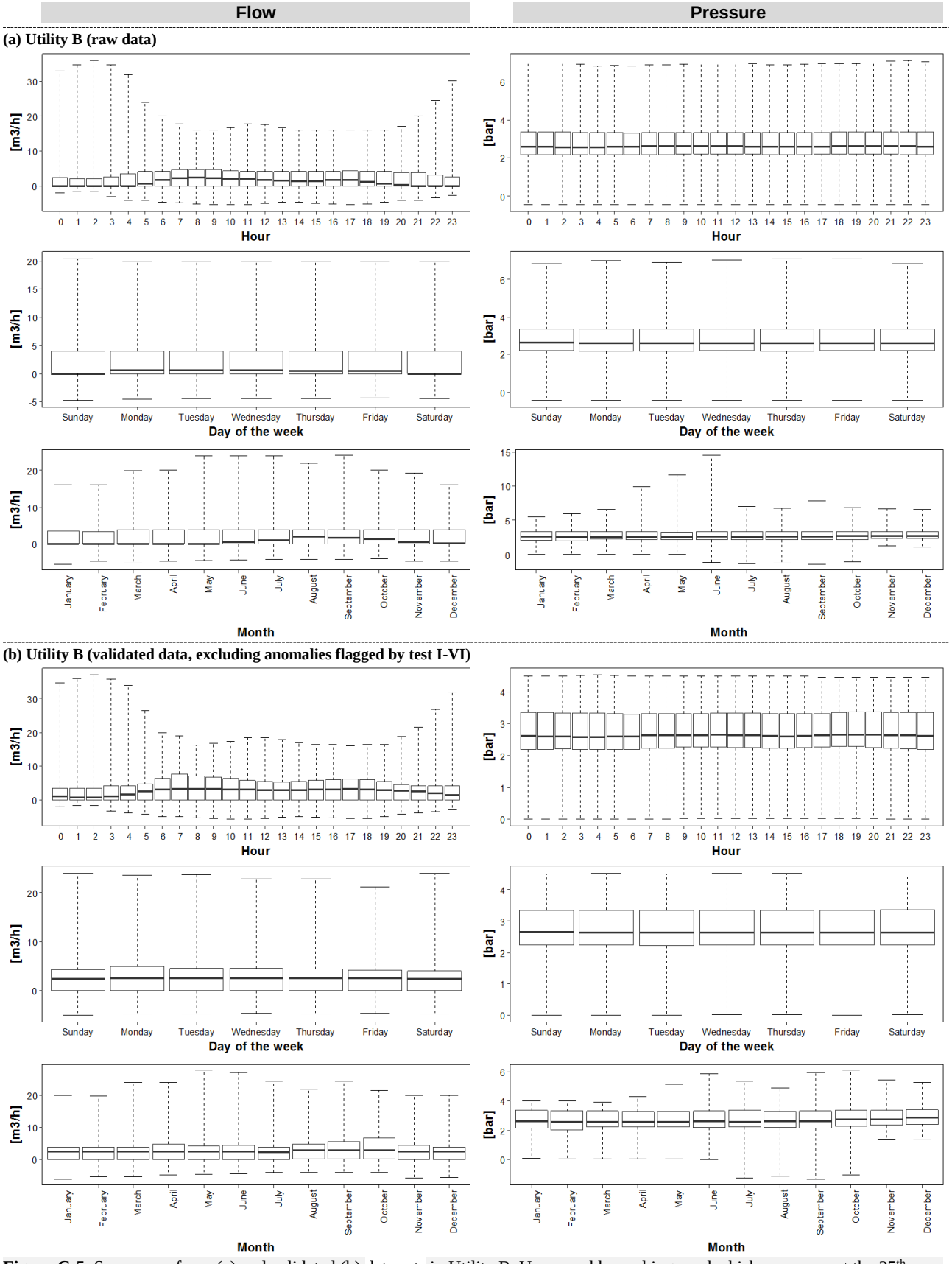


Figure C-5. Summary of raw (a) and validated (b) data sets in Utility B. Upper and lower hinge and whiskers represent the 25th, 75th, 2.5th and 97.5th percentiles. The line across the box displays the median. Except non-numeric values, no outliers have been removed from the raw data analysis. Values below the 2.5th and above the 97.5th percentiles are not displayed.

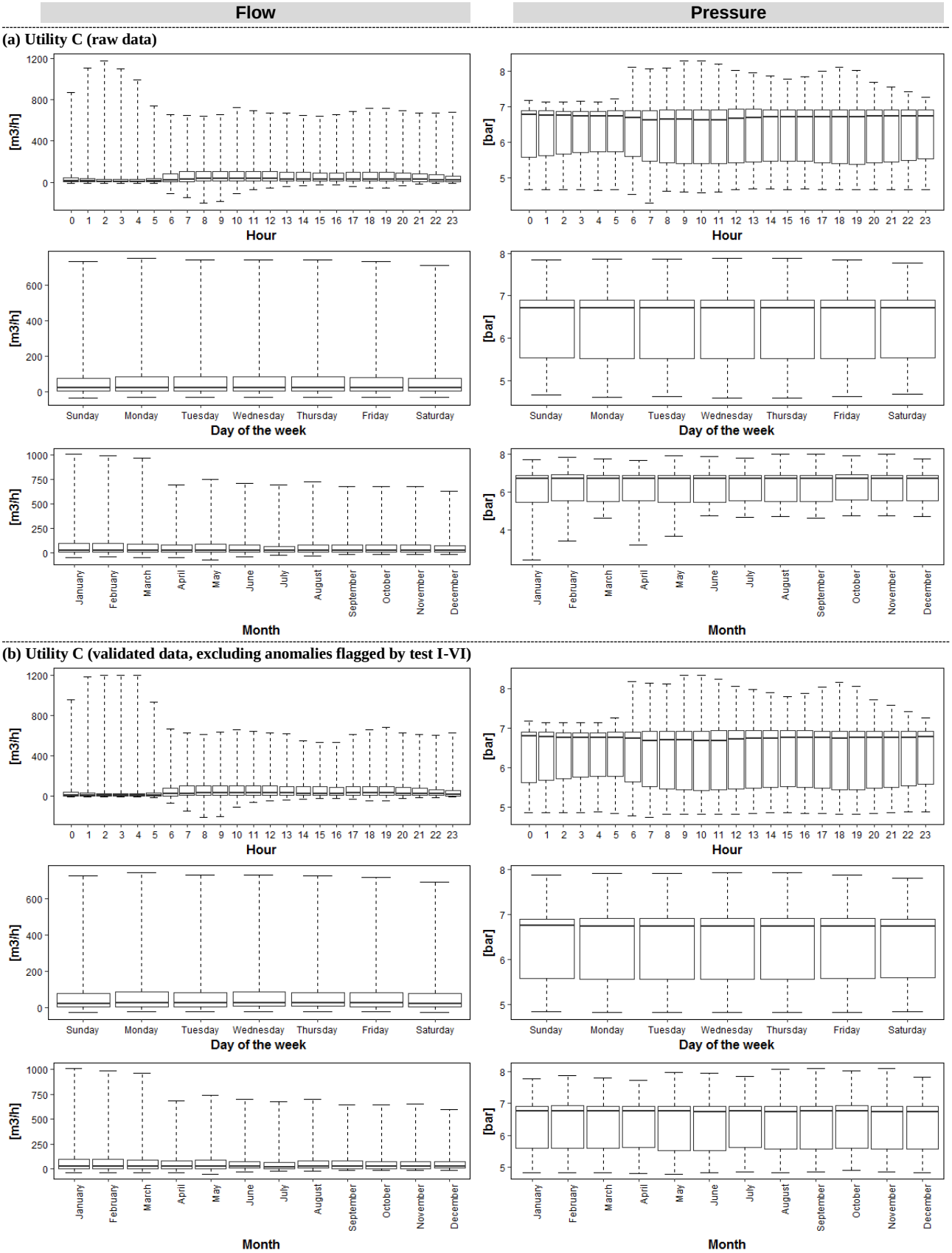


Figure C-6. Summary of raw (a) and validated (b) data sets in Utility C. Upper and lower hinge and whiskers represent the 25th, 75th, 2.5th and 97.5th percentiles. The line across the box displays the median. Except non-numeric values, no outliers have been removed from the raw data analysis. All values below the 2.5th and above the 97.5th percentiles are not displayed.

D. Sensitivity of parameters applied in the anomaly testing

Section 2.2, 4.1 and 4.2 of the main paper

It is almost inevitable that erroneously flagged data will be contained in the malfunction indicator database (MAID). In general, incorrect flagging of meter data occurs when test parameters are chosen that are too sensitive. Moreover, the range (test III) and rate of change test (test IV) depend on the historical distribution of the data. In such a case, it is difficult to avoid flagging of type 3 anomalies if the analysed data set covered periods where the system behaviour changed drastically. Thus, to reduce false alarm rates and flagging of type 3 anomalies, parameters had to be chosen carefully in selected tests of the anomaly testing framework. First, the results of different parameter combinations for tests III–VI are shown and discussed, following which examples are given of parameter combinations affecting the applied conditions of test VII.

D.1 Tests III–VI: Parameter sensitivity

Figure D-1 and Figure D-2 illustrate the sensitivity of the anomaly tests III–VI to the variation of parameters within the selected tests for all flow and pressure meters, respectively. Each column illustrates how the average percentage of flagged data points changed by a given parameter set. In the range test (test III) and change in rate test (test IV), higher percentile rates in combination with increasing the values of α/β and λ led as intended to a lower number of flagged values.

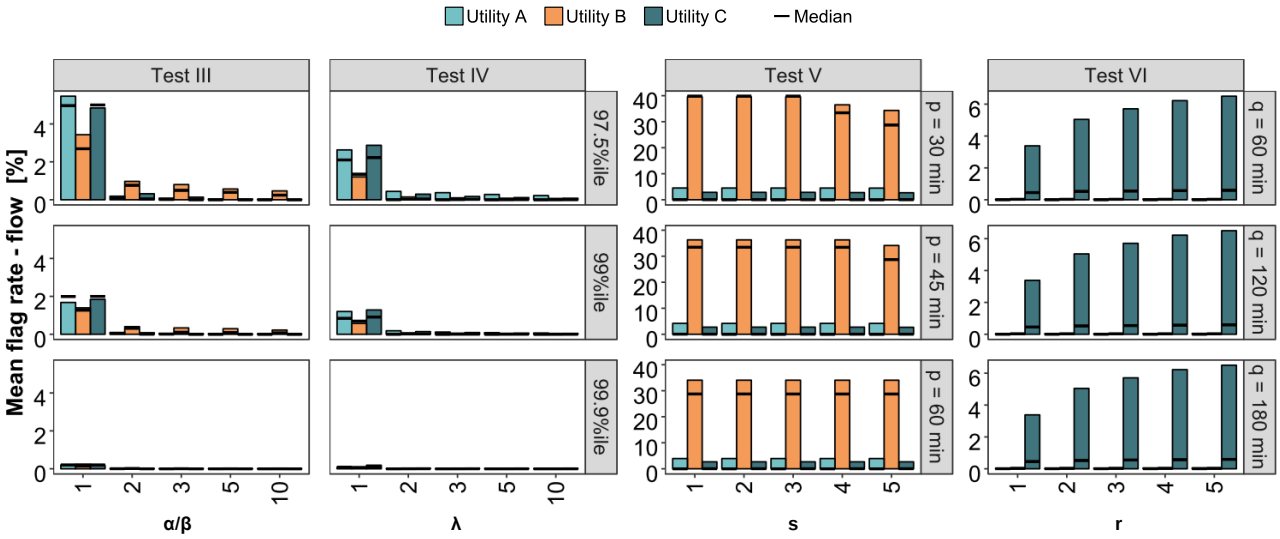


Figure D-1. Effect on error rates of all flow meters by varying test parameters on the range test (test III), change in rate test (test IV), flatline test (test V), and timestamp inconsistency test (test VI), based on the raw data from three utilities.

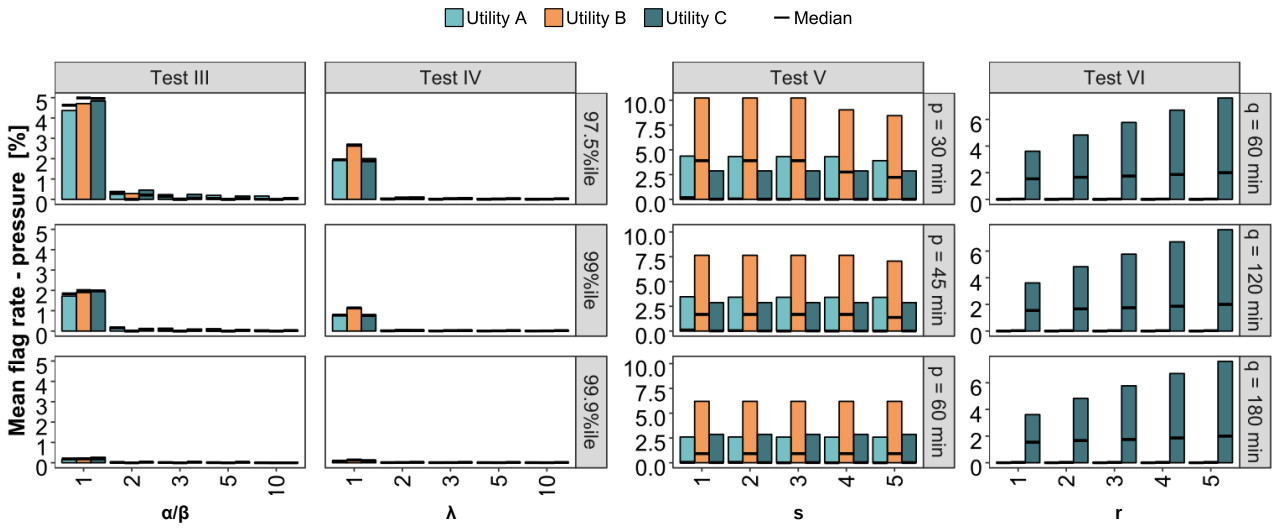


Figure D-2. Effect on error rates of all pressure meters by varying test parameters on the range test (test III), change in rate test (test IV), flatline test (test V), and timestamp inconsistency test (test VI), based on the raw data from three utilities.

In general, the figures illustrate that certain types of anomaly vary between the utilities. In the case of near real-time applications, however, it is important for the utility to identify a set of parameters that does not generate too many false alarms while correctly identifying anomalies. As the variation of parameters has a clear impact on the total number of data points, the parameters need to be fine-tuned. A utility should not only fine-tune the test parameters based on the measured parameters, but also change the test parameters independently for certain meter groups. For example, meters with varying sampling intervals or different objectives (e.g. DMA inlet or emergency pump monitoring) will probably need different optimal parameter settings. The flatline test column (test V) shows that varying the length of time and number of consecutive steps having an identical value has less influence on the mean flag rate, meaning that the flatline segments captured by these parameter sets in general occur over a longer period. The figure also illustrates that utility B has a notably higher frequency of flatline segments than the other two. It is likely that decreasing *p* would increase the error rate in utilities A and C notably, as a larger share of their data was measured at a higher sampling frequency. In this study, the originally stored number of significant digits in each meter set was included in the flatline test. Likely, more flatline segments would have been detected, when a reduced number of significant digits was included in the test. Moreover, the timestamp inconsistency test (test VI) shows that, for utility C only, varying parameter *r* has a visible influence on the mean rate of flagged values. This can be explained by the fact that the data collection system in utility C deletes equal measurements of less than 15 minutes duration.

D.2 Test VII: Sensitivity of conditions

In the timestamp drift test (test VII), the number of meters where a drift had been identified was initially determined by two conditions (Sec. 2.2). The sensitivity of test VII was assessed by varying the four parameters used to raise a flag in the conditions, namely $C_{weekly}(d)$, the number of reference weeks w included, $P_{weekly}(d)$ and the number of subsequent days d. Figure D-3 and Figure D-4 illustrate the effect of changing these parameters on the total number of meters with a drift for all flow and pressure measurements, respectively. As with the sensitivity of tests III–VI, an increase in the individual parameters led to a decrease in the number of identified anomalies, i.e. meters with drift. Also, this test needs to be fine-tuned by sensitivity; certain patterns in a utility might occur that would raise the false alarm rate. It seems that an increase in $P_{weekly}(d)$ had the largest impact on the overall number of meters where a drift was identified. A similar effect was seen for $C_{weekly}(d)$, because an increasing threshold also reduces the amount of data available in the test. The application of rather loose conditions shows that most drifts were identified in utilities A and C, which is linked to the total number of meters installed. Interestingly, increasing the parameters has only a small effect on the overall number of meters with a drift in utility B. A clear change in the number of meters is first seen when $P_{weekly}(d)$ > 3 hours.

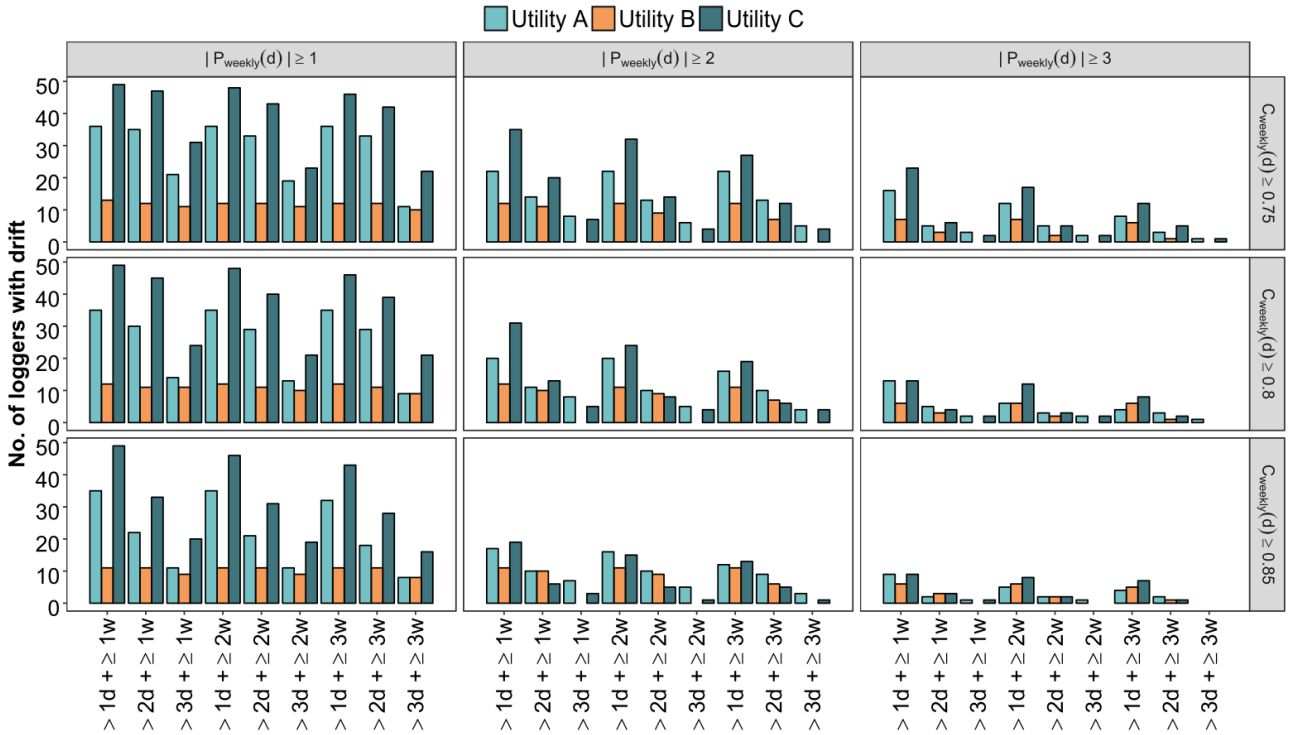


Figure D-3. Sensitivity analysis of the timestamp drift test (test VII) for all flow meters. Four parameters were varied: 1) the weekly correlation value [$C_{weekly}(d)$], before the data was accepted in the test; 2) the threshold hour before a test identifies a logger as drifting [$P_{weekly}(d)$]; 3) the number of consecutive days d in which the test has to identify a drift before a flag is raised (1d–3d); and 4) the number of reference weeks w (1w-3w) that have to agree on a drift before it is flagged.

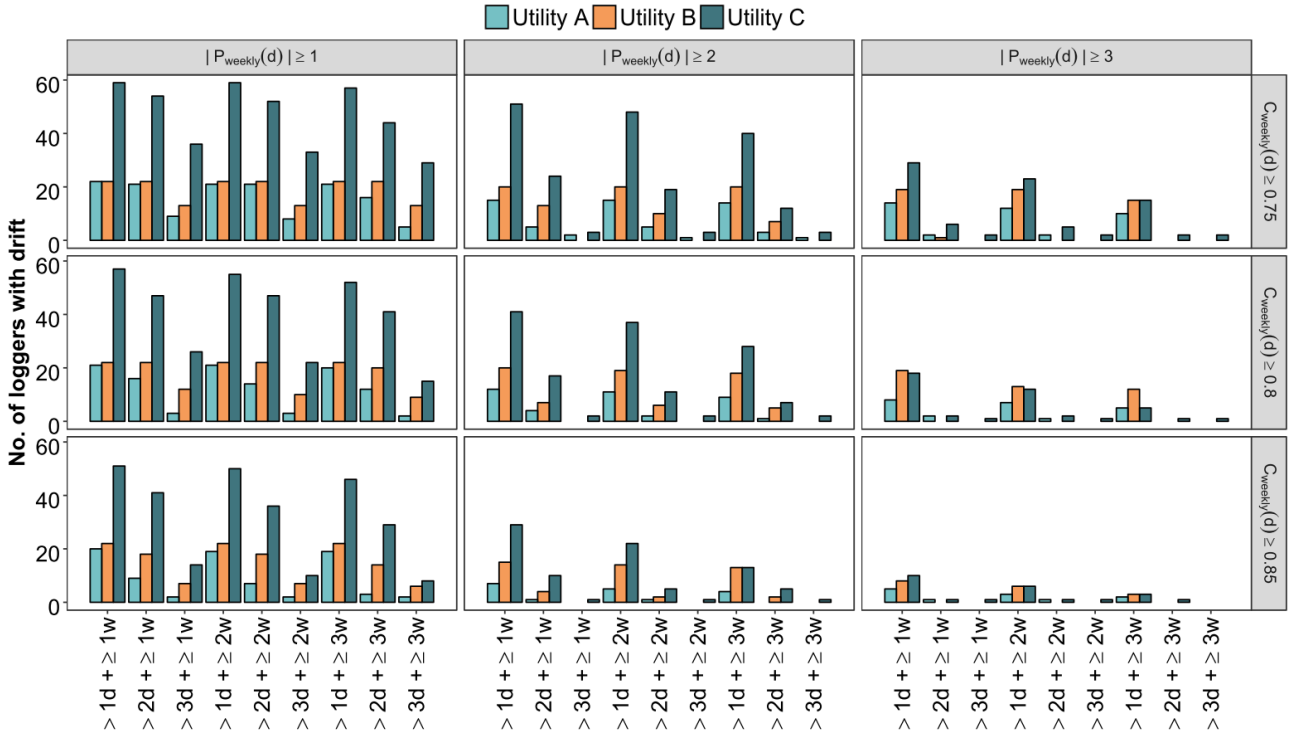


Figure D-4. Sensitivity analysis of the timestamp drift test (test VII) for all pressure meters. Four parameters were varied: 1) the weekly correlation value [$C_{weekly}(d)$], before the data was accepted in the test; 2) the threshold hour before a test identifies a logger as drifting [$P_{weekly}(d)$]; 3) the number of consecutive days d in which the test has to identify a drift before a flag is raised (1d–3d); and 4) the number of reference weeks w (1w-3w) that have to agree on a drift before it is flagged.

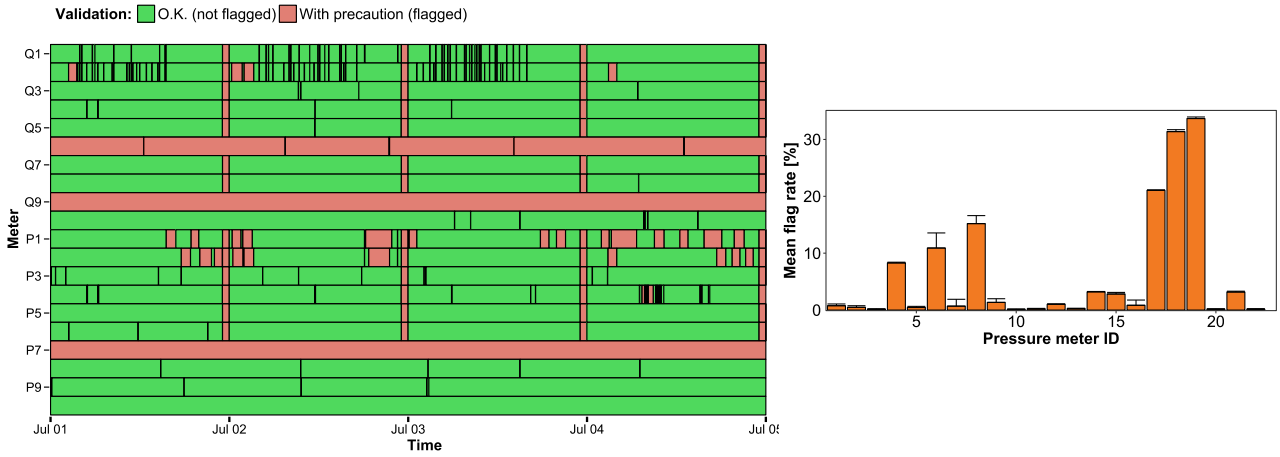
D.3 Test VII: Additional notes

The timestamp drift test is still under development and its reliability and feasibility need to be improved, as only a relatively low percentage of the collected data could be run by the test with the applied conditions. Another opportunity for improvement is where the test identifies drifts that, for example, reflect correct changes in the DMA set-up. Also, future applications should flag data at a lower scale than ±2 hours. This would make it possible to detect incorrect daylight-saving time transitions on a daily operation. It would be interesting to analyse whether certain events, such as a sudden drop in pressure or fluctuation in flow, where the exact time is known, can be used to verify the occurrence of drifts identified by test VII. Furthermore, drifts could be verified by using the temporal and spatial redundancy between meters, e.g. by including a measure of similarity between similar time series. Since test VII is dependent on a regular recurring pattern in the data, it will never be applicable for all data series; the current implementation is therefore to be seen as a first draft that illustrates the usefulness of such a test, and we expect to be able to improve it.

E. Example of data validation and analysis for operational use

Section 4.1 of the main paper

Figure E-1 illustrates two examples of data validation and analysis for operational use. Figure E-1a provides an example of the 'flag status' based on all anomaly tests of 20 sensors in utility A over a period of four days. Flagged data was found from single (minute) data points for up to several days of consecutive flagged data (e.g. pressure meter P7). For example, short periods of invalidated data are visible at a higher rate in the flow meters Q1 and Q2. Q6, Q9 and P7 indicate 'dirty data', of which Q9 and P7 likely have no data available at all.



(a) Daily data verification (utility A). **(b) Flatline anomaly test analysis (utility B).**

Figure E-1. Example of anomaly visualization for operational use. (a) Raw meter data validation from ten flow (Q) and pressure (P) meters in utility A in July 2015. (b) Mean flag rate based on all raw data points for the flatline test in pressure meters of utility B. Whiskers display the total flag rate based on all anomaly tests in the individual pressure meter.

It can be beneficial to focus on the individual error rates of the meters (Figure E-1b). For example, identifying differences in the individual error rates can be used to highlight meters not being sensitive enough to detect certain flow regimes at the installed location. In our case, the pressure meters P17–P19 tended to have a significant higher rate of flatlines than the remaining meters, which calls for inspection of these meters.

F. Extended Jaccard coefficient analysis

Section 4.3 of the main paper

Figure F-1 illustrates the Jaccard coefficients based on flags determined by tests I–VI for all utilities. We merged the anomalies of tests I and II into one group owing to the low number of flags in both categories. In the case that a meter contains flags in a certain anomaly process, the Jaccard coefficient by itself results in a high similarity of 1. If a meter data set contains no anomalies, no Jaccard coefficient, i.e. a value of zero, is computed. This results in the almost full red straight lines such as those seen in the range test of utilities A and B. In case of utility B, only a certain similarity pattern is visible for the flatline test of flow meters. This pattern is likely to be due to the high rate of flatline segments already shown in Figure 3 in the main paper. Also, Figure 3 shows that the number of missing data points (not based on time) was relatively low for utility B; however, the Jaccard coefficient of around 0.5 in Figure F-1, covering almost the entire timestamp inconsistency plot, indicates that large parts of the system lack data or went offline at the same time. High rate of change test and flatline test similarities are seen between six pressure meters (ID > 90, Figure F-1). All six pressure meters are positioned in an emergency pumping station. Thus, under normal circumstances all meters should log more or less constant values and change drastically if pumping starts. This might be reflected in terms of flags captured by the two tests. According to utility C, the meters are connected to the same programmable logical controller before being sent to the raw database, potentially being the source of the anomaly.

Figure F-1. The Jaccard coefficient computed for six different tests from the anomaly testing, based on the available flow and pressure meters. A Jaccard coefficient between zero and one describes no and very high similarity respectively for the occurrence of flags in the applied tests. Test types: I – duplicate timestamp, II – illegitimate timestamp, III – range, IV – rate of change, V – flatline and VI – timestamp inconsistency.

G. References

DANVA (Danish Water and Wastewater Association). 2016. Water in Figures 2016. Skanderborg: DANVA. https://www.danva.dk/publikationer/benchmarking-og-statistik/water-in-figures-pdf/water-in-figures-2016/