Water Supply

IWA PUBLISHING

Water Supply Vol 22 No 7, 6370 doi: 10.2166/ws.2022.251

Direct assessment of background leakage levels for individual district metered areas (DMAs) using correspondence of demand characteristics between DMAs

Peter van Thienen

KWR Water Research Institute, P.O. Box 1072, 3430 BB Nieuwegein, The Netherlands
E-mail: peter.van.thienen@kwrwater.nl

PVT, 0000-0001-5528-845X

ABSTRACT

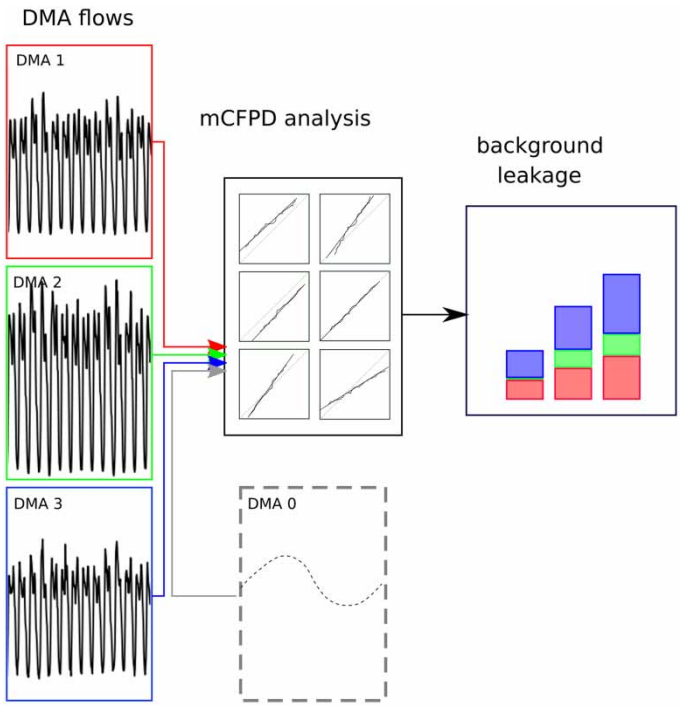
This paper proposes a new approach to rank a group of district metered areas (DMAs) in terms of background and unreported leakage rate and to quantify background/unreported leakage levels for individual DMAs in this group. This is done using an extension of the comparison of flow pattern distributions or CFPD method. The approach presented is based on an assumption of similarity in demand behavior between different DMAs. It requires no other data than net inflow timeseries for the DMAs or supply areas under consideration, and no assumptions other than that of similarity of demand. As such, it provides a low-data-requirements method for the evaluation of background and unreported leakage that does not share underlying assumptions with the commonly used mininum night flow method and may potentially present a supplement or alternative to it. The approach is validated using numerical simulations and applied to flow data of a set of DMAs from the Netherlands.

Key words: background leakage, DMAs, pattern similarity

HIGHLIGHTS

- A new method to rank DMAs in terms of background/unreported leakage
- Generates estimates for (ranges of) absolute background/unreported leakage levels
- Low data requirements
- A single central assumption that is verifiable

GRAPHICAL ABSTRACT



INTRODUCTION

Leakage continues to be a major challenge for water companies in many parts of the world (global average 30% with large regional variations, Liemberger & Wyatt 2019). With continuing population growth and urbanization, and a changing climate that is currently already affecting the demand for and the availability of water, but is likely to do so to a much larger degree in the coming decades and beyond, it is more important than ever to minimize losses of water through leakage.

Lambert (1994) distinguishes three types of leakage: background leakage (e.g. at joints, with low flow rates that are hard to detect), unreported leakage (potentially significant flows, but nobody noticed, and the water disappears into the soil or a waterway), and reported leakage. This paper focuses on the former two categories, that is to say pre-existing leakage that is difficult to detect and quantify from flow signals. This distinguishes the current approach from burst detection, which focuses on the third type of leakage. In this paper, the term background leakage is intented to encompass both unreported and true background leakage as defined by Lambert (1994).

The most commonly used methods to assess (background) leakage at the level of district metered areas (DMAs) or supply areas in practice are the minimum night flow method (Puust *et al.* 2010) and the water balance method (Farley & Trow 2003). The former hinges on an estimate for the minimum demand in a DMA during the quiet hours of the night or an assumption of zero consumption for sufficiently small DMAs. The latter requires, among other things, registration or estimates of actual demand. A third approach, the burst and background estimates method (BABE, Lambert 1994; Lambert & Morrison 1996), aims to evaluate individual components of leakage using a number of characteristics of the network and its operation. Finally, a hydraulic-model-based approach can be applied (e.g. Giustolisi *et al.* 2008; Marzola *et al.* 2022; Romero *et al.* 2022; Steffelbauer *et al.* 2022; Wu *et al.* 2022), though this does not appear to be common practice among water utilities for estimating leakage levels, and obviously this requires the availabilty of a hydraulic model of adequate quality.

This paper proposes a new approach to rank a group of DMAs in terms of background leakage rate and to quantify background leakage levels for individual DMAs in this group. This is done using an extension of the comparison of flow pattern distributions or CFPD method (Van Thienen 2012; Van Thienen & Vertommen 2016). The approach presented is based on an assumption of similarity in demand behavior between different DMAs. It requires no other data than net inflow for the DMAs or supply areas under consideration, and no assumptions other than that of similarity of demand. As such, it provides

a low-data-requirements method for the evaluation of background and unreported leakage that does not share underlying assumptions with the commonly used mininum night flow method and may potentially present a supplement or alternative to it.

The paper is structured as follows. After a brief introduction of the original CFPD method, its extension to background leakage identification is described. The method is then illustrated and tested on synthetic DMA flow data, and a sensitivity analysis is presented. Next, the validity of the CFPD approach for non-uniform and non-constant leakage behavior is determined using hydraulic models. The combined set of methods is subsequently applied to actual DMA flow data from a set of supply areas in the Netherlands. Finally, the results are discussed and conclusions are presented.

METHODS

A short introduction to the CFPD method

In this section, we give a summary of the method presented by Van Thienen (2012). Consider a supply area for which the flow rate into the area (accounting for all inflow, outflow and storage) is observed for a period of time (e.g., a day, a week, a month or an entire year) and again for a comparable period of the same length in another year. The observed patterns are likely to be similar in shape but not exactly the same. The simple CFPD procedure allows a quantitative comparison of these patterns, taking the following steps (see Figure 1):

1. Sort both data sets by magnitude from small to large. Sorted measurement *ranks*, scaled to a 0–1 range, are on the horizontal axis, while flow rates are shown on the vertical axis.
2. Plot one observed data set against the other on a CFPD plot.
3. Determine a linear best fit with slope a and intercept b.

The slope a is a scaling factor of which the physical interpretation is a change in demand consistent with the pre-existing demand pattern. Hence this change is called a *consistent* change. The intercept b is a shift in the total pattern, which is a change in demand that is not consistent with the pre-existing demand pattern. Hence this change is called an *inconsistent* change. This can be physically interpreted in terms of a change in the (constant) leakage rate. A more elaborate discussion on the interpretation of these factors is presented by Van Thienen (2012).

Note that the word *pattern* is used here in the sense of a time series that is generally repetitive to a significant degree with some variations. In general, it is preferable to construct the CFPD plot with the first period on the horizontal axis and the second on the vertical. In this case $a>1$ and/or $b>0$ corresponds to an increase in flow rate. Note that comparison of periods

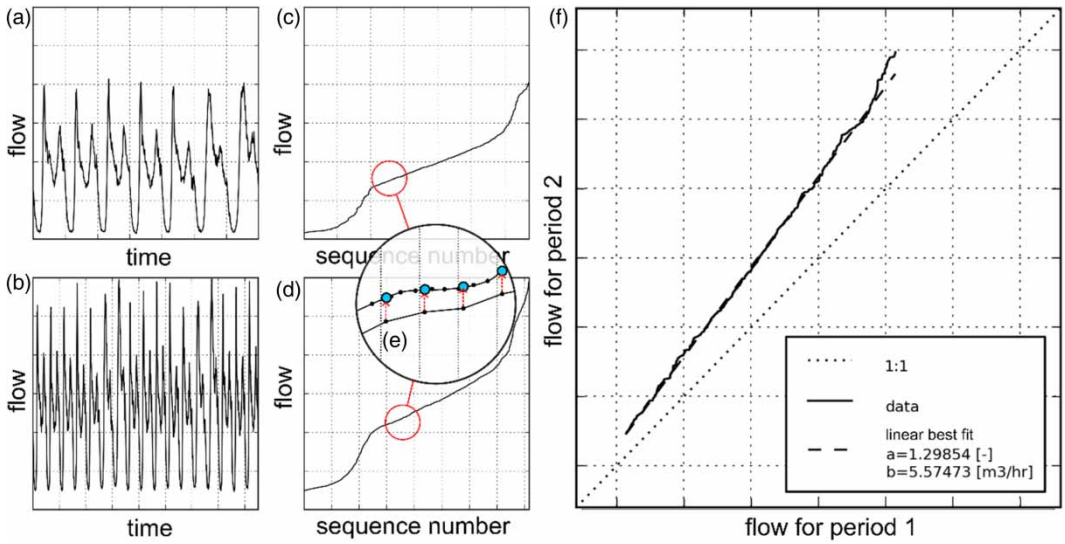


Figure 1 | Visualization of the CFPD procedure. The time component is removed from two flow time series (a, b) by sorting the data from small to large magnitude (c, d). If the time series are of unequal length, resampling of one of them is performed to make them of equal length while preserving the characteristic value distribution (e). The resulting data are plotted against each other (f) and parameters a and b are determined from the linear best fit. Note that flow and time may have any appropriate unit for these quantities.

of different lengths or sampling frequencies is also possible but requires an additional interpolation step, as depicted in Figure 1(e).

Multiple DMA CFPD comparison

Let us assume that we have flow data for three DMAs (1, 2, and 3) and that a CFPD comparison shows that demand behavior in these DMAs is comparable, that is to say, that a good linear fit is possible in the CFPD curve of each pair in this set (such as shown in Figure 1(f), however in this case comparing *DMAs* rather than periods for the same DMA). Each of these DMAs will have a somewhat different population size and background leakage level (factors a are different from 1 and factors b are different from 0). We construct CFPD curves for each pair in this set of DMAs 1, 2, and 3 for a selected time window (ideally the same for each of the DMAs to avoid seasonal signals in the data). These CFPD curves can then be described by the following expressions:

$$F_2 = a_{21}F_1 + b_{21} \tag{1a}$$

$$F_3 = a_{13}F_1 + b_{13} \tag{1b}$$

$$F_3 = a_{23}F_2 + b_{23} \tag{1c}$$

with F_1, F_2, and F_3 the sorted flow data for DMAs 1, 2, and 3, respectively, and a_{ij} and b_{ij} the CFPD coefficients for their comparisons. Note that a_{ij} and b_{ij} refer to the differences in pattern j compared to pattern i.

From (1b) and (1c), we get the following relation

$$F_2 = \frac{a_{13}}{a_{23}}F_1 + \frac{b_{13} - b_{23}}{a_{23}} \tag{2}$$

which, combined with (1a) gives us

$$a_{21} = \frac{a_{13}}{a_{23}} \tag{3a}$$

$$b_{21} = \frac{b_{13} - b_{23}}{a_{23}} \tag{3b}$$

This can be generalized to

$$a_{ij} = \frac{a_{ik}}{a_{jk}} \tag{4a}$$

$$b_{ij} = \frac{b_{ik} - b_{jk}}{a_{jk}} \tag{4b}$$

for any three-way comparison of DMAs, in which indices i, j, and k refer to three different DMAs. These expressions allow us to describe the relations between any number of DMAs, provided that their demand behavior is comparable, making it possible to perform a comparison of any number of DMAs in a single analysis (note that clustering within a group of DMAs, i.e. performing analyses on subsets rather than the complete set, will be investigated in the results section of this paper).

However, this does not yet give us absolute background leakage levels for individual DMAs. This requires an additional step. If we postulate a hypothetical DMA 0 that has no background leakage, the parameters of interest, i.e. the absolute background leakage levels b_{0j} for DMAs j are introduced.

We rewrite (4b) as

$$b_{ij}a_{jk} = b_{ik} - b_{jk} \tag{5}$$

For a set of n comparable DMAs and postulated background leakage free DMA 0, for all combinations of

$$i \in [0, n] \tag{6a}$$

$$j \in [0, n] \tag{6b}$$

$$k \in [0, n] \tag{6c}$$

we get a system of equations, that can be written as:

$$(C \cdot \vec{v}^T) \; o \; (D \cdot \vec{v}^T) = E \cdot \; \vec{v}^T - F \cdot \vec{v}^T \tag{7}$$

with C, D, E and F coefficient matrices, $\vec{v}$ a vector containing all coefficients a_{ij} and b_{ij}, and $\circ$ the Hadamard product. All matrices have shape $p \times q$, *with* $p=(n+1)^3$ the number of expressions, and $q=2(n+1)^2$ the number of CFPD coefficients a_{ij} and b_{ij}. Note that all expressions for $i=j$ are redundant. We leave them in the system to more clearly illustrate the structure of the matrices. We construct vector $\vec{v}$ in the following manner:

$$\vec{v}^T = (a_{00}, a_{01}, \; \ldots, \; a_{10}, a_{11}, \; \ldots, \; a_{20}, a_{21}, \; \ldots, \; b_{00}, b_{01}, \; \ldots, \; b_{10}, b_{11}, \; \ldots, \; b_{20}, b_{21}, \; \ldots) \tag{8}$$

This results in a clear matrix structure for C, D, E, and F, see Appendix I in the supplementary material online.

The CFPD coefficient vector $\vec{v}$ contains both parameters that can be easily determined from the DMA datasets (a_{ij} and b_{ij} for $i \neq 0$, $j \neq 0$, $i \neq j$ and for $i=j$, in which case $a_{ij}=1$ and $b_{ij}=0$ by definition), and unknowns (a_{ij} and b_{ij} for $i=0$ or $j=0$, $i \neq j$). We are particularly interested in unknowns b_{0j}, which represent the absolute background leakage levels in the real DMAs. As an additional constraint, when available, we can use an estimate for the total background leakage L of the area which comprises the set of DMAs:

$$L = \sum_{j=1}^{n} b_{0j} \tag{9}$$

As for each DMA, the net flow (i.e. inflows minus outflows) is considered, this also works for cascading DMAs.

We solve for these unknowns by applying a numerical optimization scheme (implemented in SciPy, Virtanen *et al.* 2020), minimizing the residual between model and data. We impose the boundary condition that all $b_{0j} \geq 0$, i.e. no negative background leakage rates. Note that the parameters a_{0j} have no physical meaning, since there is no natural pattern scale for the postulated DMA 0. The parameters a_{0j} are constrained by requiring that $a_{01}=1$. This sets the scale for the postulated undisturbed pattern, and thus fixes the other a_{0j} parameters as well.

Testing and sensitivity analysis

There are several sources of error in this analysis that can have a significant impact on its results. The effects of DMA patterns not showing the same behavior (in the sense that CFPD analyses result in a linear relation) are investigated by adding noise to the flow signals. Also, the effects of choosing a wrong value for the total background leakage level L are determined.

Constant flow rate leakages

First, we look at leaks that have a more or less constant flow rate. This situation is representative of networks that have small pressure variations throughout the day due to implemented pressure management. It has been shown for simple test cases (Van Thienen 2012) that introducing leakage pressure dependence does not invalidate the CFPD approach. This will be further validated for a range of relevant conditions using hydraulic models below.

We start by generating a large number of flow patterns from a single, unperturbed flow signal that is postulated to be leak free (Figure 2). In this case, we have chosen a pattern with 0 minimum night flow, but this is not necessarily the case. These patterns are then transformed using the CFPD paradigm, i.e., first scaled and then shifted, the latter operation representing the addition of a leak that has a constant flow rate. A large number of different scaling factors a (range 1–10) and offset factors b (range 0–100) are chosen randomly. Normally distributed noise is added. Clusters of DMAs are analyzed using the procedure described above, as a function of noise level and error in the estimated DMA cluster background leakage. The complete procedure is described in the pseudocode in Figure 3.

The results of this analysis are shown in Figure 4 through Figure 7, for DMA clusters of 2, 3, 5 and 8 DMAs, respectively. The mean error in the recovered background leakage for the DMA cluster and its standard deviation are plotted in Figures 4–7 subfigures a and b. These are presented as a percentage of the actual DMA cluster background leakage rate. The mean error in the recovered background leakage for individual DMAs and its standard deviation are plotted in Figures 4–7 subfigures c

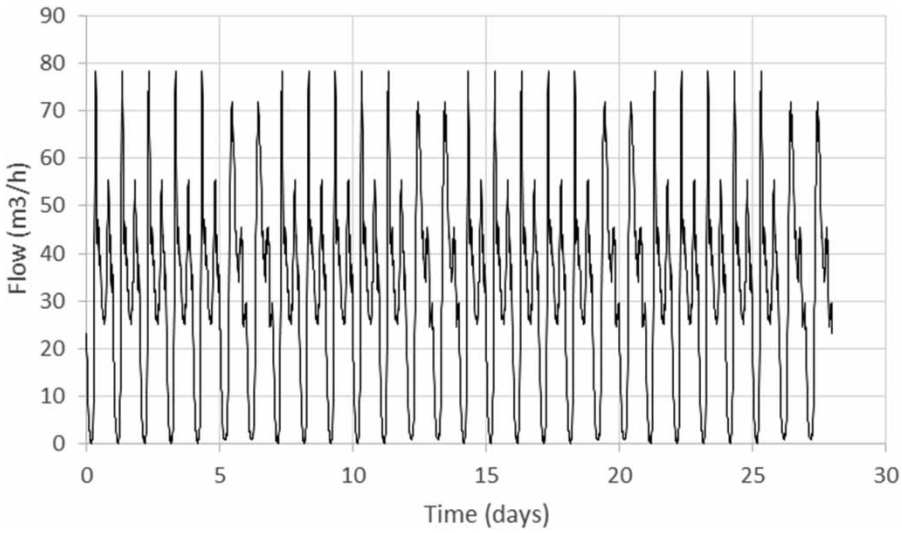


Figure 2 | Unperturbed base flow pattern.

and d. These are presented as a percentage of the mean DMA flow, since presenting them in the same manner as previously described would result in extreme magnification of errors in low background leakage DMAs.

The resulting images are all quite similar for the varying number of DMAs. Any error that is included in the estimate for the total background leakage is directly reflected in the sum of the recovered background leakage levels (subfigures a, c). This means that the multiple DMA CFPD (mCFPD) method is able to reconstruct the actual background leakages in all DMAs provided that the estimated total background leakage is accurate. If this is not the case, i.e. if the summed non-revenue water (NRW) error is nonzero, this error propagates directly into the recovered values for the background leakage levels of the DMAs. The robustness of this recovery declines somewhat for greater errors in the total NRW estimate, in both directions, but remains relatively small, showing a standard error which is mostly below 10% of the actual value for the total recovered background leakage levels for a cluster (subfigures b) and mostly within 5% of the mean flow for the individual DMA background leakage levels (subfigures d). None of the tests shows any significant sensitivity to the noise that was added to the signals. It must be noted, however, that using other probability density functions for the noise signal, in particular skewed ones (e.g., resulting from lower sensitivity and larger error at low flow rates), is likely to result in a degradation of the performance for increasing noise levels.

Solvability and uncertainty

Occasionally, the set of flow signals is such that the optimizer cannot find a good solution, even for synthetic examples for which a true solution is available and known. The main expression of this behavior is in large SSR (sum of squares of

```
for number_of_DMAs in [2,3,5,8]:
        for noise_level in [0.0, 0.01, 0.015, 0.025, 0.05, 0.1, 0.25]:
                for summed_background_leakage_error in [-0.5, -0.25, -0.1, 0.0, 0.1, 0.25, 0.5]:
                        for sample in [1:50]:
                                for DMA in number_of_DMAs:
                                        generate DMA flow time series
                                perform mCFPD analysis to recover background leakage level for each DMA
                                compare to true values
                        generate statistics for the batch of samples
```

Figure 3 | Pseudocode of sensitivity analysis for constant flow rate leaks. The parameter number_of_DMAs refers to the DMA cluster size.

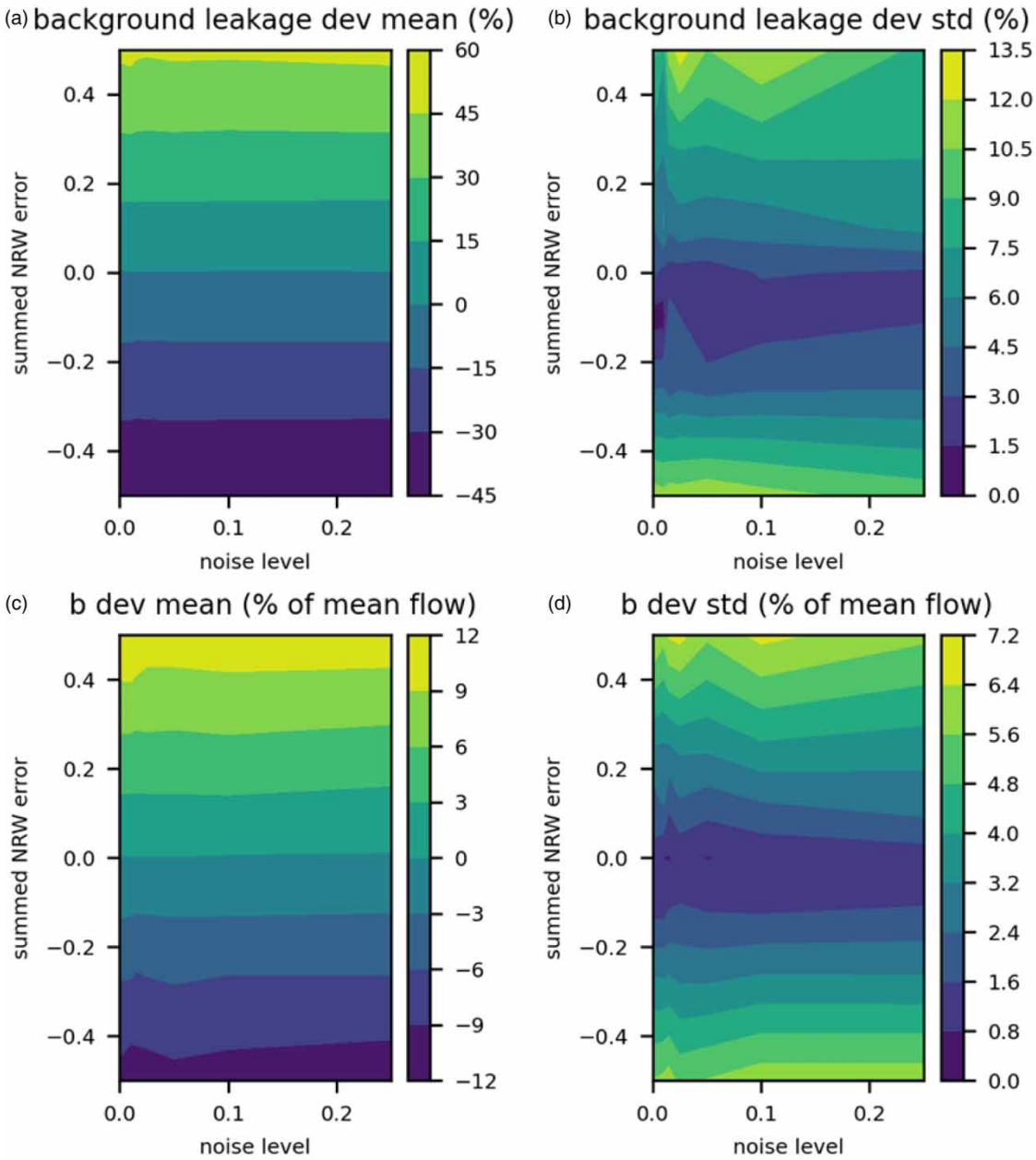


Figure 4 | Sensitivity analysis for constant flow rate leaks, for two DMAs. (a) mean error in recovered background leakage level for DMA clusters, relative to the actual background leakage level; (b) standard deviation of the error in recovered background leakage level for DMA clusters; (c) mean error in recovered background leakage level for individual DMAs, relative to the mean DMA flow; (d) standard deviation of the error in recovered background leakage level for individual DMAs. All subfigures plot results as a function of the error in the estimated total background leakage level L, see expression (9), and the imposed noise level in the flow data. Note that summed non-revenue water (NRW) error and noise levels are dimensionless fractions.

residuals) values, which should therefore be monitored and considered in relation to plausible background leakage level numbers. This behavior is further illustrated in Appendix II in the supplementary material.

As can be seen in the results of the previous section, both the estimated background leakage levels for individual DMAs and consequently the sum of these estimates depend strongly on the estimate of the background leakage level L that is fed into the problem as prior information (Equation (9)). Handling and propagation of this uncertainty is discussed in Appendix III in the supplementary material.

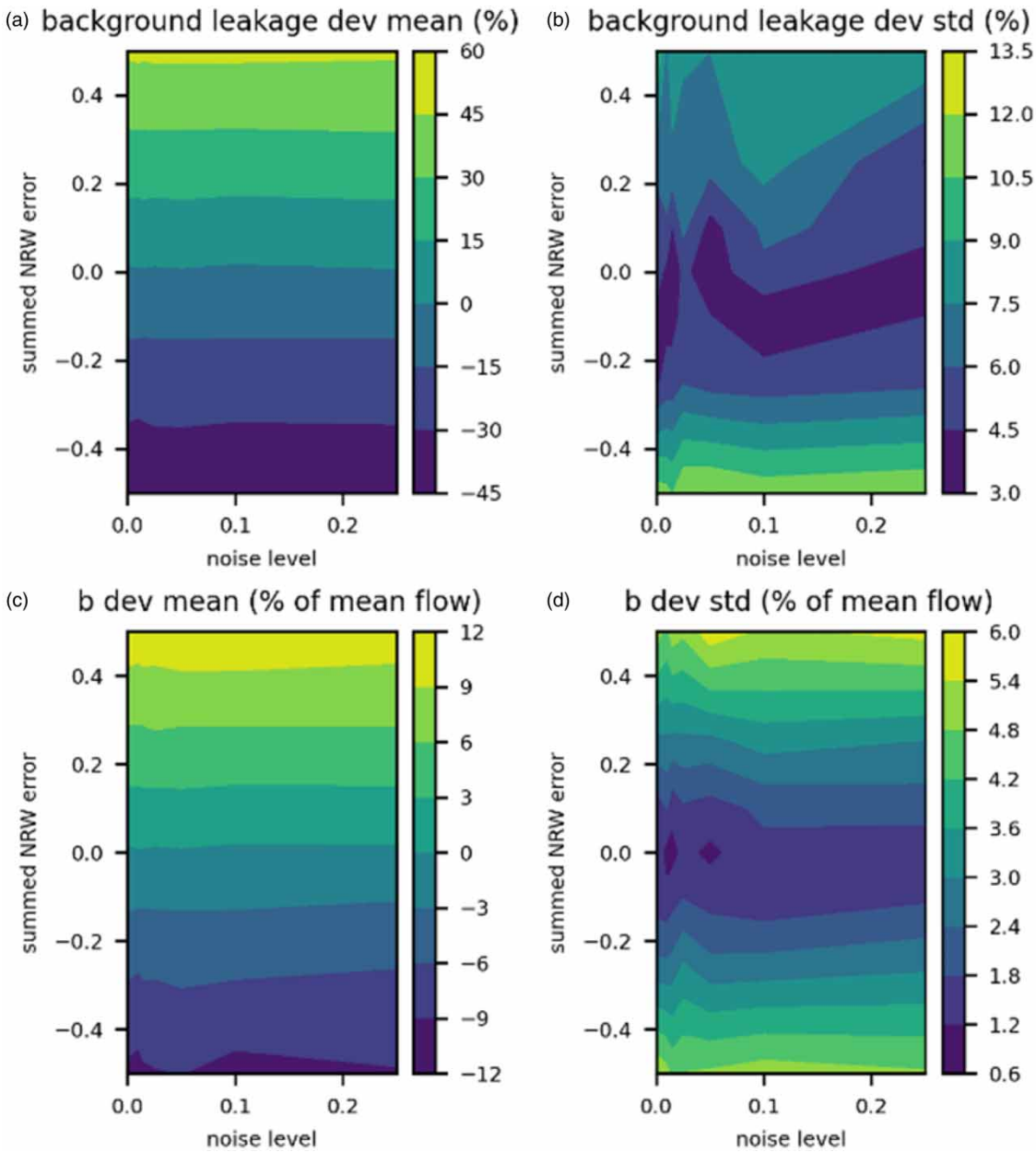


Figure 5 | Sensitivity analysis for constant flow rate leaks, for three DMAs. For an explanation of the subfigures, see Figure 4.

Verification of the applicability of the CFPD model for complex leakage behavior

Before applying the mCFPD method described above to real-world data, first a number of background leakage scenarios are investigated in a set of hydraulic models. The purpose of this is to investigate to what degree the CFPD model holds when a more complex leakage model is applied, that is to say, whether we can apply and interpret the mCFPD approach that is built on the CFPD paradigm.

A set of network models for real-world systems with different characteristics was obtained, see Table 1. The first is a modification of an existing network, in which a number of looped structures have been opened up to create a more branched topology. This had originally been proposed to improve the network self-cleaning. For the present application, this variant of the network has been selected because it is expected to show larger pressure variations and thus variations in background leakage through the day. The model represents part of a Dutch town, with about 1,000 connections.

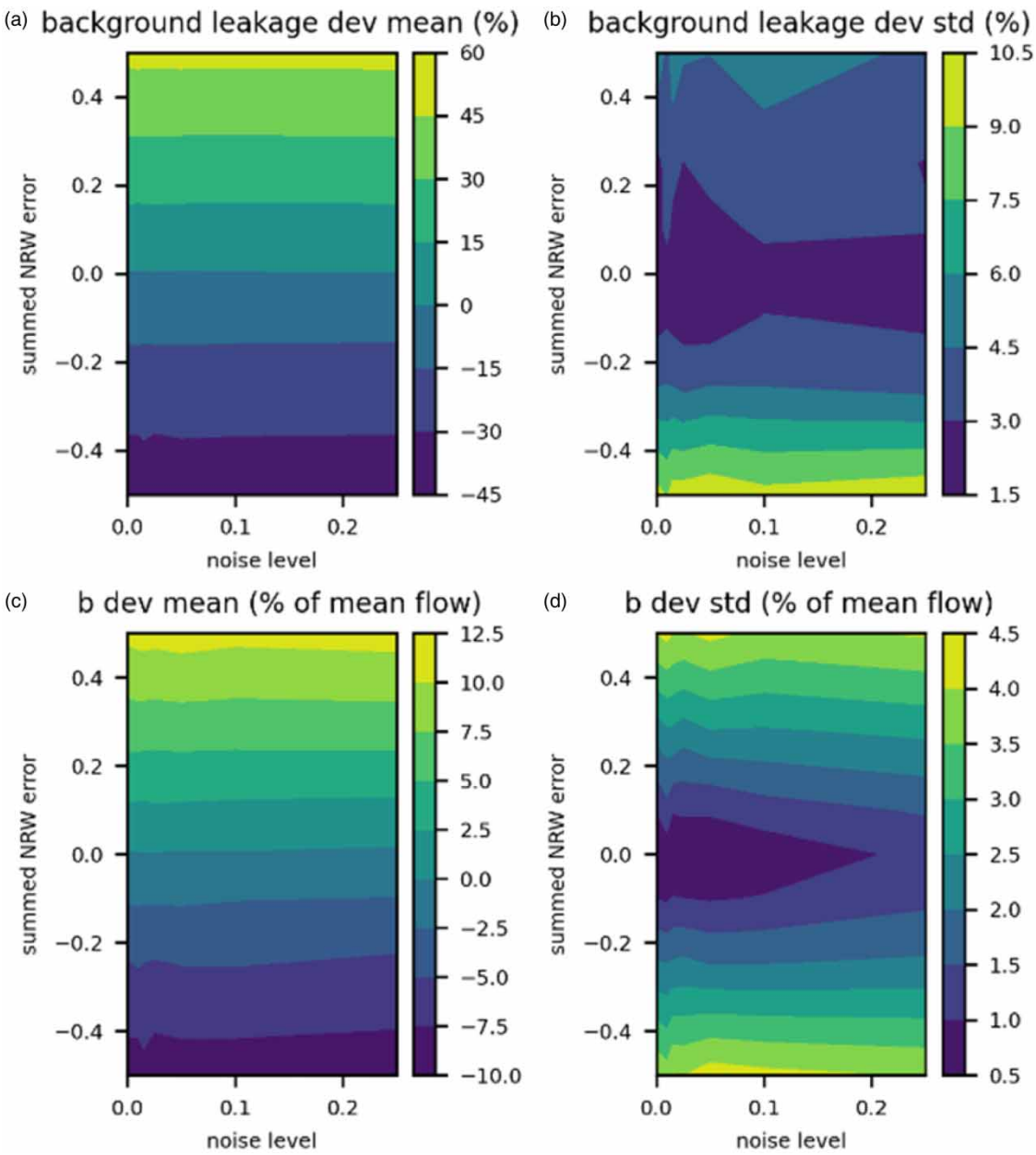


Figure 6 | Sensitivity analysis for constant flow rate leaks, for five DMAs. For an explanation of the subfigures, see Figure 4.

The second is the network of a city of approximately 77,000 inhabitants. These models have a relatively flat topography and little pressure variations in the network throughout the day.

In order to investigate the effects of more significant spatial and temporal pressure variations, for each of these models, a series of alternatives with synthetic topography (sloping west-east with varying gradients) and synthetic head profiles at the source (cosine pattern with varying extreme values at midnight and midday) were generated. In each of these modified models, a synthetic background leakage scenario was generated by randomly adding emitters whose characteristics were selected randomly within a specified range to a specified fraction of 10% of the junctions in the network model. Hydraulic simulations were performed using EPAnet 2.2 (Rossman *et al.* 2020). In each case, a CFPD analysis was performed comparing hydraulic simulation with and without added leakage (but using the same topography and head function scenario).

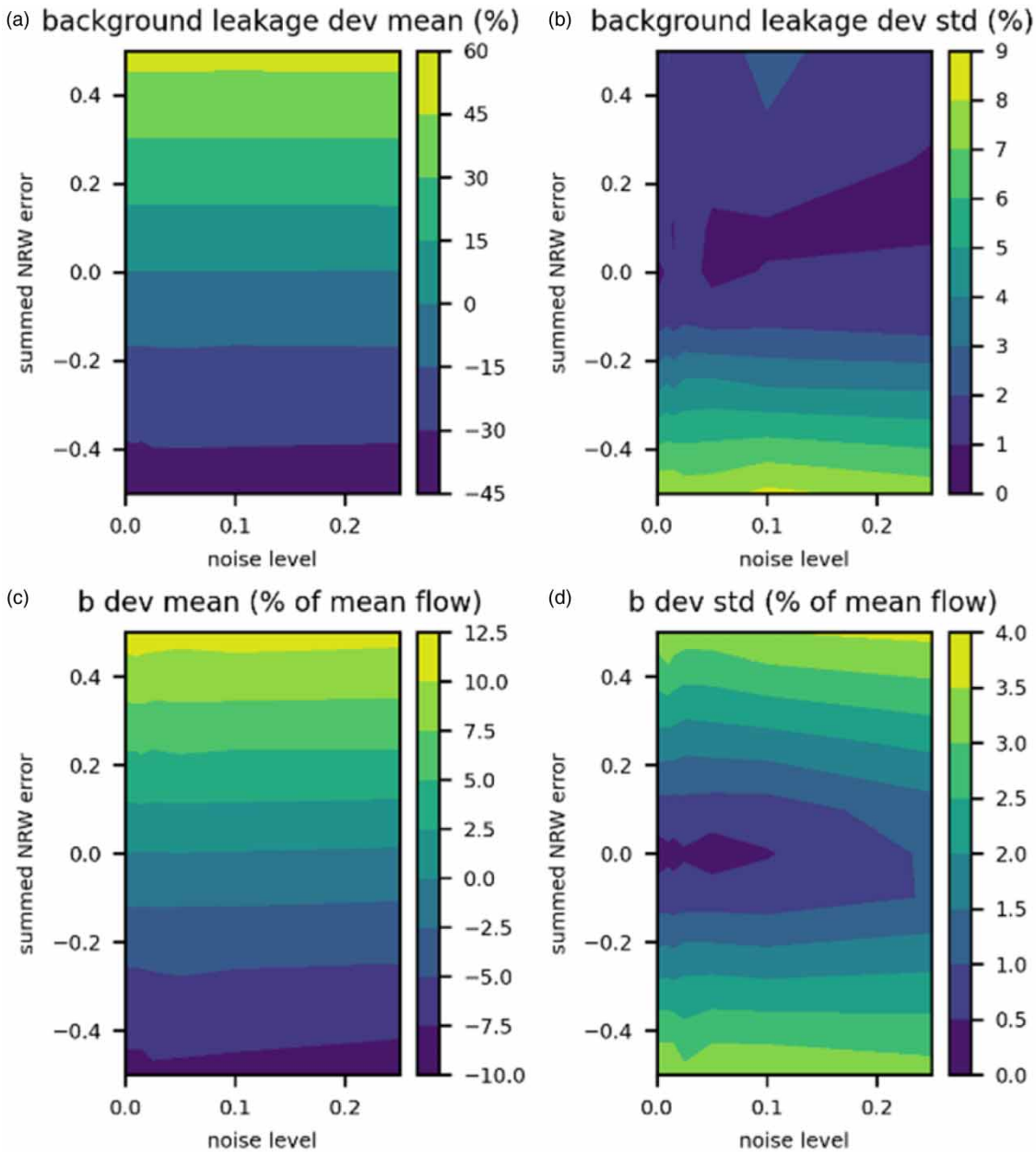


Figure 7 | Sensitivity analysis for constant flow rate leaks, for eight DMAs. For an explanation of the subfigures, see Figure 4.

The results of this exercise are shown in Figure 8 for Network Model 1 and in Figure 9 for Network Model 2. These include the standard error in the determination of the intercept factor b, which is formulated as follows (Seltman 2018):

$$se_b = s\sqrt{\frac{\sum x^2}{n\sum (x^2) - (\sum x)^2}} \tag{10}$$

with

$$s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n-2}} \tag{11}$$

Table 1 | Network model characteristics

	Network Model 1 (Sittard branched)	Network Model 2 (Roosendaal)
# junctions	497	15489
# pipes	474	13355
# feeds	1	1
layout		
elevation parameter space	80 m (W)–30 m (E) 30 m (W)–80 m (E)	0 m (W)–30 m (E) 30 m (W)–0 m (E)
feeder head parameter space	cosine extreme 0 h: 100 m–150 m cosine extreme 12 h: 150 m–100 m	cosine extreme 0 h: 40 m–60 m cosine extreme 12 h: 60 m–40 m

Both figures show a similar situation as a function of the elevation profile and head function amplitudes. Both mean background leakage levels over a 24 hour period (subfigures a) and CFPD reconstructed mean background leakage levels (subfigures c) indicate a noisy signal. This is because the background leakage levels are randomly generated for each instance of the hydraulic model with a different combination of elevation profile and head function. However, the important observation is that they represent more or less the same noisy pattern. This shows that even in a case where there is significant variation in the background leakage level through the day (represented by a higher standard error in subfigure b), the CFPD recovery of the mean background leakage (*b* factor in subfigure c) is still good. The observed errors (subfigure d) are well within +/−20% for the cases studied here, and show some asymmetry with respect to the elevation profile. This is presumably related to asymmetry in the network layout, i.e., high topography related pressures on the extensive east side of the network will have a higher proportion of randomly distributed leakages than the more sparse western side. The larger number of leakages in a high pressure part of the network will result in a disproportionate increase in leakage volume. Error parameters for the CFPD fit (subfigures e–g) visually correlate well with the variability of the daily changes in the background leakage level (subfigure b).

From these diagrams, we can conclude that for the hydraulic models studied and within the bounds of the prescribed parameter range, the CFPD approach catches compounded leakage signals in complex networks to well within +/−20% of their true values even when topography and source head variations through the day cause significant variations in the instantaneous background leakage level. This justifies the use of the CFPD methodology as a basis for the mCFPD approach.

Time windowing and clustering for multi DMA comparison

Application of the basic CFPD method only gives meaningful results if the input data are not disturbed by anomalies, such as occasional fillings of swimming pools, erratic behavior by large volume customers, and other exceptional events. Therefore, in order to obtain a successful multi DMA CFPD analysis, we need to find a time window in a long time series that avoids these

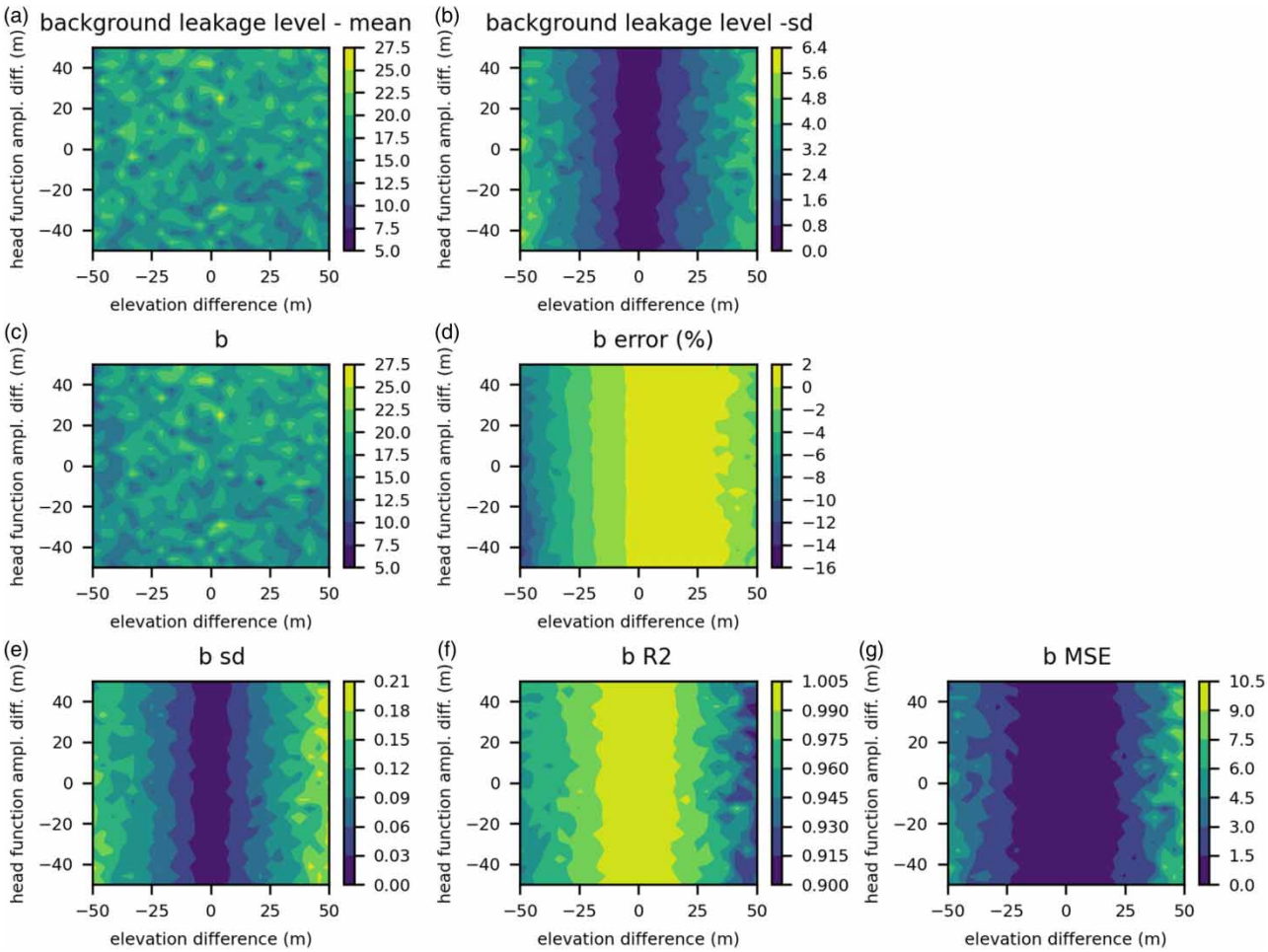


Figure 8 | Pressure variation sensitivity test results for Model 1. All sample nodes in the parameter space spanned by elevation slope (difference) and head function amplitudes (difference) represent randomly sampled background leakage levels. (a) mean background leakage level through a 24 hour period; (b) standard deviation of the background leakage level through a 24 hour period; (c) reconstructed mean background leakage level using CFPD; (d) difference between the true mean background leakage level and the recovered level; (e) standard error in the intercept value of the CFPD linear regression; (f) coefficient of determination for the CFPD fit; (g) mean square error for the CFPD fit. Note that the numbers along the vertical axis represent the difference between the midday and midnight extremes of the cosine function that has been prescribed as head pattern.

anomalies and gets the best CFPD fit (assumed to be the most reliable). This is done by evaluating all possible time windows of length l in the dataset, starting from the first window of length l in the timeseries and progressively moving the window one time step. For each of these windows CFPD analyses are performed between all DMAs. By assessing which time window results in the greatest mean R^2, that is for which the CFPD fits are the best, the most appropriate time window is identified.

Also, when comparing a larger set of DMAs, it is possible or even likely that not all will exhibit exactly the same demand behavior (meaning that not all CFPD curves will provide very good linear fits). Within a larger set of DMAs, smaller subsets of DMAs may be present that do exhibit very similar demand behavior. Therefore, all possible combinations of DMAs in the datasets in up to n subsets are considered individually as described above, and the combination of subsets that has the lowest mean of mean standard errors in the b coefficient is selected. This will be illustrated in the applications to real DMA data section below. Note that the determination of optimal clusters requires the use of the same time window for all (based on an analysis for the complete set of DMAs) if the analysis results at different clustering levels are to be compared. This will be illustrated in the field data case below.

Validation and application on real DMA data

In order to further validate the approach described in this paper (in addition to the validation on synthetic flow data described above), one would want to have a flow dataset for which the true background leakage levels are known. Estimates from

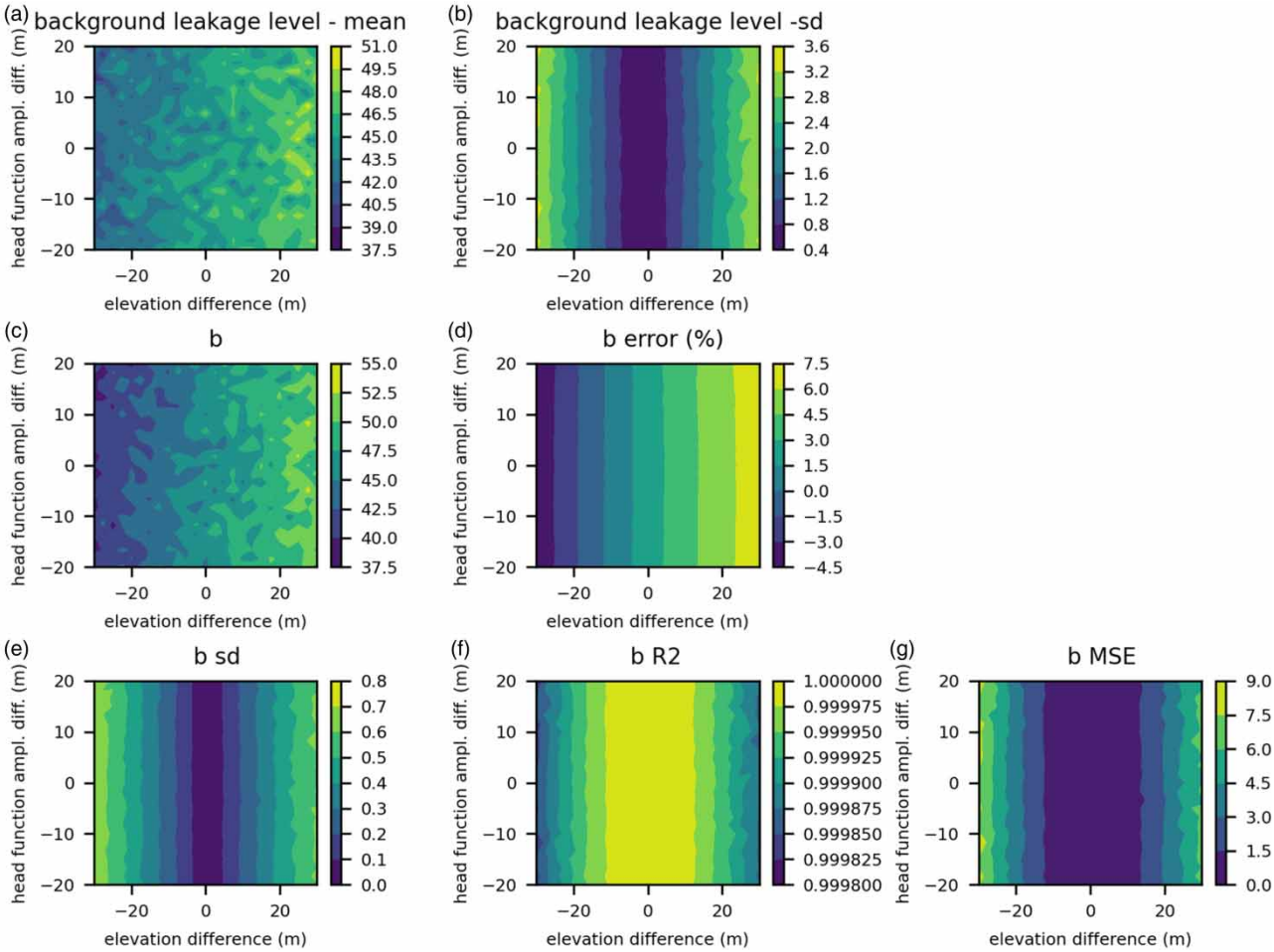


Figure 9 | Pressure variation sensitivity test results for Model 2. For an explanation of the subfigures, see Figure 8.

minimum night flow analyses may provide an indication, which may be used to verify that the results of the mCFPD analysis are plausible, but do not provide a ground truth. Application of the method in many different DMAs in addition to traditional methods may help to build confidence in the mCFPD approach. In the following paragraphs, we describe the application of the mCFPD method to a dataset from the Netherlands. These analyses are not presented as a validation, but as an illustration of the approach and associated challenges when dealing with real data.

Description of DMAs and dataset

Flow data into nine anonymized supply areas (which we designate WA…WI) were kindly provided for the year 2019 by the WML water company, which supplies the southeast of the Netherlands. Data from two DMAs consist of mutually dependent derived values that have for this reason been excluded from the analysis (see Table 2). The data have a time resolution of one measurement per hour.

Analysis and results

In this case, since we have seven DMAs available, the consideration of the degree to which the flow patterns in the different DMAs are comparable in the sense of CFPD is considered implicitly employing the clustering method that has been described above. The complete set of DMAs has been considered as a single cluster, two clusters or three clusters. A 14 day period which has the best correspondence in the CFPD framework was identified (November 19 – December 2, 2020), as described above. The results for the single cluster analysis are presented in Figure 10; those for the two cluster analysis in Figures 11 and 12; and those for the three cluster analysis in Figure 13–15. These figures show the allocation of the total background leakage estimate by the mCFPD analysis for the entire cluster to the different DMAs that comprise the cluster. Also, the SSR is shown

Table 2 | Characteristics of flow data supplied by WML

	WA	WB	WC	WF	WG	WH	WI
total supply (10^3 m^3, 2019)	306.8	359.0	215.5	244.5	266.2	576.8	143.2
mean flow (m^3/h)	341.3	399.3	239.7	272.0	296.1	641.6	159.3

for the complete range (from zero to the summed minimum night flow for the period under consideration). The point at which this curve flattens out or bounces back up in the lower left of each of these diagrams corresponds to the model in which one (or theoretically possibly more than one) of the DMAs in the cluster has a zero background leakage level. We can consider this model to be our estimate for the minimum background leakage levels for all the DMAs in the cluster. These numbers have been compiled in Table 3.

The overview provided in Table 3 shows us a number of findings. To start with, the approach applied here gives a minimum background leakage estimate for each subcluster of DMAs with the smallest value for each subcluster being 0. Any existing offset between subclusters will not be expressed here.

This means that results for different clustering within a set cannot be directly compared (Table 3a), but must be corrected.

To illustrate this effect, let us consider the following simplified example: DMAs W, X, Y and Z each have different background leakage levels, of which X's is smallest. The mCFPD analysis will give a lower bound estimate for X's background leakage of 0. If we subsequently consider two subclusters W,Y and X,Z separately, the lower bound estimate for X will again be 0, but also that of Y in the other subcluster if that DMA has a smaller background leakage level than W. Therefore, in order to compare the numbers of the two subclusters, the offset between the subclusters has to be determined from the full cluster.

The offset correction term can be calculated (Table 3b) as the difference of the mean background leakage level of all the DMAs in a cluster and that of the same DMAs in the full set. This results in mostly comparable background leakage levels (Table 3c) for the different clustering levels (similar numbers in each column of Table 3c). It follows from the approach that at a higher level of subdivision, the similarities between the DMA patterns within the cluster must be greater, and therefore the determination of the *b* factors in the CFPD analysis more reliable. We observe that the differences between the single-cluster, two-cluster and three-cluster analyses are relatively small, i.e., the analysis is internally consistent for the three subdivisions.

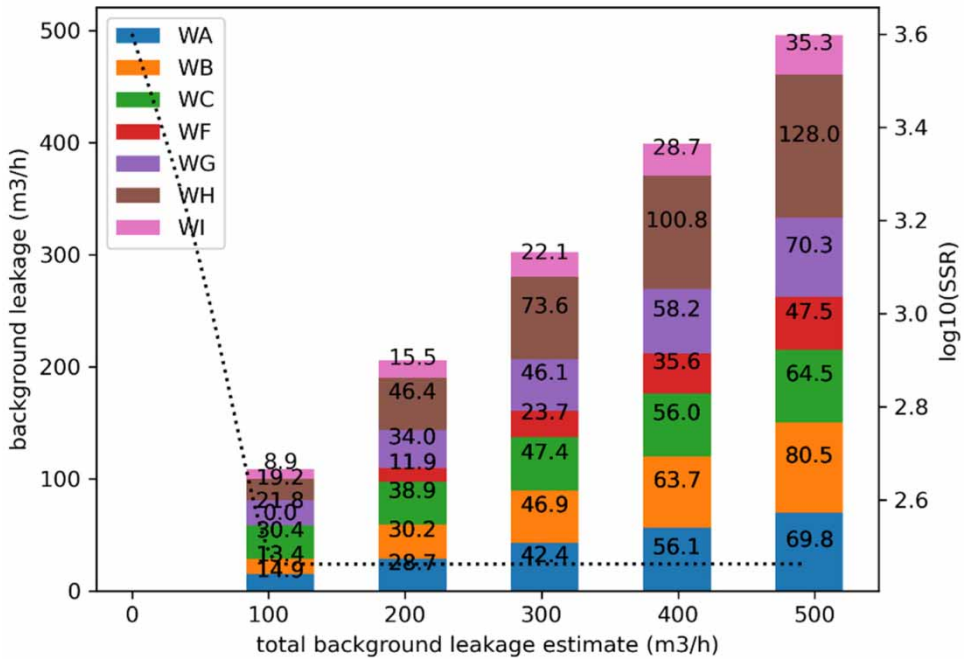


Figure 10 | mCFPD analysis results for the complete set of seven DMAs. The dotted curve indicates the SSR for the analysis.

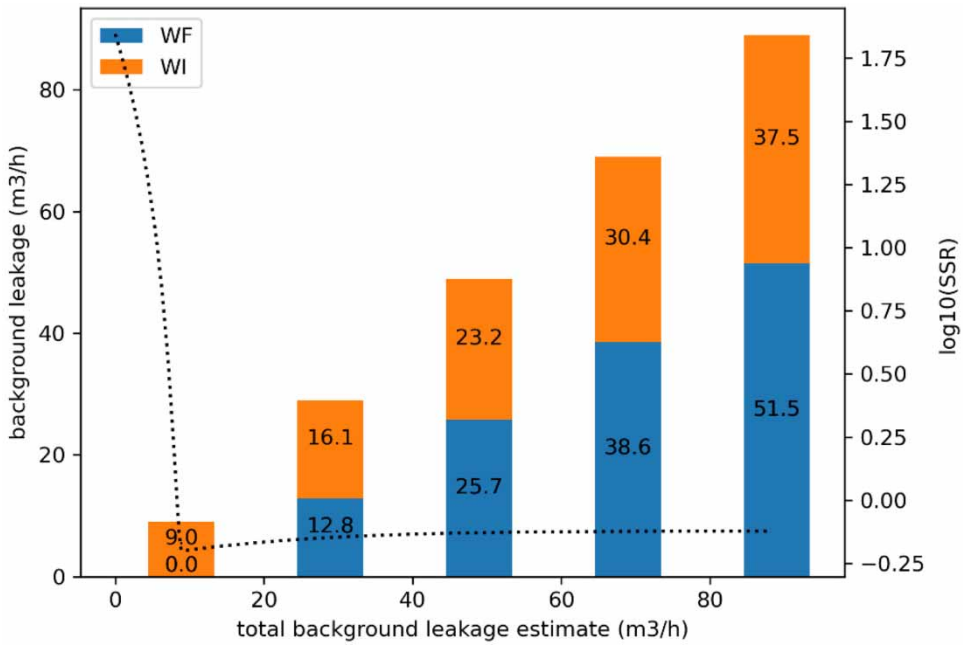


Figure 11 | mCFPD analysis results for the first cluster of a 2-cluster mCFPD analysis. The dotted curve indicates the SSR for the analysis.

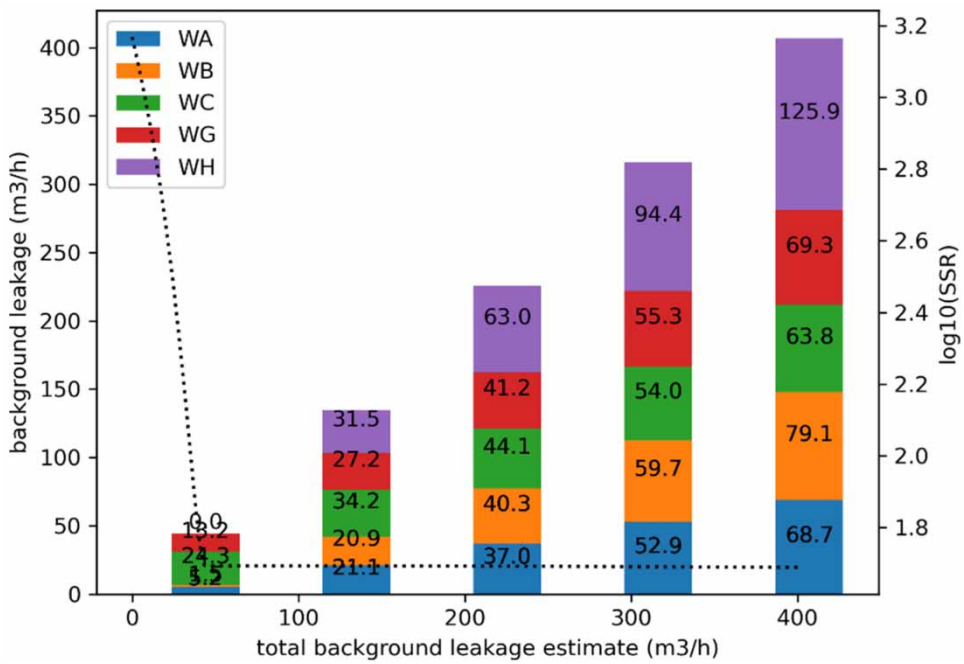


Figure 12 | mCFPD analysis results for the second cluster of a 2-cluster mCFPD analysis. The dotted curve indicates the SSR for the analysis.

DISCUSSION AND CONCLUSIONS

An alternative approach to estimating background (i.e., true background and unreported) leakage levels that relies solely on flow measurements at DMA inlets has been presented in this paper. Its main advantages with respect to other methods are the use of the complete 24 h signal (which in theory should provide a stronger statistical significance to the findings), its

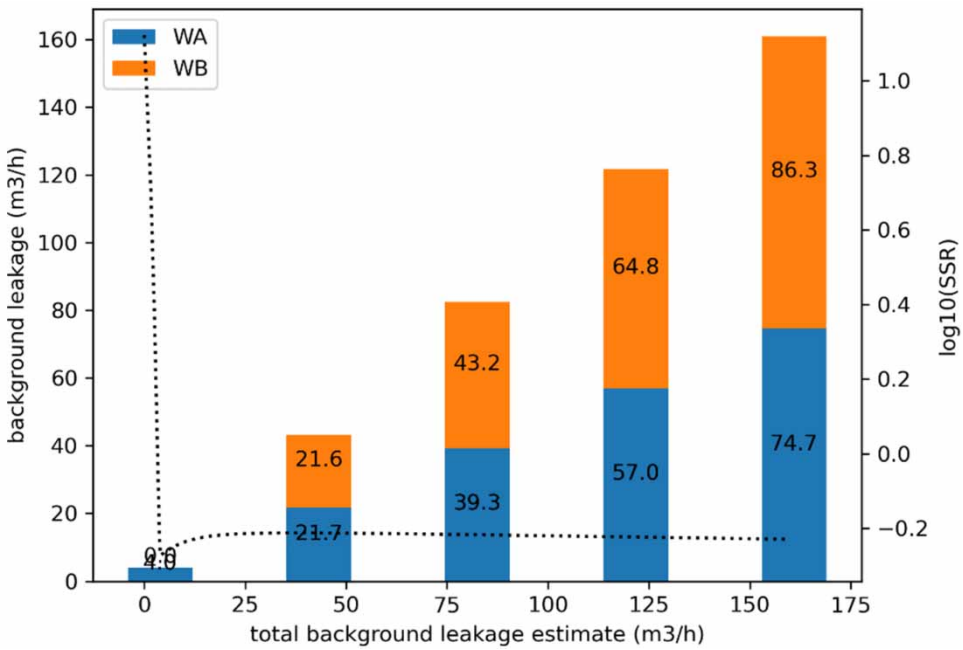


Figure 13 | mCFPD analysis results for the first cluster of a 3-cluster mCFPD analysis. The dotted curve indicates the SSR for the analysis.

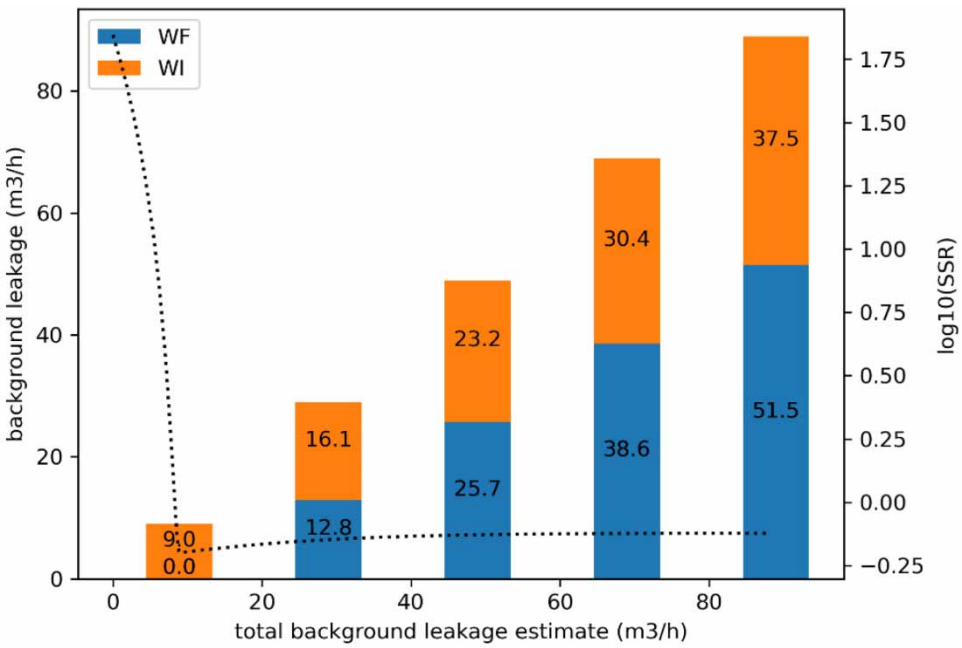


Figure 14 | mCFPD analysis results for the second cluster of a 3-cluster mCFPD analysis. The dotted curve indicates the SSR for the analysis.

insensitivity to unbiased noise, and the absence of the need for any estimations or assumptions except the central assumption of the method. This central assumption is that all DMAs within a set or subsets of DMAs have similar demand behavior. In practice it will be easy to find both situations in which this assumption holds very well (e.g. neighboring areas in a city with similar economic and demographic characteristics), as well as situations where it does not (e.g. comparing a mostly

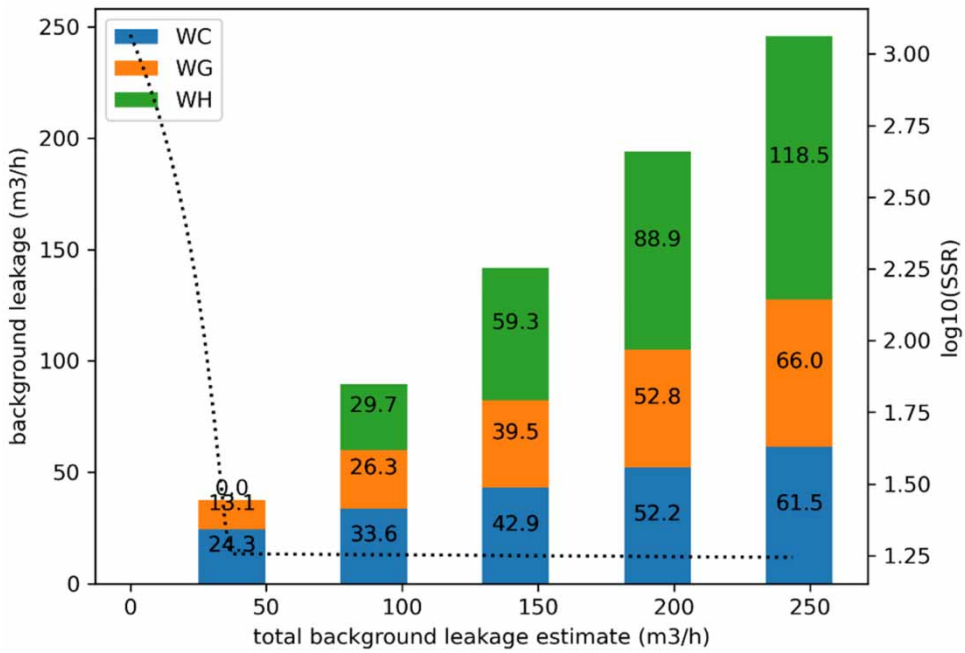


Figure 15 | mCFPD analysis results for the third cluster of a 3-cluster mCFPD analysis. The dotted curve indicates the SSR for the analysis.

Table 3 | Summary of the mCFPD analysis results for the WML dataset, presented in Figure 10 through Figure 15

# clusters	WA	WB	WF	WI	WC	WG	WH
	(a) Minimum background leakage levels (m^3/h)						
1	14.92	13.38	0.03	8.91	30.36	21.82	19.23
2	5.18	1.45	0.00	8.96	24.31	13.15	0.03
3	4.04	0.00	0.00	8.96	24.31	13.11	0.04
	(b) cluster offsets relative to the single cluster case (m^3/h)						
2	11.12		−0.01		11.12		
3	12.13		−0.01		11.32		
	(c) corrected background leakage levels (m^3/h)						
1	14.92	13.38	0.03	8.91	30.36	21.82	19.23
2	16.30	12.57	−0.01	8.95	35.43	24.27	11.15
3	16.17	12.13	−0.01	8.95	35.62	24.43	11.36
	(d) minimum background leakage levels as a percentage of the mean flow						
1	4.6%	3.4%	0.0%	5.6%	3.5%	7.3%	3.0%
2	5.0%	3.2%	0.0%	5.6%	5.8%	8.1%	1.7%
3	4.9%	3.1%	0.0%	5.6%	5.8%	8.2%	1.8%

(a) Minimum background leakage levels for the seven DMAs, based on analysis as a single, two or three clusters. (b) Offsets of the cluster base level with respect to the base level of the single cluster analysis. (c) Minimum background leakage levels including offsets. (d) Minimum background leakage levels as a percentage of the mean flow of the individual areas. Grey tones indicate clusters.

commercial area to a mostly residential area or comparing areas with non-neglegible constant or erratic industrial consumption) and therefore DMAs for which the mCFPD approach may be suitable and those for which it is clearly not. An approach to verify this correspondence and minimize its effect on the analysis has been presented as well.

The low data requirements (sensors on DMA inflows/outflows only, low time resolution) of the method and the absence of any assumptions other than the central one may make the method particularly useful for developing countries. That is, provided that they

do not have intermittent supply, which is still not uncommon in the developing world (Charalambous & Laspidou 2017), and which will strongly affect flow measurements and patterns (Totsuka *et al.* 2004; Criminisi *et al.* 2009; Weston *et al.* 2022). At this point, it is not clear which degree of intermittency of supply, if any, is acceptable, and to what degree secondary effects of intermittent water supply, such as the use of personal water storage tanks, affect the applicability of the mCFPD method. This requires further investigation.

Even though the approach presented here does not provide the information on where the background leakage is in a DMA, it does provide a basis for prioritizing DMAs for leakage localization and network rehabilitation efforts for water utilities.

The minimum night flow (MNF) method appears to be the most commonly used approach to determine background and unreported leakage levels, and allows its users to do the same things as described in the previous paragraph. MNF does hinge on assumptions with respect to night consumption, which may include a regional component (Amoatey *et al.* 2018). The mCFPD method is based on a completely different and independent assumption, as described above. The combined application of both MNF and mCFPD may strengthen the credibility of the outcomes of both analyses (in case they align) or give reason for a further investigation of the applicability of the underlying assumption. This will also give insight into which of the assumtions (i.e. those for MNF and mCFPD, respectively) may be critical in specific practical situations.

It must be noted that the quality of the output of this method relies on the quality of the input data. That is to say, systematic errors in the flow measurements are directly reflected in the analysis results. Even though noise with a non-skewed distribution may be expected to average out in the CFPD analysis and therefore not significantly affect (m)CFPD outputs, offsets in the input data will be directly reflected in the b factors (background leakage estimates). More complex measurement error behavior may be reflected in both the a and b factors. As with any method, this is to be taken into consideration when interpreting the results.

Sensitivity tests and application to synthetic signals from hydraulic models demonstrate the validity of the method in ideal circumstances. Subsequent application of the method to real data shows the feasibility of the analysis. However, due to the lack of reliable true background leakage values, this does not suffice as a validation. The presentation of the method validated on synthetic data in this paper reflects the author's hope and expectation that attempts at validation will be done by researchers in the field with access to a range of suitable datasets.

Suggested future work

This paper describes a newly proposed method and provides a basic analysis of its sensitivity and applicability. It is quite clear, however, that more experience in applying this method is necessary to grow confidence in its applicability and generate insights into the practical bounds of the method. In addition to application to a range of field data sets for validation purposes, a number of questions and ideas remain for further investigation:

- As flow meters are sometimes selected for the expected flow during peak hours rather than the minimum night flow, accuracy issues in the lower flow range may provide an additional source of uncertainty for minimum night flow analyses. Even though these issues may also arise in the mCFPD approach, it would be worthwhile to investigate if excluding the parts of the flow data that are known to be within the low accuracy range of the flow meters used can improve the quality of the analysis results.
- It would be very valuable to obtain a better understanding of what makes flow patterns similar or dissimilar in the CFPD framework, in order to better understand the bounds of application of the mCFPD method.
- Using statistical filtering on the input data (e.g., using the 95% most conforming data) or other types of data preprocessing may make the method more robust.
- Understanding needs to be built on to what degree the method is applicable under conditions of intermittent supply.

ACKNOWLEDGEMENTS

The author would like to thank Brabant Water and WML water companies for supplying the data and models presented in this paper and Dick Bos (Brabant Water) and Henk Vogelaar (WML) for their support in understanding the data and models. Also, fruitful comments on an early version of the paper and discussions with Dragan Savić, as well as constructive reviews by two anonymous reviewers are gratefully acknowledged.

DATA AVAILABILITY STATEMENT

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

CONFLICT OF INTEREST

The author declares there is no conflict.

REFERENCES

Amoatey, P. K., Minke, R. & Steinmetz, H. 2018 Leakage estimation in developing country water networks based on water balance, minimum night flow and component analysis methods. *Water Practice and Technology* **13** (1), 96–105. https://doi.org/10.2166/wpt.2018.005.

Charalambous, B. & Laspidou, C. 2017 *Dealing with the Complex Interrelation of Intermittent Supply and Water Losses*, 1st edn. IWA Publishing, London.

Criminisi, A., Fontanazza, C. M., Freni, G. & Loggia, G. L. 2009 Evaluation of the apparent losses caused by water meter under-registration in intermittent water supply. *Water Science and Technology* **60** (9), 2373–2382.

Farley, M. & Trow, S. 2003 *Losses in Water Distribution Networks*. IWA Publishing, London.

Giustolisi, O., Savic, D. A. & Kapelan, Z. 2008 Pressure-driven demand and leakage simulation for water distribution networks. *Journal of Hydraulic Engineering* **134** (5), 626–635. doi: 10.1061/(ASCE)0733-9429(2008)134:5(626).

Lambert, A. 1994 Accounting for losses: the bursts and background concept. *Journal of CIWEM* **8**, 205–214.

Lambert, A. & Morrison, J. A. E. 1996 Recent developments in application of 'bursts and background estimates' concepts for leakage management. *Journal of CIWEM* **10**, 100–104.

Liemberger, R. & Wyatt, A. 2019 Quantifying the global non-revenue water problem. *Water Science & Technology: Water Supply* **19** (3). doi: 10.2166/ws.2018.129.

Marzola, I., Mazzoni, F., Alvisi, S. & Franchini, M. 2022 Leakage detection and localization in a water distribution network through comparison of observed and simulated pressure data. *Journal of Water Resources Planning and Management* **148** (10). doi: 10.1061/(ASCE)WR.1943-5452.0001503.

Puust, R., Kapelan, Z., Savic, D. & Koppel, T. 2010 A review of methods for leakage management in pipe networks. *Urban Water Journal* **7** (1), 25–45.

Romero, L., Blesa, J., Puig, V. & Cembrano, G. 2022 Clustering-Learning approach to the localization of leaks in water distribution networks. *Journal of Water Resources Planning and Management* **128** (3). doi: 10.1061/(ASCE)WR.1943-5452.0001527.

Rossman, L., Woo, H., Tryby, M., Shang, F., Janke, R. & Haxton, T. 2020 *EPANET 2.2 User Manual*. U.S. Environmental Protection Agency, EPA/600/R-20/133, Washington, DC.

Seltman, H. J. 2018 *Experimental Design and Analysis*. Available from: http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf (accessed 15 June 2022).

Steffelbauer, D. B., Deuerlein, J., Gilbert, D., Abraham, E. & Piller, O. 2022 Pressure-leak duality for leak detection and localization in water distribution systems. *Journal of Water Resources Planning and Management* **128** (3). doi: 10.1061/(ASCE)WR.1943-5452.0001515.

Totsuka, N., Trifunovic, N. & Vairavamoorthy, K. 2004 Intermittent urban water supply under water starving situations. In: *30th WEDC International Conference*. Vientiane, Lao, pp. 505–512.

Van Thienen, P. 2012 A method for quantitative discrimination in flow pattern evolution of water distribution supply areas with interpretation in terms of demand and leakage. *Journal of Hydroinformatics* **15** (1), 86–102. https://doi.org/10.2166/hydro.2012.171.

Van Thienen, P. & Vertommen, I. 2016 Automated feature recognition in CFPD analyses of DMA or supply area flow data. *Journal of Hydroinformatics* **18** (3), 514–530. https://doi.org/10.2166/hydro.2015.056.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F. & Van Mulbregt, P. & SciPy 1.0 Contributors 2020 Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods* **17** (3), 261–272.

Weston, S. L., Loubser, C., Jacobs, H. E. & Speight, V. 2022 Short-term impacts of the filling transition across elevations in intermittent water supply systems. *Urban Water Journal* 1–10. doi: 10.1080/1573062X.2022.2075764.

Wu, J., Ma, D. & Wang, W. 2022 Leakage identification in water distribution networks based on XGBoost algorithm. *Journal of Water Resources Planning and Management* **128** (3). doi: 10.1061/(ASCE)WR.1943-5452.0001523.

First received 17 May 2022; accepted in revised form 24 June 2022. Available online 1 July 2022