Simulation of a leak's growth process in water distribution systems based on growth functions

Guancheng Guo, Shuming Liu, Dailin Jia, Shanhe Wang and Xue Wu

ABSTRACT

Water loss in water distribution systems is one of the major problems faced by water utilities. The components of water losses should be accurately assessed and their priority should be determined. Generally, water balance analysis is used to quantify different components of water losses and identify the main contributor to high leakage rates. The leak flow rate is assumed to be static within a given calculation period during the calculation of real losses. Errors will inevitably arise during this process. This is mainly due to our limited understanding of a leak's growth process. To overcome this problem, the current work proposes the use of growth functions to represent a leak's growth process and establish a functional relationship between the leak flow rate and the leak duration. A leakage development model is adopted to simulate a leak's growth process and optimize the parameters of growth functions. The results show that the Richards function performs better than other growth functions and its mean absolute percentage error is 15.33%. Furthermore, the growth function could be used to calculate real losses and has the prospect of evaluating the effects of leakage detection.

Key words | growth function, leak duration, leak flow rate, real loss, water distribution system

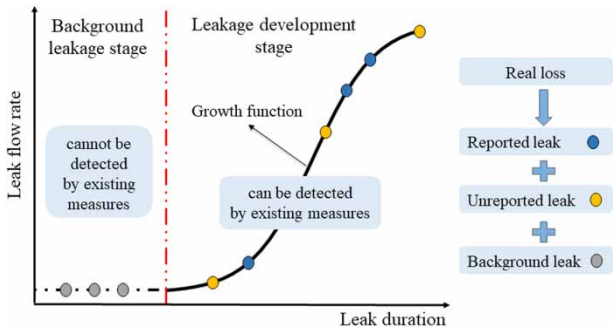
Guancheng Guo
Shuming Liu (corresponding author)
Shanhe Wang
Xue Wu
School of Environment, Tsinghua University,
Beijing 100084,
China
E-mail: *shumingliu@tsinghua.edu.cn*

Dailin Jia
Chengdu Municipal Waterworks Co., Ltd,
Chengdu 610072,
China

HIGHLIGHTS

- The growth function can be used to represent a leak's growth process.
- The leakage development model is adopted to simulate a leak's growth process.
- The proposed calculation method of real losses considers a leak's evolution process.
- The detection coefficient is proposed to evaluate leakage detection efficiency.

GRAPHICAL ABSTRACT



doi: 10.2166/aqua.2021.021

INTRODUCTION

Over the past few decades, the shortage of water resources has gradually received attention. Annually, about 20–30% of the total water supply in water distribution systems (WDS) will be lost (Al-Washali *et al.* 2016, 2020). Water loss increases the cost of water abstraction and delivery, and the use of electricity and chemicals, and may affect water quality (Eryiğit 2019; Xue *et al.* 2020). This is a challenge faced by many water utilities. It is not only an economic issue but also an environmental and safety issue (Güngör-Demirci *et al.* 2018; Guo *et al.* 2020).

A standard for quantifying and evaluating water loss is defined by the International Water Association (IWA). Water loss includes real loss and apparent loss (Al-Washali *et al.* 2020). Real loss mainly occurs in pipelines, reservoirs, and customer connections. Apparent loss is mainly due to unregistered customer meters, data processing or billing errors, and unauthorized use (Ríos *et al.* 2014; Ethem Karadirek 2019). In practice, water loss in WDS is inevitable. Reducing water loss to an acceptable range is feasible from the perspective of economic and environmental costs (Kanakoudis *et al.* 2012; Sakai *et al.* 2020). Real loss is the main form of water loss and depends on the characteristics of pipe networks. Three methods can be used to estimate real losses: (1) a top-down water balance analysis (Al-Washali *et al.* 2016); (2) a down-up minimum night flow (MNF) analysis (Alkasseh *et al.* 2013); and (3) a component analysis of leakage (Aboelnga *et al.* 2018). To effectively improve leakage control management, each component or subcomponent of real losses should be assessed, including background losses, unreported losses, and reported losses. Component analysis of leakage is a conventional method that breaks down real losses into subcomponents (Al-Washali *et al.* 2020).

Component analysis of leakage is also called burst and background estimates (BABE), which is an empirical model to analyze a certain part of real losses (Lambert 1994). In this method, real losses mainly include water losses caused by three types of leakage (Lambert *et al.* 1999). The first is background leakage. Leaks cannot be detected by advanced technologies or measures due to small leak flow rates. The second is unreported leakage. Leaks can be detected underground using leakage detection devices (e.g., noise loggers or correlators). The third is reported leakage. Leaks can be easily discovered by practitioners because water overflows to the ground. The lost volume of a leak is calculated as the leak flow rate multiplied by the calculation period (Lambert *et al.* 1999). Although background or unreported leakages have small leak flow rates, they can cause significant water losses due to their long leak duration (Aboelnga *et al.* 2018). Reported leakages often have low water losses due to their short leak duration (Lambert & Fantozzi 2005). The conventional method for calculating real losses is related to leak flow rates and calculation periods. Generally, the leak flow rate is assumed to be static and its evolution over time is ignored. The detection period (i.e., the time required for practitioners to detect all pipelines in a city) is usually used as the calculation period, which is not equal to the leak duration. This has great uncertainty because it mainly depends on the experience of practitioners. However, no verification is found in the literature for this assumption. There is limited knowledge of a leak's growth process and no lab work that studies the evolution of a leak in water supply pipes. As a result, the current calculation method of real losses is misleading and needs to be improved.

In real situations, a leak could exist for several detection periods until it is discovered, and its leak flow rate might vary over time. This is a dynamic process. In each detection period, there may be some leaks that can be detected by existing technologies or measures but eventually are not found. This is related to practitioners' experience or detection strategies. For example, it is not easy for inexperienced practitioners to find small leaks or use correlators to detect leaks under ambient noise. Furthermore, it is difficult to find a suitable calculation period for calculating real losses. Essentially, an optimal calculation period should be equal to the leak duration. There is insufficient consideration of this in the literature.

To fill the research gaps mentioned above, it is of great significance to understand a leak's growth process and clarify the mathematical relationship between the leak flow rate and the leak duration. The current work uses a growth

function (i.e., Logistic, Gompertz, and Richards functions) to represent a leak's growth process. The contributions of this study are summarized as follows:

(1) A growth function is used to describe a dynamic evolution process between the leak flow rate and the leak duration, which helps analyze a leak's growth process from a microscopic point of view.
(2) A leakage development model is adopted to simulate a leak's growth process, and a method to calculate real losses using growth functions is proposed. This helps provide a new method for accurately quantifying each component of real losses (i.e., unreported losses or reported losses).
(3) A detection coefficient is proposed to evaluate the effects of leakage detection, which helps provide a new angle to understand water utilities' leakage detection efficiency.

The remainder of this paper is organized as follows. The 'Methodology' section describes the growth function and the leakage development model. The 'Case study' section describes the data acquisition and modeling process. The 'Results and discussion' section discusses optimization results and the application of growth functions. The 'Conclusions' section summarizes this work and offers suggestions for future work.

METHODOLOGY

Growth function

Generally, pipeline leakage in WDS is caused by pipe sinking, pipe corrosion, pipe aging, and excessive external loads (Morais & de Almeida 2007). Excluding human interference (e.g., pipe bursts due to construction accidents), a leak's growth process is unidirectional and irreversible: the leak flow rate gradually increases over time until the leak is found and repaired by practitioners. In this study, it is assumed that a complete leak's growth process includes two steps, as shown in Figure 1. The first is the background leakage stage. These background leaks cannot be detected by existing technologies or measures due to low leak flow rates. The second is the leakage development stage. These

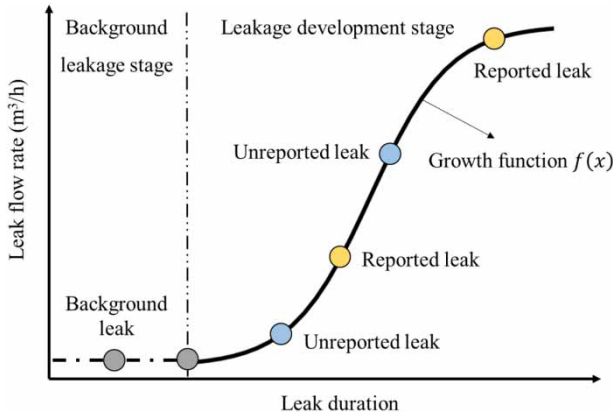


Figure 1 | Leak's growth process.

unreported or reported leaks can be discovered using existing technologies or measures, and their leak flow rates will gradually increase over time. Unreported leaks are discovered underground, and reported leaks can be easily found because water overflows to the ground.

The current work mainly investigates the leakage development stage and does not consider the background leakage stage. Assuming that a leak's growth process follows a certain model or function, the mathematical relationship between the leak flow rate and the leak duration could be established. Growth functions have been widely used to describe the growth law of an individual organism or predict the development trend of a certain technical or economic characteristic (Gallagher 2011; Rymuza 2017). Commonly used growth functions include the Logistic function, the Gompertz function, and the Richards function (Thornley & France 2005). For example, Morgan (1976) ingeniously used the Logistic function to describe the herding behavior of African elephants; Zachariadis *et al.* (1995) used the Gompertz function to simulate the growth of car ownership; Huo & Wang (2012) used the Richards function to simulate vehicle sales and stock in China. Inspired by these studies mentioned above, we try to use a growth function to represent a leak's growth process. Essentially, the leak's growth process is a complex physical, chemical, and biological reaction process caused by some microorganisms attached to the pipelines. This process is similar to the growth process of animals, plants, and microorganisms (Thornley & France 2005). These growth functions can

be expressed as

$$\text{Logistic function } f(x) = \frac{a}{(1 + e^{b-cx})} \quad (1)$$

$$\text{Gompertz function } f(x) = ae^{(-be^{-cx})} \quad (2)$$

$$\text{Richards function } f(x) = \frac{a}{(1 + e^{b-cx})^{\frac{1}{N}}} \quad (3)$$

where $f(x)$ denotes the leak flow rate (m^3/h), and x denotes the leak duration. a represents the extreme leak flow rate. b, c, and N are parameters that need to be optimized.

LEAKAGE DEVELOPMENT MODEL

Model structure

A leakage development model is developed to simulate a leak's growth process and optimize the parameters of growth functions, as shown in Figure 2. Leakage information and pipeline information are collected from a pipeline maintenance database. A real leakage dataset is divided into a low leak flow rate dataset and a high leak flow rate dataset. The main reason is that for the same probability density of leak flow rates, it is assumed that the leak duration of high leak flow rates is longer than that of low leak flow rates. In this paper, the kernel density estimation is used to calculate the probability density of leak flow rates. This is a basic data-smoothing approach inferring populations based on a finite data sample (Heidenreich *et al.* 2013). Then, a random sampling process is established to simulate a real leakage detection process. In each random sampling process, the simulated leak flow rates will increase according to growth functions. Finally, the mean square error (MSE) between real leak flow rates and simulated leak flow rates is selected as an objective function of the leakage development model. The optimal parameters of growth functions can be obtained by minimizing the objective function.

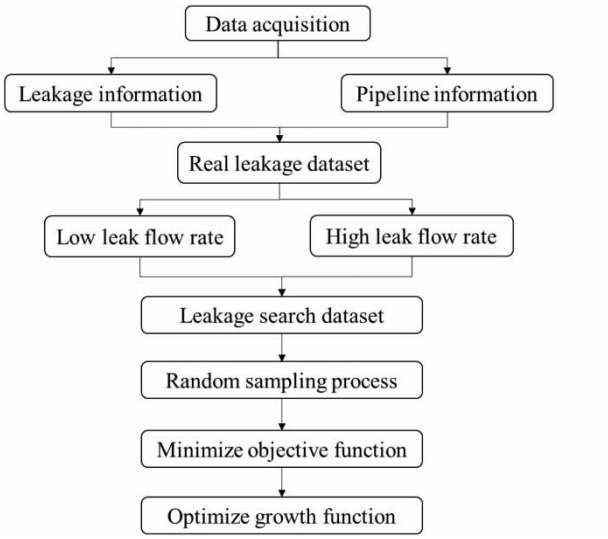


Figure 2 | Structure of the leakage development model.

Leakage search dataset

To describe the features of a leak, leakage information and pipeline information are used in this study. Leakage information includes detection date, detection period, leak type, and leak flow rate. Pipeline information includes pipe material, pipe diameter, pipe age, and pipe length. Leaks are classified into different categories according to the pipe material, pipe diameter, and pipe age. In this paper, H denotes the pipe length, h represents the detection distance (i.e., the distance between the leak and the practitioner when using a listening stick to detect leaks), and t denotes the detection period. In each detection period, the number of searches is H/h. In T years (i.e., the total time of data collection), the number of searches is m and the number of leaks that have been found is n. Then, real leakage dataset M can be expressed as

$$M = \{Q_i \mid Q_i < Q_{i+1}, \quad i = 1, 2, \ldots, n-1\} \quad (4)$$

where Q_i represents the leak flow rate, and n is the number of leaks that have been discovered in T years.

Furthermore, the probability density estimation of leak flow rates in dataset M is $P(Q_i)$. The maximum probability density estimation is $P_{\max}(Q_l)$, which corresponds to the lth leak in dataset M. Then, dataset M is divided into a low leak flow rate dataset $M1$ and a high leak flow rate

dataset $M2$. They can be expressed as

$$M1 = \{Q_{i1} \mid Q_{i1} < Q_{i1+1}, \quad i1 = 1, 2, \ldots, l\} \tag{5}$$

$$M2 = \{Q_{i2} \mid Q_{i2} < Q_{i2+1}, \quad i2 = l+1, l+2, \ldots, n-1\} \tag{6}$$

where $i1$ represents the $i1$th leak in dataset $M1$, and $i2$ represents the $i2$th leak in dataset $M2$.

Finally, a leakage search dataset that consists of zero points and nonzero points is established. The number of zero points is m-n and their corresponding values are equal to 0. This indicates that no leaks are discovered. The number of nonzero points is n and their corresponding values are equal to leak flow rates Q_i in dataset M. This indicates that leaks are discovered.

Random sampling process

Although water utilities have formulated detailed leakage detection arrangements, there are still many uncertainties in practice. Whether a leak can be found depends not only on its size or leak flow rate, but also on the experience of practitioners, the precision of detection devices, the intensity of ambient noise, and other factors. This study establishes a random sampling process to simulate a real leakage detection process, as shown in Figure 3. The statistical probability distribution of real leak flow rates in T years represents the overall level of the probability distribution of leak flow rates in each detection period. Each random sampling process represents a real leakage detection process in a detection period. The probability distribution of simulated leak flow rates is random and can represent the probability distribution of leak flow rates during each

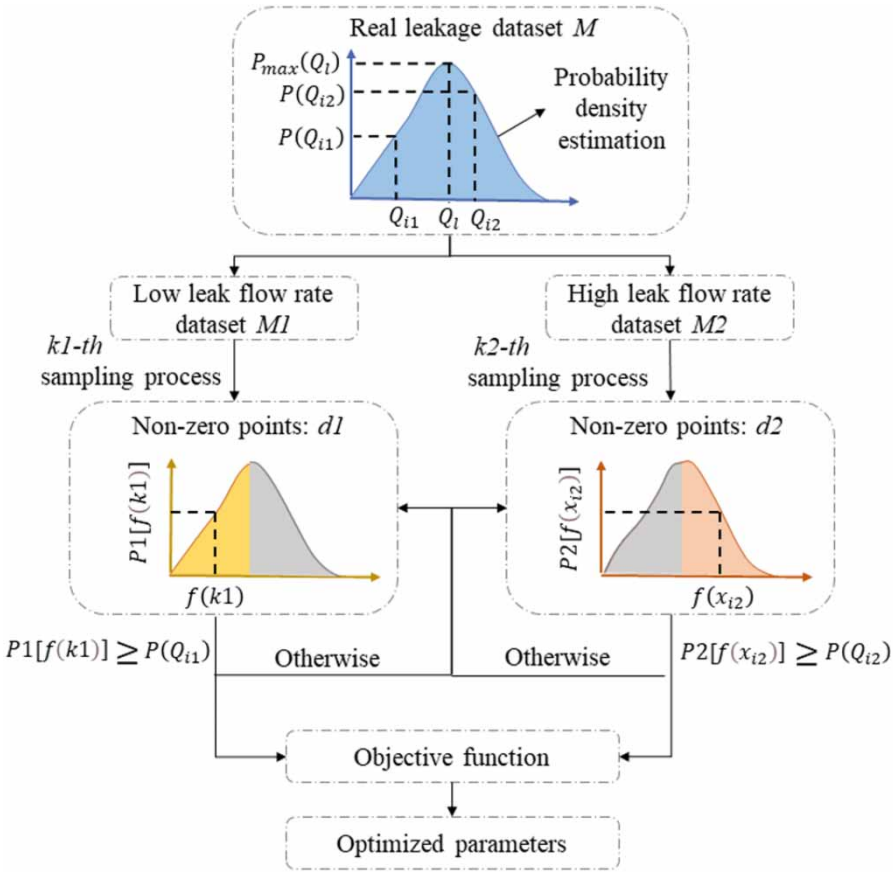


Figure 3 | Flowchart of the random sampling process.

random sampling process. In the leakage development model, by comparing the probability distribution of real leak flow rates and simulated leak flow rates, it can be determined whether a leak is found. The random sampling process contains three parts.

Firstly, a real leak A is selected from dataset $M1$. Its leak flow rate is Q_{i1} and its probability density estimation is $P(Q_{i1})$. In each random sampling process, H/h samples are randomly selected from the leakage search dataset. The classical Knuth–Durstenfeld shuffle algorithm (Fisher & Yates 1963) is used to ensure that the zero and nonzero points in the leakage search dataset can be selected equally. The random sampling process in dataset $M1$ is as follows.

Step I: In the $k1$th random sampling process $(1 \le k1 \le [x_l] + 1)$, H/h samples are randomly selected from the leakage search dataset, which contains $d1$ nonzero points. This indicates that practitioners implemented H/h searches and found $d1$ leaks in the kth detection period.
Step II: The simulated leak flow rate of leak A is $f(k1)$ and its probability density estimation is $P1[f(k1)]$. The value of $f(k1)$ is calculated by growth functions. If $P1[f(k1)]$ is larger than $P(Q_{i1})$, the real leak A is found. Otherwise, continue to repeat step I to step II. This indicates that when the probability density of simulated leak flow rates is larger than the average probability density of real leak flow rates, the real leak A in dataset $M1$ will be discovered.
Step III: The sum of squared errors of real leaks in dataset $M1$ is calculated (i.e., $\sum_{i1=1}^{l} [f(k1) - Q_{i1}]^2$).

Secondly, a real leak B is selected from dataset $M2$. Its leak flow rate is Q_{i2} and its probability density estimation is $P(Q_{i2})$. In dataset $M2$, the leak duration before the real leak B is $x_p (p = l, l+1, \ldots, n-1)$. The random sampling process in dataset $M2$ is as follows:

Step I: In the $k2$th random sampling process $(1 \le k2 \le (T/t) - [x_l] - 1)$, H/h samples are randomly selected from the leakage search dataset, which contains $d2$ nonzero points. The leak duration of leak B is $x_{i2} = x_p + k2$ $(i2 = l+1, \ldots, n;\ p = l, \ldots, n-1)$.
Step II: The simulated leak flow rate of leak B is $f(x_{i2})$ and its probability density estimation is $P2[f(x_{i2})]$. If $P2[f(x_{i2})]$ is larger than $P(Q_{i2})$, the real leak B is found. Otherwise, continue to repeat step I to step II.
Step III: The sum of squared errors of real leaks in dataset $M2$ is calculated (i.e., $\sum_{i2=l+1}^{n} [f(x_{i2}) - Q_{i2}]^2$).

Thirdly, the objective function of the leakage development model is the sum of MSE between real leak flow rates and simulated leak flow rates. When the objective function reaches the minimum value, the corresponding parameters of growth functions are optimal. It can be expressed as

$$\text{Objective function} = \frac{\sum_{i1=1}^{l} [f(k1) - Q_{i1}]^2 + \sum_{i2=l+1}^{n} [f(x_{i2}) - Q_{i2}]^2}{n} \tag{7}$$

CASE STUDY

Data acquisition

Pipeline information and leakage information are collected in City DC from 2007 to 2016. In City DC, practitioners detect leaks on all water supply pipelines every 3 months by using listening sticks or noise correlators. The average detection distance is 100 m, and the detection period is 3 months. Table 1 presents the main pipeline information in City DC.

Table 2 presents the main leakage information in City DC. The leak flow rate is calculated according to the orifice type function (Puust *et al.* 2010).

Parameter selection of models

The leakage development model is developed based on the leakage search dataset from 2007 to 2016. Different growth functions have different parameters. Essentially, the Richards function can be regarded as a combination of the Logistic function and the Gompertz function. When parameter N is equal to 1, the Richards function is the Logistic function. When parameter N tends to zero, the Richards function is close to the Gompertz function. Parameter a is equal to the maximum leak flow rate. Parameter b ranges from 1 to 20 with an increase of 1. Parameters c and N range from 0 to 1 with an increase of 0.1.

Table 1 | Pipeline information

Pipe diameter	Pipe material	Pipe age (Year)	Pipe length (m)	Detection distance (m)	Detection period (months)	Number of leaks
DN100	Cast iron	5–15	1,009,300	100	3	987
DN100	Steel	5–10	49,400	100	3	238
DN150	Cast iron	5–15	125,800	100	3	226
DN200	Cast iron	5–20	407,000	100	3	406
DN300	Cast iron	5–20	998,900	100	3	195
≥DN400	Cement	5–20	1,382,900	100	3	220
≥DN400	Cast iron	5–20	330,900	100	3	188

Table 2 | Leakage information

Date	Leak type	Leak position	Leak flow rate (m^3/h)
09 Jan 2007	DN100 cast iron	Transverse crack	11.82
18 Apr 2010	DN800 cement	Longitudinal crack	41.21
⋮	⋮	⋮	⋮
07 Aug 2014	DN100 steel	Pipe connection leak	5.21
21 Dec 2016	DN400 cast iron	Tee leak	19.31

For the probability density estimation, the uniform function, the triangular function, and the Gaussian function are commonly used kernel functions. Prakasa Rao (1983) proved that different kernel functions have little effect on the non-parametric density estimation. The Gaussian function is selected as a kernel function and the optimal bandwidth can be calculated according to an empirical method (Silverman 1988). The optimal bandwidth is $\varphi = 1.06\sigma n^{-0.2}$, where σ is the standard deviation of samples, and n is the number of samples. Table 3 shows the optimized parameters for different leakage development models.

Performance indicator of models

Four indicators are used to evaluate the performance of leakage development models (Bennett *et al.* 2013): mean absolute error (MAE), mean absolute percentage error (MAPE), relative entropy (RE), and Nash–Sutcliffe model efficiency (NSE).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y - Y'_i| \tag{8}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - Y'_i|}{|Y_i|} \tag{9}$$

$$\mathrm{RE} = \sum P(x)\log\frac{P(x)}{Q(x)} \tag{10}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y'_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y_i})^2} \tag{11}$$

Table 3 | Optimized parameters for leakage development models

Leak type	Growth function	*H/h*	φ	Q_l	*a*	*b*	*c*	*N*
DN100 cast iron	Richards	10,093	2.65	5.74	85	3	0.20	0.8
DN100 steel	Richards	494	4.70	5.85	78	3	0.20	0.7
DN150 cast iron	Richards	1,258	4.69	6.81	86	3	0.20	0.85
DN200 cast iron	Richards	4,047	4.31	6.64	97	3	0.2	0.8
DN300 cast iron	Richards	9,989	6.08	7.55	92	3	0.2	0.75
≥DN400 cement	Richards	13,829	6.51	12.41	110	1	0.20	0.20
≥DN400 cast iron	Richards	3,309	9.27	10.72	110	1	0.20	0.20

where Y_i is a real leak flow rate, and $\overline{Y_i}$ is a mean leak flow rate. Y_i' is a simulated leak flow rate, and n is the number of leaks. $P(x)$ is the probability distribution of real leak flow rates, and $Q(x)$ is the probability distribution of simulated leak flow rates. The RE indicator is used to evaluate the similarity of probability distributions. The closer the RE is to 0, the closer the probability distribution is. The closer the NSE is to 1, the more accurate the model is.

RESULTS AND DISCUSSION

Optimization results

Table 4 presents the performance of different leakage development models. The results show that the Richards function performs better than the Logistic and Gompertz functions. Its average MAPE is 15.33%. The Gompertz function performs worse than other growth functions. Its average MAPE is 22.83%. This indicates that the Richards function is more suitable for simulating the leak's growth process. We can also observe that MAPE values for three growth functions are relatively high. The main reason is that the prediction error of the low leak flow rate may lead to a larger MAPE value. This indicates that the performance of leakage development models for low leak flow rates still needs to be improved. For ≥DN400 cement pipelines, the model performance is poor because there are fewer leaks that can be used to optimize growth functions. For the same leak type, the leakage development model needs data for at least 200 leaks to obtain a satisfactory result.

The growth function and the probability distribution of leak flow rates for different pipelines are similar, as shown in Figures 4–6. This is an ideal growth curve. A leak is discovered with gradually increasing leak flow rates under natural conditions. In practice, leaks discovered by practitioners could be located anywhere on growth functions. Most leak flow rates are less than 25 m^3/h and the leak duration is less than 2 years. A few leak flow rates are more than 50 m^3/h and the leak duration exceeds 5 years. The potential reasons are given as follows: (1) some pipelines may not be detected by practitioners and (2) large ambient

Table 4 | Performance of leakage development models

						Parameters			
Leak type	**Growth function**	**MAE (m^3/h)**	**MAPE (%)**	**RE**	**NSE**	***a***	***b***	***c***	***N***
DN100 cast iron	Logistic	0.05	17.65	0.015	0.97	85	5	0.30	
	Gompertz	0.06	20.08	0.022	0.96	85	15	0.20	
	Richards	0.04	16.92	0.012	0.98	85	3	0.20	0.80
DN100 steel	Logistic	0.08	11.97	0.005	0.99	78	5	0.30	
	Gompertz	0.12	16.86	0.011	0.98	78	10	0.20	
	Richards	0.06	10.74	0.001	0.99	78	3	0.20	0.70
DN150 cast iron	Logistic	0.12	13.97	0.007	0.98	86	6	0.30	
	Gompertz	0.17	23.47	0.021	0.96	86	11	0.20	
	Richards	0.09	11.66	0.004	0.99	86	3	0.20	0.85
DN200 cast iron	Logistic	0.10	19.96	0.011	0.98	97	7	0.30	
	Gompertz	0.20	19.86	0.009	0.97	97	7	0.20	
	Richards	0.08	18.36	0.008	0.99	97	3	0.20	0.80
DN300 cast iron	Logistic	0.13	13.52	0.005	0.99	92	6	0.30	
	Gompertz	0.16	23.54	0.005	0.98	92	13	0.20	
	Richards	0.12	11.80	0.002	0.99	92	3	0.20	0.75
≥DN400 cement	Logistic	0.17	17.97	0.006	0.98	110	6	0.30	
	Gompertz	0.19	19.81	0.006	0.98	110	16	0.20	
	Richards	0.16	16.93	0.003	0.98	110	1	0.20	0.20
≥DN400 cast iron	Logistic	0.20	23.07	0.004	0.99	110	10	0.40	
	Gompertz	0.27	36.20	0.011	0.98	110	14	0.20	
	Richards	0.19	20.91	0.003	0.99	110	1	0.20	0.20

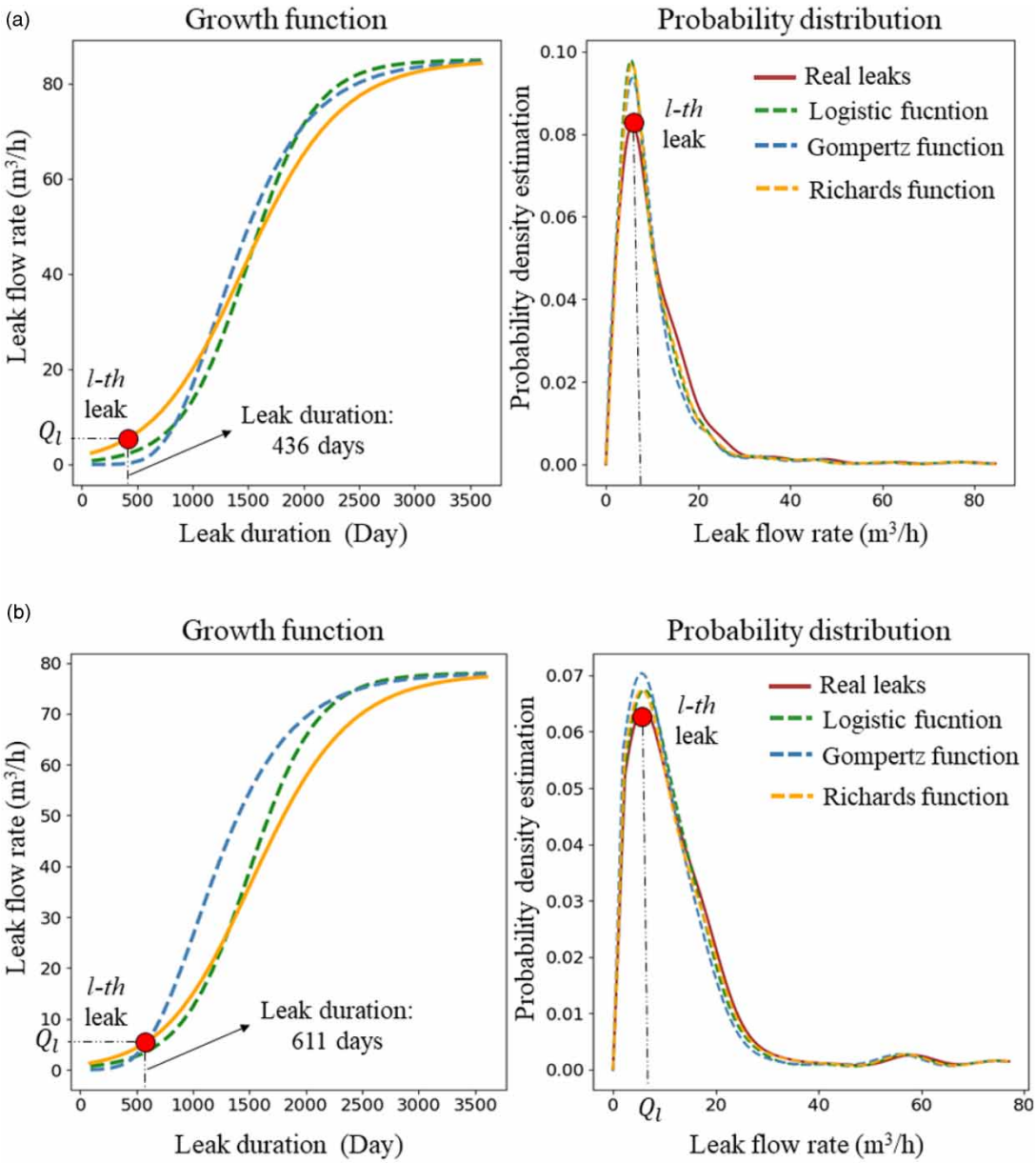


Figure 4 | Growth function and probability distribution for (a) DN100 cast iron and (b) DN100 steel pipes.

noise could reduce leakage detection accuracy. Furthermore, the probability distribution of simulated leak flow rates using the Richards function is close to the probability distribution of real leak flow rates, which further demonstrates that the Richards function outperforms other growth functions. For the same pipe material, the larger the pipe diameter, the longer the average leak duration. For the same pipe diameter, the steel and cement pipes have a longer average leak duration than cast iron pipes due to strong corrosion resistance. The leak's growth process is related to pipe properties.

Rationality of the random sampling process

In the leakage development model, the probability distribution of leak flow rates is selected as an indicator to stop the random sampling process. When a simulated leak has a higher kernel density estimation than a real leak, it

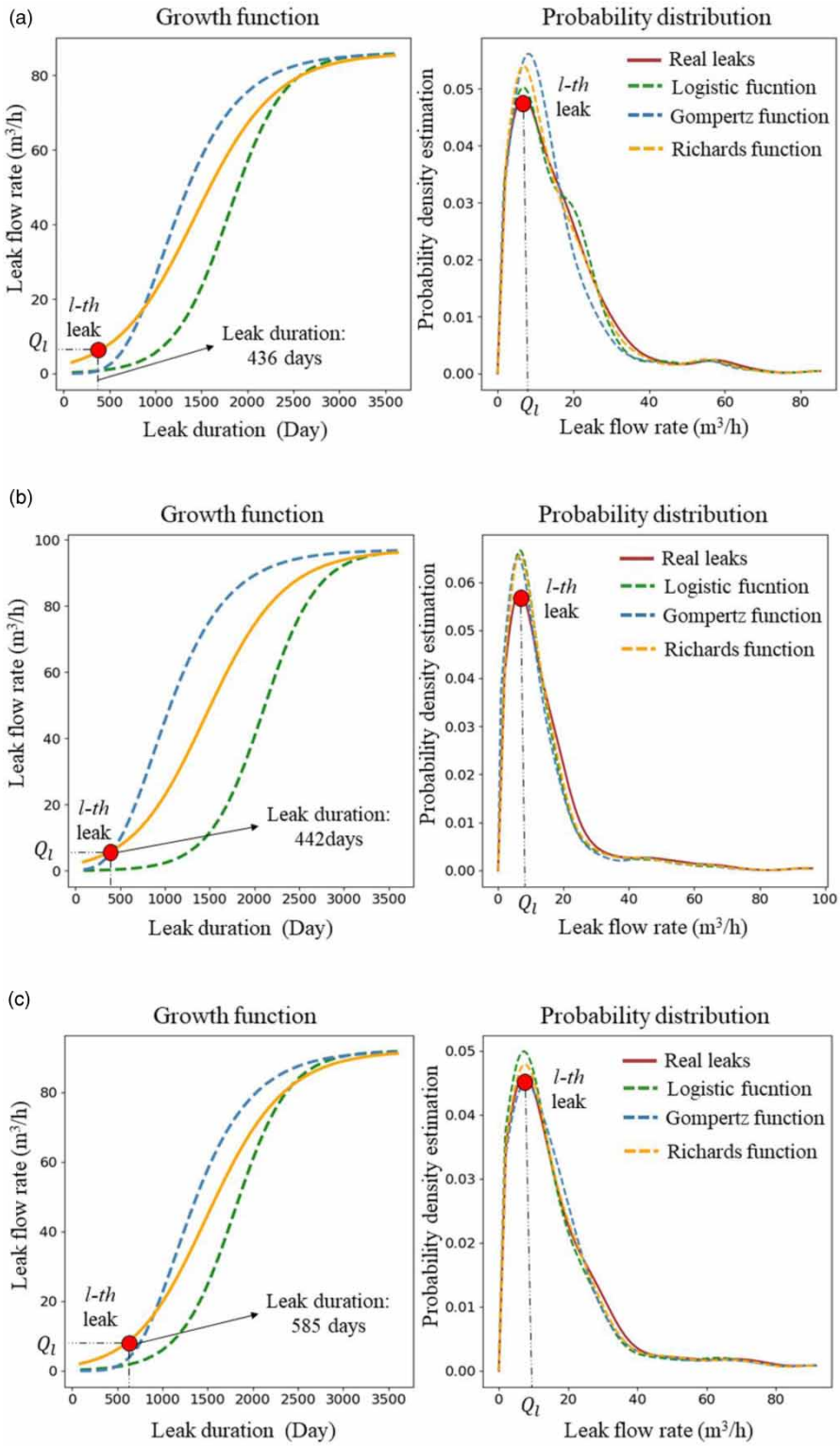


Figure 5 | Growth function and probability distribution for (a) DN150, (b) DN200, and (c) DN300 cast iron pipes.

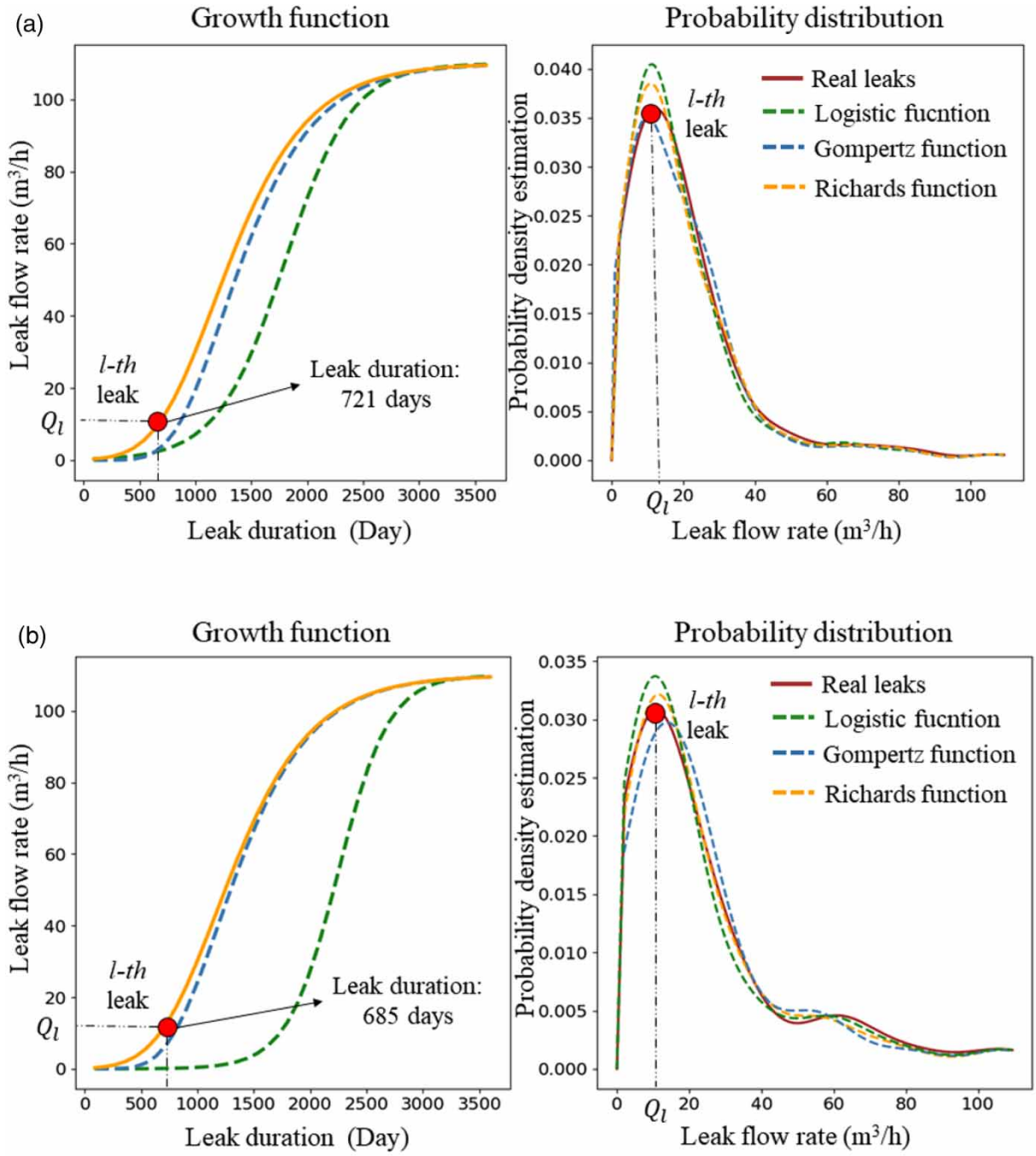


Figure 6 | Growth function and probability distribution for (a) ≥DN400 cement and (b) ≥DN400 cast iron pipes.

indicates that the probability of the simulated leak occurring is higher than the historical average probability. In this case, the simulated leak is considered detected. The frequency (i.e., $(d1 + d2)/(H/h)$) converges to the probability P based on the law of large numbers. This can be expressed as

$$\lim_{H/h\to\infty}\left\{\left|\frac{d1+d2}{H/h}-P\right|\geq\varepsilon\right\}=0 \tag{12}$$

where $d1$ and $d2$ obey a Binomial distribution, i.e., $d1 + d2 \sim B(H/h, P)$, and ε is an error coefficient. According to Chebyshev's theorem, the following equation is satisfied

$$P\left(\left|\frac{d1+d2}{H/h}-P\right|\geq\varepsilon\right)\leq\frac{\mathrm{Var}\left(\frac{d1+d2}{H/h}\right)}{\varepsilon^2}=\frac{P(1-P)h}{H\varepsilon^2} \quad (13)$$

Using the DN100 cast iron pipeline as an example, $d1 + d2 \sim B(10093, 0.0024)$, and ε is equal to 0.01. It is given in the form

$$P\left(\left|\frac{d1+d2}{H/h}-P\right|\geq\varepsilon\right)\leq\frac{P(1-P)h}{H\varepsilon^2}=0.0024 \quad (14)$$

when the number of samples is equal to 10,093, the probability of a large deviation is less than 0.24%. The larger the number of samples, the smaller the deviation of the random sampling process. For each random sampling process, the frequency of leaks at a certain leak flow rate can be regarded as the probability of leaks. The random sampling process is reasonable and can represent the average probability distribution of leak flow rates in dataset M.

Application of the growth function

The growth function is used to establish the mathematical relationship between the leak flow rate and the leak duration. In this study, real losses mainly include unreported losses and reported losses, which can be calculated by integrating growth functions. For example, a leak was discovered on a DN100 cast iron pipe on 5 September 2013, and its leak flow rate was 5.74 m^3/h. According to the Richards function, this leak appeared on 27 June 2012, and its leak duration was 435 days. The total real loss caused was 35,544 m^3. However, according to the conventional method (i.e., component analysis of leakage) for calculating real losses, the leak flow rate is considered a constant value during a given calculation period (Al-Washali *et al.* 2020). The total real loss is equal to the leak flow rate multiplied by the calculation period. If the calculation period is equal to 3 months, the total real loss caused is 12,398 m^3.

Figure 7 shows the comparison of real losses under different scenarios. For scenario 1, unreported leak A is discovered in the current detection period. Areas A1 and A2 represent real losses calculated by the leak flow rate multiplied by the calculation period, and area A2 represents real losses calculated by integrating growth functions. In this case, real losses calculated by the conventional method are larger than real losses calculated by growth functions. For scenario 2, unreported leak B is discovered in the next detection period, and real losses for the current detection period are calculated by integrating growth functions. However, the conventional method ignores real losses caused by unreported leak B. For scenario 3, reported leak C is discovered in the current detection period. Area C1 represents real losses calculated by the leak flow rate multiplied by the repair duration, and area C2 represents real losses calculated by integrating growth functions. In this case, real losses calculated by the conventional method are not necessarily greater than real losses calculated by

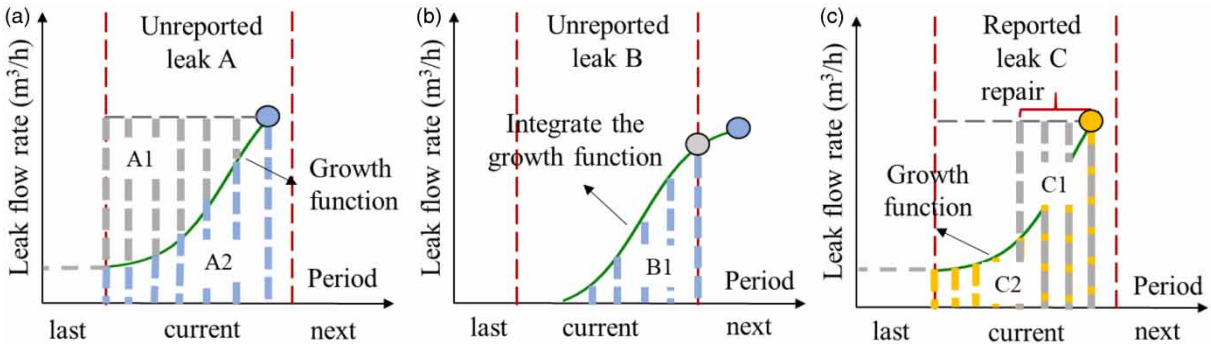


Figure 7 | Comparison of real losses under different scenarios.

growth functions. The results illustrate that the conventional method ignores a leak's growth process, so errors will inevitably occur when calculating real losses. The proposed method for calculating real losses by growth functions considers the dynamic evolution process of leak flow rates, which is in line with the actual situation. A novel calculation method of real losses proposed in this paper is as follows:

(1) For unreported leakage, real losses can be calculated by integrating growth functions, which includes leaks that have been discovered by existing technologies (e.g., unreported leak A in Figure 7) and leaks that can be detected by existing technologies but are not discovered (e.g., unreported leak B in Figure 7).
(2) For reported leakage, real losses caused by human factors are equal to the leak flow rate multiplied by the repair duration, and real losses caused by nonhuman factors can be calculated by integrating growth functions (e.g., reported leak C in Figure 7).
(3) For the current detection period, real losses caused by unreported leak A and reported leak C correspond to detected losses, and real losses caused by unreported leak B correspond to undetected losses.

Table 5 presents a comparison of real losses between the conventional method and the proposed method using growth functions. For different calculation periods (i.e., 3, 6, 9, and 12 months), the errors between the conventional method and the proposed method are different. When the calculation period is equal to 3 months, the errors are larger than those of other calculation periods. When the calculation period is equal to 6 months, real losses calculated by the proposed method are close to real losses calculated by the conventional method. During the same calculation period, the number of leaks used to calculate real losses in the conventional method is less than the number of leaks used to calculate real losses in the proposed method. The main reason is that many leaks that can be detected by existing technologies or measures are not discovered (e.g., unreported leak B in Figure 7). For the conventional method, the calculation period is uncertain and mainly depends on the experience of practitioners. Our proposed method for calculating real losses will not be influenced by these factors. In a certain calculation period, practitioners discover fewer leaks, which does not mean that there are no other leaks on the pipelines. According to our assumptions, a leak does not occur suddenly, but will gradually

Table 5 | Real loss between the conventional method and the proposed method

				DN100 cast iron pipe		DN100 steel pipe		
Year	Period	Month	Method	Real loss (m^3)	Nb	Real loss (m^3)	Nb	Error (%)
2008	3	1–3	C	952,000	35	596,151	8	26.13[a]
			P	1,200,810	188	492,288	82	17.42[b]
		4–6	C	631,028	29	280,287	10	82.51[a]
			P	1,151,710	189	508,537	84	81.43[b]
		7–9	C	249,768	14	168,784	12	387.92[a]
			P	1,218,662	177	497,633	78	194.83[b]
		10–12	C	770,691	21	161,926	12	59.16[a]
			P	1,226,623	182	426,349	66	163.3[b]
	6	1–6	C	3,166,058	64	1,752,878	18	24.88[a]
			P	2,378,467	224	1,011,927	92	42.27[b]
		7–12	C	2,040,919	35	661,421	24	20.48[a]
			P	2,458,897	196	929,054	78	40.46[b]
	9	1–9	C	5,498,392	78	3,135,670	30	34.34[a]
			P	3,610,198	241	1,515,302	96	51.68[b]
	12	1–12	C	10,413,955	99	4,828,598	42	53.42[a]
			P	4,850,432	260	1,946,722	96	59.68[b]

Note: C indicates the conventional component analysis of leakage. P indicates the proposed method using growth functions. Nb indicates the number of leaks that have been used to calculate real losses.

[a]The calculation errors of real losses between the conventional method and the proposed method for DN100 cast iron pipes.

[b]The calculation errors of real losses between the conventional method and the proposed method for DN100 steel pipes.

grow until it is discovered. The number of leaks used to calculate real losses in each calculation period should be stable. The proposed method may be more in line with the actual situation than the conventional method.

Figure 8 presents calculation errors for real losses calculated using the conventional method and the proposed method from 2007 to 2014 in City DC. The results show that when the calculation period is equal to 6 months, the absolute percentage error is low. When the calculation period is equal to 3 months, the absolute percentage error has a large deviation. The number of leaks discovered by practitioners during a short calculation period has great uncertainty because it may be affected by factors such as practitioners' experience, equipment performance, and ambient noise. This indicates that 6 months may be a suitable calculation period for the conventional method. In real cases, it is difficult to select an optimal calculation period for calculating real losses. The leak duration cannot be obtained for the conventional method, which may lead to calculation errors of real losses. However, the proposed method can avoid the influence of the calculation period on real losses.

Meanwhile, the growth function can be used to evaluate the effects of leakage detection. For example, Table 6 shows real losses in DN100 cast iron pipes in City DC. The detection coefficient is equal to detected losses divided by real losses, which can reflect the leakage detection efficiency of practitioners. The higher the detection coefficient, the higher the detection efficiency of practitioners. The results show that detection coefficients gradually increase from 2007 to 2014, which indicates that the detection effect has improved during these years. For some areas with low detection coefficients, water utilities should further improve leakage detection strategies (e.g., increase the frequency of leakage detection).

CONCLUSIONS

This study investigates the potential of using a growth function to represent a leak's growth process. The growth function can be used to calculate real losses and assess leakage detection efficiency. The following conclusions can be drawn:

(1) The growth function can be used to represent a leak's growth process. The Richards function is more suitable

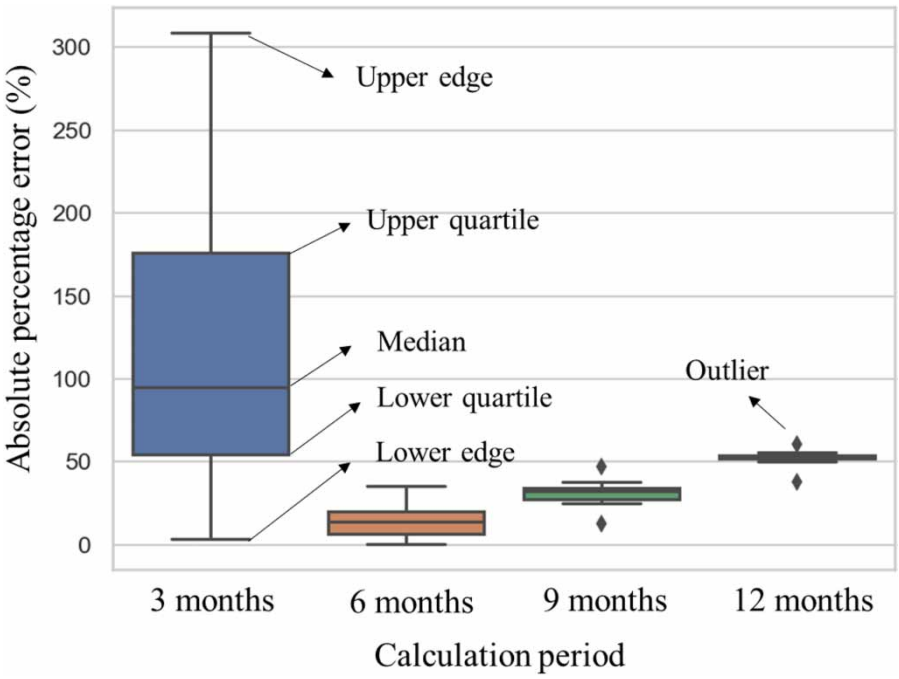


Figure 8 | Box plot of calculation errors for real losses from 2007 to 2014 in City DC.

Table 6 | Real loss in DN100 cast iron pipes in City DC

Year	Real loss (m^3)	Number of leaks	Detected loss (m^3)	Undetected loss (m^3)	Detection efficiency
2007	4,772,672	238	1,011,505	3,761,167	0.27
2008	4,850,432	260	1,241,443	3,608,989	0.34
2009	4,864,127	227	996,908	3,867,219	0.26
2010	4,603,657	218	1,199,618	3,404,039	0.35
2011	4,038,503	214	988,037	3,050,466	0.32
2012	3,880,504	238	1,194,472	2,686,032	0.44
2013	4,422,274	253	1,453,573	2,968,701	0.49
2014	4,457,025	240	1,570,633	2,886,391	0.54

for establishing the mathematical relationship between the leak flow rate and the leak duration than the Logistic and Gompertz functions.

(2) The leakage development model can be used to simulate a leak's growth process and optimize the parameters of growth functions. The random sampling process can be used to simulate a real leakage detection process, and its rationality has been verified.

(3) The proposed calculation method of real losses considers a leak's dynamic evolution process, which helps provide a new angle to further understand each component of real losses. However, the conventional method ignores a leak's growth process, which may lead to large errors.

(4) The detection coefficient may be used as an indicator to evaluate leakage detection efficiency. Water utilities could formulate leakage detection measures based on the detection coefficients of different regions.

It is suggested that future works test the leakage development model on a large leakage dataset. For small leak flow rates, consider shortening the horizontal time scale and make accurate assessments. In addition, the calculation method of real losses based on growth functions still needs to be improved and is compared with the actual real losses. Further research could investigate the influence of pipe properties (e.g., pipe age) on the proposed method.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (51879139).

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY STATEMENT

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

REFERENCES

Aboelnga, H., Saidan, M., Al-Weshah, R., Sturm, M., Ribbe, L. & Frechen, F.-B. 2018 Component analysis for optimal leakage management in Madaba, Jordan. *Journal of Water Supply: Research and Technology – AQUA* **67** (4), 384–396.

Alkasseh, J. M. A., Adlan, M. N., Abustan, I., Aziz, H. A. & Hanif, A. B. M. 2013 Applying minimum night flow to estimate water loss using statistical modeling: a case study in Kinta Valley, Malaysia. *Water Resources Management* **27** (5), 1439–1455.

Al-Washali, T., Sharma, S. & Kennedy, M. 2016 Methods of assessment of water losses in water supply systems: a review. *Water Resources Management* **30** (14), 4985–5001.

Al-Washali, T., Sharma, S., Lupoja, R., Al-Nozaily, F., Haidera, M. & Kennedy, M. 2020 Assessment of water losses in distribution networks: methods, applications, uncertainties, and implications in intermittent supply. *Resources, Conservation and Recycling* **152**, 104515–104526.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D. & Andreassian, V. 2013 Characterising performance of environmental models. *Environmental Modelling & Software* **40**, 1–20.

Eryiğit, M. 2019 Water loss detection in water distribution networks by using modified Clonalg. *Journal of Water Supply: Research and Technology – AQUA* **68** (4), 253–263.

Ethem Karadirek, I. 2019 An experimental analysis on accuracy of customer water meters under various flow rates and water pressures. *Journal of Water Supply: Research and Technology – AQUA* **69** (1), 18–27.

Fisher, R. & Yates, F. 1963 *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver and Boyd, Edinburgh.

Gallagher, B. 2011 Peak oil analyzed with a logistic function and idealized Hubbert curve. *Energy Policy* **39** (2), 790–802.

Güngör-Demirci, G., Lee, J., Keck, J., Guzzetta, R. & Yang, P. 2018 Determinants of non-revenue water for a water utility in California. *Journal of Water Supply: Research and Technology – AQUA* **67** (3), 270–278.

Guo, G., Yu, X., Liu, S., Xu, X., Ma, Z., Wang, X., Huang, Y. & Smith, K. 2020 Novel leakage detection and localization method based on line spectrum pair and cubic interpolation search. *Water Resources Management* **34**, 3895–3911.

Heidenreich, N. B., Schindler, A. & Sperlich, S. 2013 Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *ASTA – Advances in Statistical Analysis* **97** (4), 403–433.

Huo, H. & Wang, M. 2012 Modeling future vehicle sales and stock in China. *Energy Policy* **43**, 17–29.

Kanakoudis, V., Tsitsifli, S. & Papadopoulou, A. 2012 Integrating the carbon and water footprints' costs in the water framework directive 2000/60/EC full water cost recovery concept: basic principles towards their reliable calculation and socially just allocation. *Water* **4** (1), 45–62.

Lambert, A. 1994 Accounting for losses – the bursts and background concept. *Journal of the Institution of Water and Environmental Management* **8** (2), 205–214.

Lambert, A. O. & Fantozzi, M. 2005 Recent advances in calculating economic intervention frequency for active leakage control, and implications for calculation of economic leakage levels. *International Conference on Water Economics, Statistics and Finance* **5** (6), 263–271.

Lambert, A. O., Brown, T. G., Takizawa, M. & Weimer, D. 1999 A review of performance indicators for real losses from water supply systems. *Journal of Water Supply: Research and Technology – AQUA* **48** (6), 227–237.

Morais, D. C. & de Almeida, A. T. 2007 Group decision-making for leakage management strategy of water network. *Resources, Conservation and Recycling* **52** (2), 441–459.

Morgan, B. J. T. 1976 Stochastic models of grouping changes. *Advances in Applied Probability* **8** (1), 30–57.

Prakasa Rao, B. L. S. 1983 Nonparametric functional estimation. *Journal of the American Statistical Association* **81** (393), 483–512.

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

Ríos, J. C., Santos-Tellez, R. U., Rodríguez, P. H., Leyva, E. A. & Martínez, V. N. 2014 Methodology for the identification of apparent losses in water distribution networks. *Procedia Engineering* **70**, 238–247.

Rymuza, K. 2017 Application of a logistic function to describe the growth of Fodder Galega. *Journal of Ecological Engineering* **18** (1), 125–131.

Sakai, H., Satake, M., Arai, Y. & Takizawa, S. 2020 Report cards for aging and maintenance assessment of water-supply infrastructure. *Journal of Water Supply: Research and Technology – AQUA* **69** (4), 355–364.

Silverman, B. W. 1988 Density estimation for statistics and data analysis. *Journal of the American Statistical Association* **83** (401), 269–270.

Thornley, J. H. M. & France, J. 2005 An open-ended logistic-based growth function. *Ecological Modelling* **184** (2–4), 257–261.

Xue, Z., Tao, L., Fuchun, J., Riehle, E., Xiang, H., Bowen, N. & Singh, R. P. 2020 Application of acoustic intelligent leak detection in an urban water supply pipe network. *Journal of Water Supply: Research and Technology – AQUA* **69** (5), 512–520.

Zachariadis, T., Samaras, Z. & Zierock, K. H. 1995 Dynamic modeling of vehicle populations: an engineering approach for emissions calculations. *Technological Forecasting & Social Change* **50** (2), 135–149.

First received 31 January 2021; accepted in revised form 14 March 2021. Available online 25 March 2021