Water Resources Management (2024) 38:603–619
https://doi.org/10.1007/s11269-023-03687-7

Investigating the Impact of Cumulative Pressure-Induced Stress on Machine Learning Models for Pipe Breaks

Charalampos Konstantinou[1] · Carlos Jara-Arriagada[2] · Ivan Stoianov[2]

Received: 20 June 2023 / Accepted: 27 November 2023 / Published online: 30 December 2023

Abstract

Significant financial resources are needed for the maintenance and rehabilitation of water supply networks (WSNs) to prevent pipe breaks. The causes and mechanisms for pipe breaks vary between different WSNs. However, it is commonly acknowledged that the operational management and water pressure influence significantly the frequency of pipe breaks. Pipe breaks occur when the water pressure exceeds the tensile strength of a pipe, or due to repetitive pressure cycles that result in fatigue-related failures. Considering these pipe failure modes, a new metric known as cumulative pressure-induced stress has been introduced. This metric takes into account both static and dynamic pressure components that contribute to pipe breaks, including mean pressure and the magnitude and frequency of pressure fluctuations, respectively. The impact of CPIS on pipe breaks has not been extensively investigated. Consequently, this study investigates and evaluates the impact of this metric when incorporated as an explanatory variable in Random Forest (RF) models that analyse the key causes of pipe breaks in two WSNs. Different RF models were developed both with and without incorporating pressure components. Subsequently, the performance of these models and the significance of each input variable were assessed. The results of this study suggest that CPIS is an important variable, especially in cases where pressure-related factors play a significant role in pipe breaks. Consequently, incorporating CPIS has shown a notable improvement in the accuracy of pipe break models.

Keywords Random forests · Cumulative pressure-induced stress · Mean pressure · Dynamic pressure · Water supply networks · Asset management

✉ Carlos Jara-Arriagada
cij18@imperial.ac.uk

Charalampos Konstantinou
konstantinou.charalampos@ucy.ac.cy

Ivan Stoianov
ivan.stoianov@imperial.ac.uk

[1] Department of Civil and Environmental Engineering, University of Cyprus, 1 Panepistimiou Avenue, Nicosia 1678, Cyprus

[2] Department of Civil and Environmental Engineering, Imperial College London, Exhibition Road, London SW7 2AZ, United Kingdom

1 Introduction

Examining the causes of pipe breaks is a significant area of research in the field of water supply network (WSN) analysis. This interest in understanding the causes of pipe breaks is primarily driven by the significant financial and social costs associated with these breaks (Folkman 2018). Water pipes are typically designed to withstand predetermined internal and external loadings throughout their lifespan. However, various factors beyond conventional design considerations can contribute to pipe breaks, and are the topic of ongoing research. These factors may include operational conditions, environmental influences, and intrinsic properties of the pipes (Konstantinou and Stoianov 2020; Pouri and Heidarimozaffar 2022).

In the water industry, it is widely recognized that operational conditions, particularly water pressure, have an impact on pipe breaks. However, past research has not been conclusive to assess the extent of their impact (Barton et al. 2019). This lack of understanding can be attributed in part to the limited operational data recorded by water utilities (Barton et al. 2019, 2020). Additionally, to comprehensively understand the effects of water pressure and its dynamic components, it is essential to capture operational data at a high temporal resolution and with sufficient spatial coverage. This practice is not widely adopted among water utilities, but it is gradually evolving due to the increasing availability of battery-powered, high-resolution pressure monitoring devices. As a result, most prior studies have relied on data from low-resolution hydraulic models, limited assessments of average water pressure, or alternative indicators such as pipe pressure ratings for their analysis (Winkler et al. 2018; Fan et al. 2022; Moslehi and Jalili_Ghazizadeh 2020; Martínez García et al. 2020).

A few studies have linked dynamic water pressure conditions (i.e. pressure transients) with pressure-induced fatigue failures in water supply pipes (Jara-Arriagada and Stoianov 2023; Jiang et al. 2019; Xing and Sela 2019; Huang et al. 2020; Lee et al. 2023). Fatigue refers to the degradation of a material's structural resistance due to repetitive cyclic loadings. With regards to pipes, this degradation occurs as cracks propagate within the material with each cycle of loading. The rate at which these cracks propagate is significantly influenced by the magnitude of pressure fluctuations and mean pressure (Jara-Arriagada and Stoianov 2023).

Acknowledging the fluctuating dynamics of water pressure within WSNs, Hoskins and Stoianov (2017) introduced a novel metric termed cumulative pressure-induced stress (CPIS). This metric incorporates components such as the frequency, magnitude and mean pressure of water pressure fluctuations over time. The CPIS metric is designed to account for the aggregate stress exerted by these pressure components, providing a way to assess the potential for pipe fatigue and consequently, the likelihood of pipe breaks. It offers a comprehensive view of the fatigue risks in WSNs using a singular metric (Hoskins and Stoianov 2017). The metric has been introduced in statistical models by Rezaei (2017) showing that there was a positive but weak correlation between pipe breaks and the CPIS. A potential reason for these past research findings might be the use of models that did not adequately address the complexity and non-linear nature of the problem, coupled with the short duration of monitoring periods for pressure fluctuations. While there is an observable association between CPIS and pipe breaks, further research is required to establish a more conclusive and statistically significant relationship.

Building upon the insights from the studies of Hoskins and Stoianov (2017) and Rezaei (2017), the main contribution of this paper is the application of a machine learning framework designed to systematically analyze the link between CPIS and pipe breaks.

Acknowledging the role of water pressure in these breaks, we hypothesize that the relevance of CPIS increases as pressure-related factors, such as average and fluctuating water pressures play a more significant role in causing pipe breaks, regardless of the specific characteristics of pipes and networks. Under these circumstances, CPIS might effectively replace other pressure indicators. This hypothesis is a central focus of our research. Additionally, the paper introduces further novelties and contributions. It applies bespoke Random Forest models using two extensive datasets from different WSNs. This method offers improved flexibility in capturing complex, non-linear processes, a step beyond traditional methods for evaluating CPIS. Moreover, Random Forest modeling allows for the determination of crucial importance metrics, enhancing our ability to understand the relevance and interactions of variables within these models.

The study utilized one dataset (NetA) specifically designed to evaluate the ability of CPIS to capture information from other pressure-related metrics. Subsequently, a second, larger dataset (NetB) was employed to assess the impact of CPIS on model performance. These analyses provide valuable insights into the role of water pressure fluctuations in pipe breaks and highlight the necessity for developing performance metrics to monitor and evaluate water pressure dynamics. Fundamentally, this study adds to the understanding of the relationship between CPIS and pipe breaks, thereby providing water management professionals with critical information and knowledge to justify continuous monitoring and assessment of water pressure fluctuations.

2 Cumulative Pressure-Induced Stress

The cumulative pressure-induced stress (CPIS) has been proposed as a comprehensive metric for assessing the impact of both gradual and sudden pressure variations on pipe deterioration (Hoskins and Stoianov 2017). The main drive for developing the CPIS metric arises from the insufficient attention given to monitoring pressure transients, and particularly transients with medium to small amplitude. Although high amplitude cyclic loadings can result in quick crack growth and failure, low amplitude cyclic loadings may lead to crack initiation and slow propagation over time. Therefore, both low and high amplitude cycle loadings are important factors to consider when evaluating the fatigue behavior of pipes in relation to pipe breaks.

Transient events of different amplitudes in water distribution networks are shown in Fig. 1. The acquired data indicates that pressure fluctuations of large amplitude are rare compared to the frequent occurrences of medium to small amplitude events in water distribution systems. The cumulative impact of these medium to small events may result in the initiation and growth of cracks if sufficient cyclic loadings accumulate over time. As outlined by Hoskins and Stoianov (2017), the CPIS can be defined as:

$$CPIS = f\,(\text{mean pressure, amplitude of pressure cycles, cycle counts}) \tag{1}$$

Equation (1) defines the functional relationship between CPIS and three critical factors: mean pressure, amplitude of pressure cycles, and cycle counts. It offers a quantitative method to assess and measure the cumulative stress on pipes resulting from the combined impact of these factors. The functional form of the CPIS was further detailed by Rezaei (2017), who presented CPIS as a metric composed of two variables as follows:

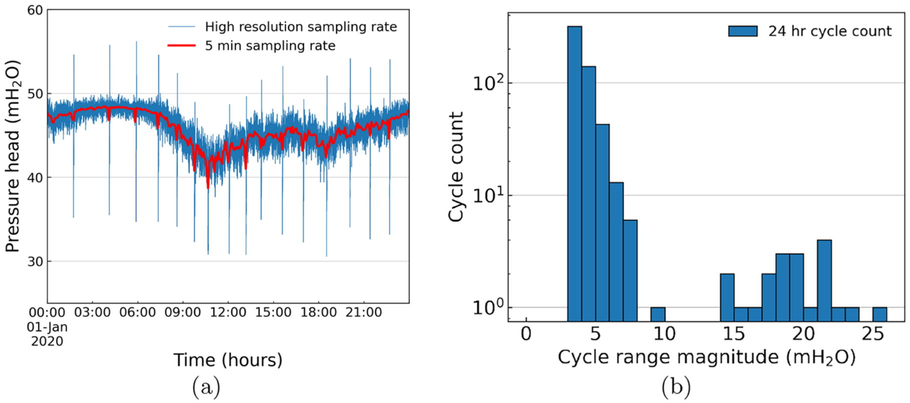


Fig. 1 **a** Comparison between 5 min sampling rate and high-resolution pressure monitoring in a 24hr sampling period. **b** Histogram of cycle counts with a cycle range threshold of 3 mH_2O

$$CPIS = f(P_{mean}, DP) \tag{2}$$

where, P_{mean} is the diurnal mean pressure in the pipe and DP, dynamic pressure, is a proposed metric that combines the amplitude of pressure cycles and cycle counts as follows:

$$DP = \sum_{i=1}^{M}(\sigma_i \cdot n_i) \tag{3}$$

where M is the number of different cyclic loading events, σ_i is the amplitude of pressure variations i, and n_i is the number of pressure variations with amplitude i.

The specific functional form of the CPIS (Eq. 2) is assumed to be implicitly learned when the variables comprising CPIS are decoupled and included individually within a statistical or machine learning model. The development of the DP metric is in line with the principle of linear damage accumulation, following Miner's rule, which is the industry standard for fatigue failure analysis of components under cyclic loadings.

High-resolution monitoring devices are essential for acquiring pressure data from multiple pipes within a network to guarantee extensive spatial and temporal coverage. Water pressure should be monitored at sampling rates over 100 samples per second to aid in identifying the sources of pressure transients. However, monitoring at high frequencies may also capture a substantial amount of noise signals, which can be removed by applying a hysteresis filter. After this filtering, the rainflow cycle counting algorithm is used to produce a structured array of the cycle amplitudes and event occurrence frequencies. Dynamic Pressure (DP) and mean pressure are then calculated at specific locations.

Graph theory is employed to identify the shortest hydraulic paths that pressure waves follow within a network. Then, an algorithm is applied to spatially extrapolate CPIS from the pipes where measurements were taken to the rest of the pipes in the network. This algorithm is based on pre-existing knowledge about potential sources of pressure fluctuations. More information on this implementation can be found in Hoskins and Stoianov (2017) and Rezaei (2017).

3 Methodology

Machine learning models, specifically Random Forest models, are applied on two comprehensive datasets to evaluate the power of CPIS to infer pipe breaks. An overview of the proposed methodology is presented in Fig. 2.

3.1 Data Curation and Exploratory Analysis

Two datasets, NetA and NetB, of historic pipe breaks and associated pipe properties have been provided by two water utilities. The data collected include environmental, operational and pipe intrinsic factors. Details of these factors are shown in Fig. 2. NetA is characterized by its detailed and good quality data, and the primary focus for its use was to compare the explanatory power of CPIS against other pressure metrics. NetA was also used to assess the impact of the cycle amplitude threshold (minimum cycle amplitude loading) in the explanatory power of the CPIS. The insights gained from NetA were then utilized to collect more targeted information from NetB. Furthermore, NetB is notably larger in scale compared to NetA. This approach allows for a comprehensive analysis that benefits from the strengths of both datasets.

High-resolution pressure time series were collected from high-frequency pressure monitoring devices over a period of 2-3 weeks for each network. The water pressure variables used in the analysis were obtained from measurements recorded by InflowSense™ pressure monitoring devices. These devices were strategically positioned at chosen locations within

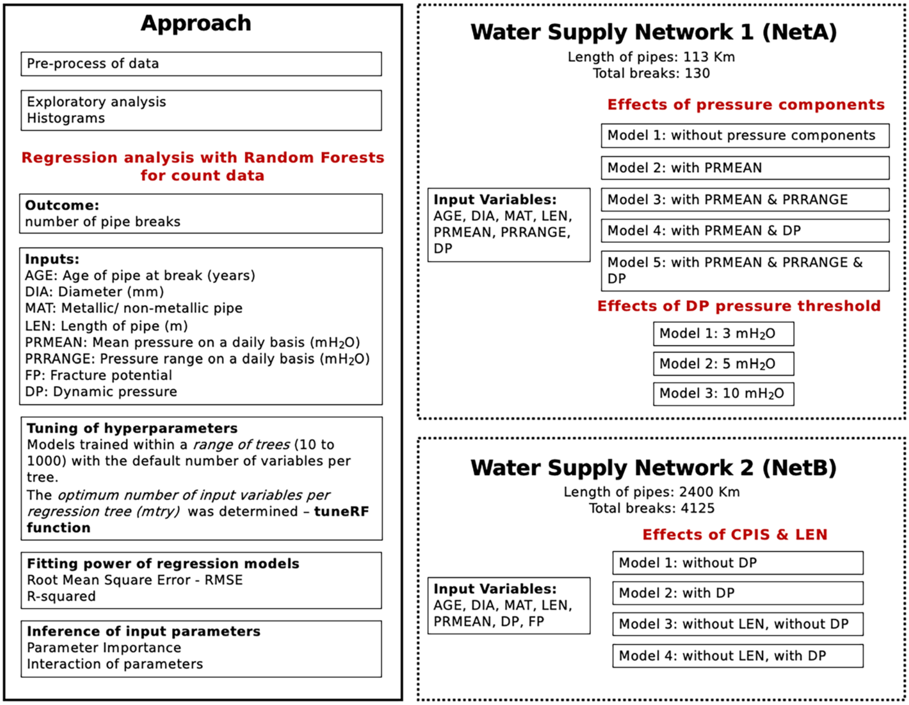

Approach
Pre-process of data
Exploratory analysis
Histograms
Regression analysis with Random Forests for count data
Outcome:
number of pipe breaks
Inputs:
AGE: Age of pipe at break (years)
DIA: Diameter (mm)
MAT: Metallic/ non-metallic pipe
LEN: Length of pipe (m)
PRMEAN: Mean pressure on a daily basis (mH_2O)
PRRANGE: Pressure range on a daily basis (mH_2O)
FP: Fracture potential
DP: Dynamic pressure
Tuning of hyperparameters
Models trained within a *range of trees* (10 to 1000) with the default number of variables per tree.
The *optimum number of input variables per regression tree (mtry)* was determined – **tuneRF function**
Fitting power of regression models
Root Mean Square Error - RMSE
R-squared
Inference of input parameters
Parameter Importance
Interaction of parameters
Water Supply Network 1 (NetA)
Length of pipes: 113 Km
Total breaks: 130
Effects of pressure components
Input Variables:
AGE, DIA, MAT, LEN, PRMEAN, PRRANGE, DP
Model 1: without pressure components
Model 2: with PRMEAN
Model 3: with PRMEAN & PRRANGE
Model 4: with PRMEAN & DP
Model 5: with PRMEAN & PRRANGE & DP
Effects of DP pressure threshold
Model 1: 3 mH_2O
Model 2: 5 mH_2O
Model 3: 10 mH_2O
Water Supply Network 2 (NetB)
Length of pipes: 2400 Km
Total breaks: 4125
Effects of CPIS & LEN
Input Variables:
AGE, DIA, MAT, LEN, PRMEAN, DP, FP
Model 1: without DP
Model 2: with DP
Model 3: without LEN, without DP
Model 4: without LEN, with DP

Fig. 2 The approach followed in the two WDSs

the networks to ensure adequate spatial coverage. The mean pressure data was utilized to calibrate hydraulic models, enabling the estimation of mean pressure values in all pipes throughout the networks. As for the dynamic pressure metric, an energy dissipation linear model was utilized for its extrapolation throughout the networks (Hoskins and Stoianov 2017; Rezaei 2017). Efforts were made to ensure a homogenized distribution of pressure monitoring devices throughout the networks. This distribution helps to provide a more representative and balanced estimation of dynamic pressure across the networks.

3.2 Random Forest Regression & Variable Performance

The Random Forest (RF) algorithm (Liaw and Wiener 2002), adjusted for count data, was chosen for this study based on the work done by Konstantinou and Stoianov (2020). In their study, the dependent variable of the models was the number of breaks per pipe segment. The study suggested that the algorithm offered superior fitting capabilities compared to other machine learning algorithms. In addition, the RF algorithm has been used for a variety of water resources problems such as the prediction of coastal flooding and rainfall prediction (Sadler et al. 2018; Faramarzzadeh et al. 2023).

A forest consists of multiple decision trees and each tree is built with a random subset of covariates (Liu et al. 2010). The conventional regression tree tends to overfit the input data set (Sadler et al. 2018), while RF overcomes this problem via its randomness feature. According to the algorithm developers (Breiman 2001; Breiman and Cutler 2004), there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error since it is estimated internally, during the run with the use of the out-of-bag error (OOB). The OOB error is a unique feature of RF that uses the samples that were not used in training a particular tree in the forest to measure the error (roughly one-third of the cases) and has been shown to be impartial in numerous testing.

RF has two main hyper-parameters (see Fig. 2), the number of trees to grow in the forest (range of trees), and the optimum number of predictor variables randomly sampled as candidates per regression tree ('mtry'). A sensitivity study was conducted to select the appropriate hyper-parameters. The models were first trained within a range of trees (10 to 1000) with the default number of variables per tree (one-third of the variables, or max number of variables). Then, the optimum number of predictor variables per regression tree was evaluated using the 'tuneRF' function in the 'randomForest' R package (Liaw and Wiener 2002). The function develops RF models using a range of 'mtry' parameters, one at a time, and determines the appropriate value by minimizing the the OBB error within the Random Forest. A 5-fold cross validation was also conducted for hyper-parameter tuning to confirm the values of mtry and ntree, yielding the same results.

The models were fitted using the entire dataset with the objective of conducting a causality analysis on the historical data. The goal was to identify cause-and-effect relationships between the input variables, assess the strength of these relationships, and explore interactions among different variables. The emphasis was on learning from the current dataset rather than making predictions. The use of OOB, enables the model to be trained, fitted, and validated simultaneously, while mitigating the risk of overfitting. This approach helps ensuring the reliability and generalization of the model's performance (Hastie et al. 2009).

The count models' performance was assessed using the R-squared and root mean squared error (RMSE) metrics, derived from the comparison of predicted and actual break values. These metrics are commonly employed to evaluate the accuracy of statistical regression models, as shown in Fig. 2. Following the model development, the contribution of each feature to the model was calculated by determining its contribution in each tree of the model. This calculation involved measuring the total decrease in node impurities resulting from splitting on the variable averaged over all trees (Gini index). The interactions between covariates were also of interest especially for the pressure components and were also evaluated. The interactions between two variables are defined via the prediction function decomposition and are measured by Friedman's H-statistic (Molnar 2018).

4 Case Study - Description of the Networks

Two datasets, water supply network 1 (NetA) and water supply network 2 (NetB), were prepared and utilized for the analysis. Both datasets have the total number of breaks experienced by a pipe as the dependent variable. Apart from the water pressure variables, which were derived from high-frequency pressure monitoring, all other variables were obtained from the records maintained by the water utility. The investigations conducted on each dataset are depicted in Fig. 2. NetA served as a 'training' case study, where a thorough analysis was conducted to offer insights and recommendations on defining the CPIS and its relationship with variables representing pressure components. Different thresholds of event amplitudes for the determination of DP have been also used to assess their effects on the regression fit performance. Following the conclusions drawn from the analysis on NetA, NetB was utilized as a 'validation' case study.

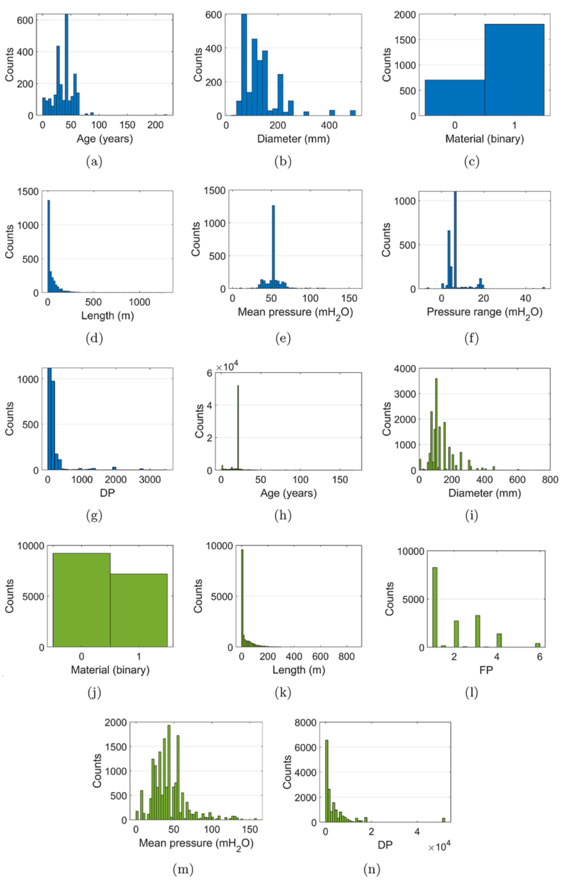
4.1 NetA Description

The total length of pipes is about 113 Km. A total of 2501 entries was obtained with 100 entries representing pipe segments with breaks (4% of all observations - each record might include more than one break). The information obtained regarding the breaks contains the number of breaks for each segment. The pipe age falls in the range between 0 and 100 years. The pipe segments are divided such that most of the lengths are about 50 m (see Fig. 3(a)). In this network, most of the material is metallic accounting for approximately 70% of the total pipes and the non-metallic material is mainly plastic. Both the mean pressure and the pressure variations have very similar values in all pipe segments and are not distributed evenly. The mean pressure is around 50 mH_2O while the pressure range is around 5 to 10 mH_2O. The DP values are concentrated around 500 (Fig. 3).

4.2 NetB Description

The second network consists of a larger sample containing approximately 79750 observations with 977 pipe break records (each record might include more than one break). The breaks represent only 1.2% of the total documented observations. The total pipe length for this set is 2400 km. Most of the pipe segments have length of 0-50 m and the average diameter is 80-120 mm (Fig. 3). Most of the pipe material in this area is asbestos cement, however, the metallic and non-metallic pipes are almost equal. The histograms of the pressure components show that the mean pressure follows again a normal distribution and the CPIS has significantly larger values compared to NetA.

Fig. 3 Classification of network characteristics for **a**-**g** NetA and **h**-**n** NetB



5 Results and Discussion

5.1 NetA

In NetA, the significance of CPIS is assessed after determining the role of the pressure components. Part of the analysis is the parameters' importance and the interaction terms strength which is calculated with the Friedman's H-statistic (Molnar 2018) including the pipe breaks based on the developed regression fit.

5.1.1 Investigation of Pressure Components Effects

First, the overall role of the pressure components is assessed by fitting RFs with and without pressure components. Once the level of influence is determined, the PRRANGE and DP effects are compared. These are 'similar' metrics, however, DP carries more information (the event cycles). Since the sample size is smaller in this network compared to NetB, the number of breaks is also smaller. All the pipes experience 0 to 4 breaks with one pipe segment only experiencing 11 breaks. The outlier of 11 breaks was not removed from the sample as the Random Forest approach appropriately deals with extreme values. Models were developed for the following combinations of pressure components:

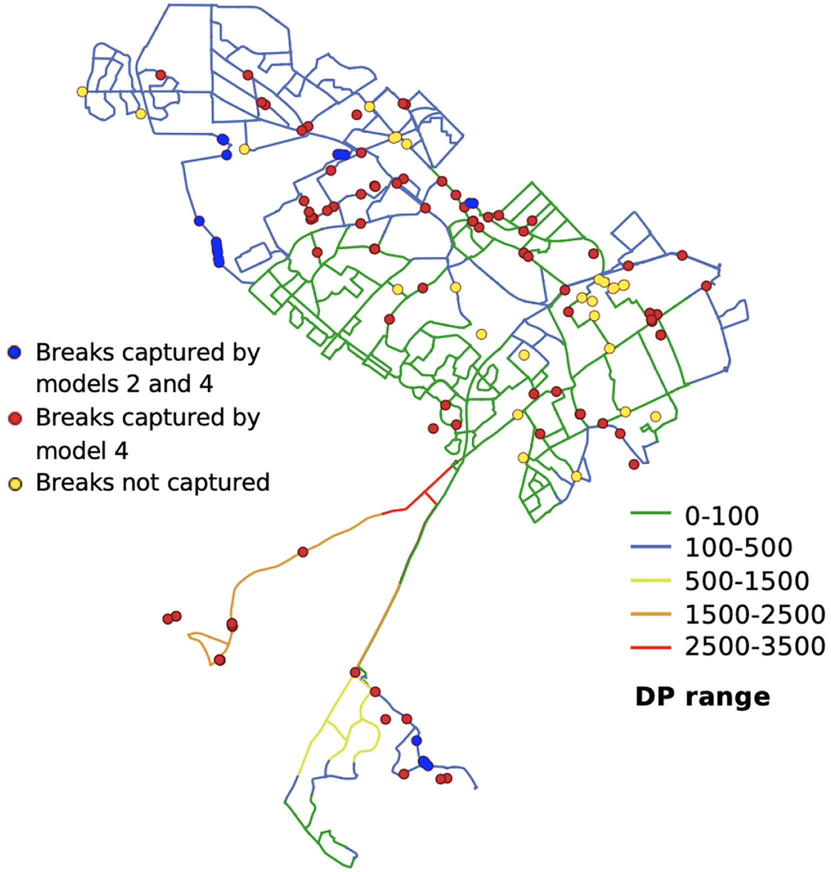
- Model 1: without pressure components
- Model 2: with PRMEAN
- Model 3: with PRMEAN and PRRANGE
- Model 4: with PRMEAN and DP (CPIS)
- Model 5: with PRMEAN, PRRANGE, and DP

Models 1 and 2 are compared to demonstrate the impact of different pressure metrics in the network. Models 3 and 4 are of particular interest as DP effectively replaces PRRANGE and includes additional information such as event cycles associated with amplitudes. Model 5, on the other hand, is not a realistic scenario as DP includes the information present in PRRANGE. However, it is included to determine if any information loss occurs when replacing PRRANGE with DP.

The models were assessed based on the R-squared and RMSE values from the plots of predicted breaks versus actual breaks. In the model without pressure components the R-squared value was 0.56 (RMSE of 0.2805) and became 0.62 when PRMEAN was included (RMSE of 0.2708). Therefore, the loading components of this network are important as the model becomes significantly more accurate once the first pressure component of mean pressure is added. This is the first operational factor that is included in the regression analysis beyond the environmental and network characteristics factors. The combination of PRMEAN and PRRANGE increases even more the accuracy of the fitted models to reach an R-squared value of 0.76. Equivalently, the root mean squared error decreases from 0.2708 to 0.2092 once PRRANGE is added in the model as a covariate (22.7% percent decrease in error). However, PRMEAN and DP (CPIS) provide by far the best fits (R-squared of 0.81 and RMSE of 0.2002); 30.5% increase in accuracy compared to Model 2 and 6.6% increase in accuracy compared to Model 4 (comparison in terms of R-squared values). The R-squared and RMSE values of all three components (model 5) are the same as in the former case (0.8113 and 0.2011, respectively) and demonstrates that there is no loss of information when removing PRRANGE. In fact, DP is a stronger pressure component as model 5 performs slightly worse compared to model 4.

Figure 4a shows a map illustrating the range of DP values within the network. The breaks have been marked also for different cases. The blue dots represent the breaks captured by model 2 (without the addition of DP), the blue and red dots are the pipe breaks captured by model 4 (with the addition of DP) and the yellow dots are the breaks that model 4 failed to capture. The addition of the DP as a covariate increases dramatically the performance of the model. Figure 4b shows the confusion matrix of model 2 in which DP was not used whilst, Fig. 4c presents these findings for model 4 in which DP was added. The pipes that have experienced one or more breaks have been clustered together. It is clear from the two matrices that the addition of DP helps in improving the estimations related to the pipes that have experienced a break. The false negatives decrease from 88 to 26 and the true positives increase from 9 to 71 signifying a massive improvement in the model performance once DP is added in the model.

The parameter significance is also examined for each model (Fig. 5(a)). Generally, the length is the most important parameter followed by age and pressure components. The pipe length is not a factor itself that causes a break, but it is used as a variable in statistical models. As shown in the histogram of the pipe length in Fig. 3, very few pipe segments are large, and these experience a high break rate. It was proved by Konstantinou and Stoianov



(a)

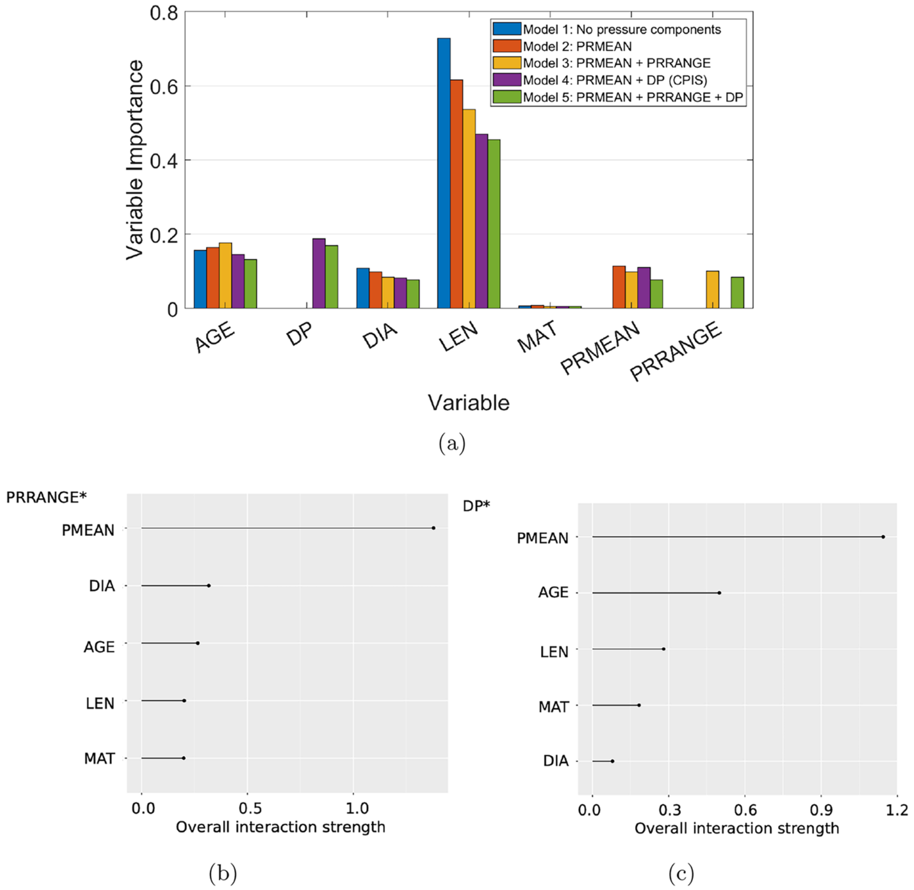
Confusion matrix- Model 2		Actual number of breaks	
		0	≥1
Estimated number of breaks	0	2402	88
	≥1	2	9

(b)

Confusion matrix- Model 4		Actual number of breaks	
		0	≥1
Estimated number of breaks	0	2400	26
	≥1	4	71

(c)

Fig. 4 **a** The range of DP values per pipe in NetA. The confusion matrix for **b** model 2 and **c** model 4



(a)

(b)

(c)

Fig. 5 **a** Variable importance for the 5 models for NetA. Overall interaction strength of **b** PRRANGE (model 3) and **c** DP (model 4) with the rest of the variables in the model

(2020) that the variable's importance relative order did not change when adding or removing the length as a variable and the results were consistent between the two cases. In other words, length does not affect the functional relationships and importance of factors affecting pipe breaks.

When the PRMEAN variable was added to the model (model 2), it emerged as the third most important covariate, following length and age. Notably, the importance of the length variable decreased significantly with the inclusion of PRMEAN. This suggests that PRMEAN contributes significantly to the model's predictive power and may have a stronger association with the dependent variable compared to the length variable. The inclusion of the PRMEAN variable likely captured relevant operational aspects that influence the occurrence of pipe breaks. Once the first pressure component was added, which is an operational factor, the significance of length decreases. When PRRANGE was added (model 3), the significance of length decreased even more, and the significance of PRRANGE was similar to PRMEAN. PRMEAN decreased as it contains part of the information of the former variable. When DP was added (model 4), it became the second most significant predictor with length's importance reducing substantially while PRMEAN remained at approximately the same levels. This was expected as amongst the three operating factors (PRMEAN, PRRANGE and DP), the DP contains more detailed information. In model 5, where both DP and PRRANGE were included, the importance of PRMEAN, PRRANGE and DP is affected the most while the rest of the variables exhibited marginal changes. This proves that there is an overlap between the information each of these covariates carries.

Finally, the strength of the interactions was also assessed for two variables (PRRANGE for model 3 and DP for model 4) with respect to the rest used in the model Fig. 5(b-c). The strongest interaction appeared to be between the pressure components in both cases proving that they have multiple and different effects on the models. DP accounts for the number and magnitude of pressure fluctuations experienced by a pipe which produce crack propagation. The mean pressure influences the rate at which crack propagation occurs. Fatigue-related failures are linked to the age, length and material type which explains the rest of the interactions. The rest of the interactions were combinations of pressure components and pipe characteristics. For model 3, the PRRANGE interacts to a lesser extent with DIA, AGE, LEN and MAT. On the other hand, for model 4, DP shows a high interaction with AGE and LEN. This explains why LEN's importance reduces from 0.75 to 0.45 (a 40% reduction) when DP is added to the model as shown in Fig. 5(a).

5.1.2 The Effects of the DP Predictors Calculated at Different Thresholds

The DP metric was calculated in three different ways for NetA. The first calculation was for cycles of amplitude at least 5 mH_2O (DP) which was shown before, the second for cycles of amplitude of at least 3 mH_2O (DP_3), and the third for cycles of amplitude of at least 10 mH_2O (DP_10). The calculated metrics show that as the threshold decreases, more information is retained in the DP metric. More of the short events amplitude events which are more frequent are captured. The threshold of 10 m excludes a vast amount of information.

The parameter importance analysis showed that as the DP becomes more detailed, its significance increases (approximately 16% for DP_10 and 18% for DP_5 and DP_3). However, the accuracy of the models remains approximately at the same levels for 3 mH_2O and 5 mH_2O indicating that lower thresholds below 5mH_2O do not add significant additional information. For 10 mH_2O, the accuracy of the model starts to decrease because it captures only the high amplitude events which are related to a lesser extent to the fatigue-related failures as these are usually high frequency events with low amplitudes. Therefore, it is not suggested to deviate much from the lower range. Between the three values, a threshold value of 5 mH_2O is reasonable to be chosen.

5.2 Development of Models for NetB

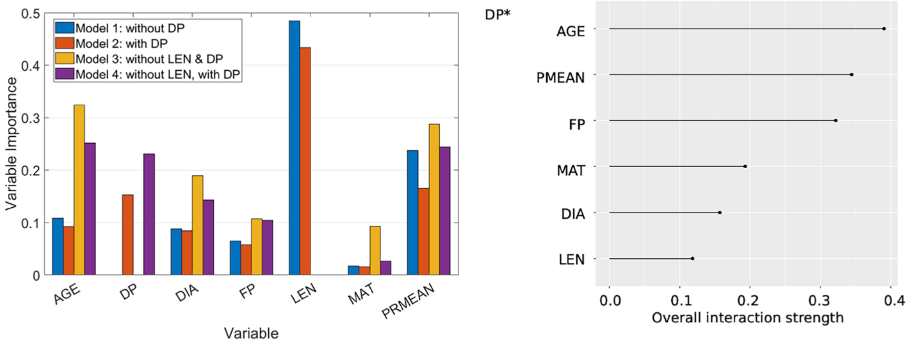
The information obtained from the analysis of NetA is used to validate the hypothesis that CPIS is an important covariate in another WSN with completely different characteristics. The focus in NetB is shifting to identifying the performance of models developed in explaining pipe breaks with and without the CPIS metric (with and without the addition of DP and mean pressure). This is also the ultimate recommendation of this study, the assessment of whether an operator needs to shift their attention to the dynamic components of the network and make plans on how to proceed with regulating the mean pressure and dynamic pressure.

The NetB dataset is much larger in comparison to NetA and is used in order to examine the sample size effects and the role of CPIS in a completely different network with very different environmental and network characteristics. Regression models were fitted with and without DP to assess the role of DP on this network, while the variable LEN has been also examined (models were developed with and without LEN) since it was shown previously to be an important parameter. Models were developed for the following cases:

- Model 1: without DP
- Model 2: with DP
- Model 3: without LEN and DP
- Model 4: without LEN, with DP

The performance of the model increased quite substantially once the DP variable was added as the R-squared value increased from 0.68 to 0.70 (3% increase) and RMSE decrease from 0.2144 to 0.2088 (2.6% decrease in error). The model developed without the addition DP shows that the most important covariate was LEN followed by PRMEAN. The results are almost identical with the case of NetA, which builds confidence about the results. When DP was added it becomes the third most important covariate with an importance percentage close to that of PRMEAN which was the second most important covariate. LEN's and PRMEAN's importance decreased (Fig. 6a).

Very similar outcomes were derived when developing models 3 and 4 (without adding LEN as a covariate). When LEN was not included as a variable DP, PRMEAN and AGE became the three most important covariates. All input covariates' importance increased proportionally without their relative ranking changing when LEN was removed from the input variables. The interaction strength of DP with respect to the other variables (Fig. 6b) demonstrates strong interactions between DP and age, DP and PRMEAN, and DP and FP. This comes in contrast to the results of NetA, however, this system is exposed to more aggressive environments with higher fracture potential (FP) when the levels of pipe age and internal pressure components are higher. This finding is in agreement with the fact that DP values in NetB were considerably higher compared to NetA. Figure 6b also shows the complexity level of this system which is much higher compared to NetA.



(a) (b)

Fig. 6 **a** Variable importance for NetB, **b** Interaction strength between DP and the rest of the variables

5.3 Assessment of the Impact of the Cumulative Pressure-Induced Stress as a Metric in Pipe Break Models

In both networks, there was no single dominant factor that was solely responsible for pipe breaks. However, among the three groups of factors (operational, environmental, and network characteristics) contributing to pipe breaks, the loading components group emerged as the most important. The size of the dataset did not have any impact on the model outcomes. In fact, the importance of the CPIS in both NetA and NetB remained relatively similar despite differences in sample size and network characteristics. This suggests that the predictive power of the model for CPIS is robust and consistent across datasets with varying sizes and network characteristics.

Based on the analysis, when loading factors such as mean pressure, maximum pressure, and pressure variations are found to be influential, the CPIS also becomes important. Correlation tests indicate that DP is not highly correlated with the other variables, suggesting that it captures effects that are not accounted for by any other covariates. As a result, DP effectively replaces all other pressure components except mean pressure. Consequently, the CPIS, defined by PRMEAN and DP, serves as a comprehensive metric for capturing these effects.

The relationship between pipe breaks and the various input variables is complex and highly non-linear. Consequently, the proposed framework, which utilizes machine learning techniques to capture these relationships, provides a fast and efficient method to guide industry operators in making informed decisions regarding their pressure management strategies. The framework's output is a recommendation regarding the criticality of PRMEAN and DP variables. This recommendation can help industry operators prioritize and focus their efforts on managing and controlling these variables effectively. By understanding the criticality of these variables, operators can implement tailored pressure management schemes to reduce the likelihood of pipe breaks and optimize the overall performance and resilience of the WSNs.

Random Forests are a powerful tool for pattern recognition, though, the algorithm's predictions are limited to the range of the input dataset. This limitation can pose challenges when encountering scenarios where a static pressure component in the future exceeds the historical mean pressure values. Fortunately, this issue is not a significant concern for CPIS since it is primarily related to fatigue-related failures driven by dynamic pressures. In such cases, the number of cycles experienced by a pipe increases with both the age of the pipe and the dynamic pressure. Therefore, the CPIS metric, which considers both the mean pressure and the dynamic pressure, can provide meaningful insights and predictions for fatigue-related failures, even in scenarios where future static pressure components exceed historical mean pressure values.

Despite its limitations, this research highlights the importance of pressure management control in WSNs. It represents a crucial step towards conducting field trials or experiments. It is recommended to closely monitor a cluster of pipes within a population that exhibits characteristics representative of the entire network. By continuously and rigorously monitoring pressures at high frequency rates, various features such as real-time fault or anomaly detection can be detected. Early warnings can also be provided during the initial stages of an event. In the wear-out phase of the material, the frequency and amplitude of the events become prominent factors contributing to the degradation process. By understanding and monitoring these factors, proactive maintenance and management strategies can be implemented to mitigate pipe breaks and extend the lifespan of the network.

6 Conclusion

This study investigated how various hydraulic loading components affect pipe breaks, specifically concentrating on dynamic components as measured by the cumulative pressure-induced stress metric (CPIS). Random Forest regressors were used to determine the significance of CPIS as a covariate for modelling pipe breaks in two water supply networks. The results showed enhanced performance across all models when dynamic pressure (DP) was incorporated, indicated by a reduction in RMSE. DP more efficiently represents the dynamic effects compared to the pressure range variable (PRRANGE), which merely suggests the possible spectrum of pressures. Furthermore, DP not only supplants PRRANGE as a covariate but also contributes extra information to the models. Therefore, the CPIS metric (PRMEAN and DP) was found to be a significant factor responsible for pipe breaks.

While the importance of adaptive pressure management is widely recognized in the water industry, the relationship between input variables and pipe breaks is notably non-linear, adding complexity to any proposed solution. Consequently, our framework provides an effective method for aiding industry professionals in making informed decisions about pressure management tactics. The outcomes of the framework include guidance on the relevance of PRMEAN and DP variables. This study demonstrates the significance of ongoing monitoring and systematic gathering of high-resolution pressure data as integral elements of a comprehensive asset management strategy for water supply networks. Future research should focus on extended monitoring campaigns, combined with proactive operational measures to reduce pressure variations. It should also incorporate a thorough examination of the condition of pipes based on information gathered during pipe break repairs. Additionally, future studies ought to conduct surveys with water management experts, utilizing questionnaires or focus groups, to evaluate their acceptance of this metric and gather their feedback on the concept.

Author Contributions Charalampos Konstantinou: Conceptualization, Methodology, Investigation, Formal analysis, Writing - Original Draft. Carlos Jara-Arriagada: Investigation, Formal analysis, Writing - Original Draft. Ivan Stoianov: Conceptualization, Formal analysis, Writing - Review & Editing, Funding acquisition, Supervision.

Funding This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) EP/P004229/1 (Dynamically Adaptive and Resilient Water Supply Networks for a Sustainable Future) and; the National Agency for Research and Development (ANID)/ Scholarship Program/ DOCTORADO BECAS CHILE/2020-72210314.

Data Availability The authors do not have permission to share data.

Declarations

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Barton NA, Farewell TS, Hallett SH (2020) Using generalized additive models to investigate the environmental effects on pipe failure in clean water networks. NPJ Clean Water 3(1):20–22. https://doi.org/10.1038/s41545-020-0077-3

Barton NA, Farewell TS, Hallett SH, Acland TF (2019) Improving pipe failure predictions: Factors effecting pipe failure in drinking water networks. Water Res 164:114926. https://doi.org/10.1016/j.watres.2019.114926

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Breiman L, Cutler A (2004) The OOB error estimate

Fan X, Wang X, Zhang X, Yu PAXB (2022) Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors. Reliab Eng Syst Safety 219:108185. https://doi.org/10.1016/j.ress.2021.108185

Faramarzzadeh M, Ehsani MR, Akbari M, Rahimi R, Moghaddam M, Behrangi A, Klöve B, Haghighi AT, Oussalah M (2023) Application of machine learning and remote sensing for gap-filling daily precipitation data of a sparsely gauged basin in East Africa. Environ Process 10. https://doi.org/10.1007/s40710-023-00625-y

Folkman S (2018) Water main break rates in the USA and Canada: a comprehensive study. Technical Report, Utah State University. https://digitalcommons.usu.edu/mae_facpub/174

Hastie T, Tibshirani R, Friedman J (2009) Elements of statistical learning, 2nd ed. https://doi.org/10.1007/978-0-387-84858-7

Hoskins A, Stoianov I (2017) Monitoring fluid dynamics. https://patents.google.com/patent/WO2017060737A1/en

Huang Y, Zheng F, Duan HF, Zhang Q (2020) Multi-objective optimal design of water distribution networks accounting for transient impacts. Water Resour Manage 34:1517–1534. https://doi.org/10.1007/s11269-020-02517-4

Jara-Arriagada C, Stoianov I (2023) Pressure-induced fatigue failures in cast iron water supply pipes. Eng Fail Anal 107731. https://doi.org/10.1016/j.engfailanal.2023.107731, https://www.sciencedirect.com/science/article/pii/S1350630723006854

Jiang R, Rathnayaka S, Shannon B, Zhao XL, Ji J, Kodikara J (2019) Analysis of failure initiation in corroded cast iron pipes under cyclic loading due to formation of through-wall cracks. Eng Fail Anal 103:238–248. https://doi.org/10.1016/j.engfailanal.2019.04.031

Konstantinou C, Stoianov I (2020) A comparative study of statistical and machine learning methods to infer causes of pipe breaks in water supply networks. Urban Water J 17(6):534–548. https://doi.org/10.1080/1573062X.2020.1800758

Lee JS, Zeng W, Lambert M, Hilditch T, Gong J (2023) Fatigue analysis of metallic-plastic-metallic pipeline systems: a numerical study. Results Eng 17:100986. https://doi.org/10.1016/j.rineng.2023.100986

Liaw A, Wiener M (2002) Classification and Regression by randomForest. Newsletter R Project News 2(3):18–22

Liu Z, Sadiq R, Najjaran H (2010) Exploring the relationship between soil properties and deterioration of metallic pipes using predictive data mining methods. J Comput Civ Eng 24(3):289–301. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000032

Martínez García D, Lee J, Keck J, Kooy J, Yang P, Wilfley B (2020) Pressure-based analysis of water main failures in California. J Water Resour Plan Manag 146(9):05020016. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001255

Molnar C (2018) IML: an R package for interpretable machine learning. J Open Source Softw 32(26):786. https://doi.org/10.21105/joss.00786

Moslehi I, Jalili_Ghazizadeh M, (2020) Pressure-pipe breaks relationship in water distribution networks: a statistical analysis. Water Resour Manage 34(9):2851–2868. https://doi.org/10.1007/s11269-020-02587-4

Pouri Z, Heidarimozaffar M (2022) Spatial analysis and failure management in water distribution networks using fuzzy inference system. Water Resour Manage 36:1783–1797. https://doi.org/10.1007/s11269-022-03104-5

Rezaei H (2017) Impact of pressure fluctuations on pipe failures in water distribution networks. PhD thesis, Imperial College London. https://doi.org/10.25560/73983

Sadler JM, Goodall JL, Morsy MM, Spencer K (2018) Modeling urban coastal flood severity from crowd-sourced flood reports using poisson regression and random forest. J Hydrol 559:43–55. https://doi.org/10.1016/j.jhydrol.2018.01.044

Winkler D, Haltmeier M, Kleidorfer M, Rauch W, Tscheikner-gratl F (2018) Pipe failure modelling for water distribution networks using boosted decision trees. Struct Infrastruct Eng 14(10):1402–1411. https://doi.org/10.1080/15732479.2018.1443145

Xing L, Sela L (2019) Unsteady pressure patterns discovery from high-frequency sensing in water distribution systems. Water Res 158:291–300. https://doi.org/10.1016/j.watres.2019.03.051

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.