Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Pipe Break Rate Assessment While Considering Physical and Operational Factors: A Methodology Based on Global Positioning System and Data Driven Techniques

Yaser Amiri-Ardakani
Graduate University of Advanced Technology

Mohammad Najafzadeh (Moha.najafzadeh@gmail.com)
Graduate University of Advanced Technology https://orcid.org/0000-0002-4100-9699

Research Article

Keywords: Break rate, Reliability evaluation, Water distribution network, Field Investigation, Artificial intelligence approaches, Available predictive techniques

Posted Date: June 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-377852/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Water Resources Management on July 23rd, 2021. See the published version at https://doi.org/10.1007/s11269-021-02911-6.

Pipe Break Rate Assessment While Considering Physical and Operational Factors: A Methodology based on Global Positioning System and Data Driven Techniques

Yaser Amiri-Ardakani[1] and Mohammad Najafzadeh[2*]

[1]Master of Science, Department of Water Engineering, Faculty of Civil and Surveying Engineering, Graduate University of Advanced Technology, P.O.Box 7631885356, Kerman, Iran. Email: yaseramiri431@gmail.com

[2*]Associate Professor, Department of Water Engineering, Faculty of Civil and Surveying Engineering, Graduate University of Advanced Technology, P.O.Box 7631885356, Kerman, Iran. Email:moha.najafzadeh@gmail.com; m.najafzadeh@kgut.ac.ir (Corresponding Author)

Abstract

Deterioration of urban Water Distribution Networks (WDNs) is one of the primary cases of water supply losses, leading to the huge expenditures on the replacement and rehabilitation of elements WDNs. An accurate prediction of pipes failure rate play a substantial role in the management of WDNs. In this study, a field study was conducted to register pipes break and relevant causes in the WDN of Yazd City, Iran. In this way, 851 water pipes were incepted and localized by the Global Positioning System (GPS) apparatus. Then, 1033 failure cases were reported in the eight zones of under study WDN during March-December 2014. Pipes break rate (BR_P) was calculated using the depth of pipe installation (h_P), number of failure (N_P), pressure of water pipes in operation (P), and age of pipe (A_P). After completing a pipe break database, robust Artificial Intelligence models, namely Multivariate Adaptive Regression Spline (MARS), Gene-Expression Programming (GEP), and M5 Model Tree were employed to extract precise formulation for the pipes break rate estimation. Results of the proposed relationships demonstrated that MARS model with Coefficient of Correlation (R) of 0.981 and Root Mean Square Error (RMSE) of 0.544 provided more satisfying efficiency than M5 model (R=0.888 and RMSE=1.096). Furthermore, statistical results indicated that MARS and GEP models had comparatively at the same accuracy level. Explicit

equations by AI models were satisfactorily comparable with those obtained by literature review in terms of various conditions: physical, operational, and environmental factors and complexity of AI models. Through a probabilistic framework for the pipes break rate, the results of first-order reliability analysis that MARS technique had highly satisfying performance when MARS-extracted-equation was assigned as a limit state function.

Keywords: Break rate; Reliability evaluation; Water distribution network; Field Investigation; Artificial intelligence approaches; Available predictive techniques

Introduction

The deterioration of pipes causing to pipe failures and leaks in urban Water Distribution Networks (WDN) has become the cornerstone of water utilities throughout the world. Pipe failures and leaks occasionally take place on account of reduction in the water-transmitting capacity associated with the pipes and water pollution in the WDNs. Water utilities are generally exposed to large amount of costs for the replacement and rehabilitation of water mains and consequently this issue makes it critical to assess the current and forthcoming conditions of the WDN for maintenance decision-making (*e.g.*, Berardi et al., 2008; Clair and Sinha, 2012; Shi et al., 2013; Stephers et al., 2020). In the case of economic and social views, expenses on the pipe breaks have significantly upward trend and as a results utility managers are put under pressure to operate annually replacement plans for deteriorated pipes which balance investment with expected benefits in a risk management issues. In this way, there is a ferocious demand for obtaining reliable pipe breaks techniques for evaluating performance of WDNs (Berardi et al., 2008).

There are a variety of deterioration-predicting-techniques to predict the condition and performance of water pipes which are classified into the four groups: deterministic, statistical, probabilistic, and soft computing techniques. Generally, there is no denying the fact that influential factors and output results related to the pipe deterioration techniques are significantly inextricably bound up with the applicability of these methodologies (Clair and Sinha, 2012). In the case of deterministic technique, most of the available predictive models were implemented by using regression analysis, composed of mechanistic-empirical analysis. The major drawback associated with deterministic models is that, in the case of practical use, these models are largely limited to a particular case study. Statistical model is intrinsically restricted while using assets with an insufficient previously-recorded information about WDN. As a major demerit, the regression-based techniques are not appropriate for modeling the real-world deterioration process of pipe infrastructure due to different restrictions of the sampling data (e.g., pipe structure, loading conditions, and environmental variables). In the probabilistic modeling, relative frequency of an event is considered which is applicable to determine the failure of a segment of WDN. Probabilistic approaches estimate a distribution and range of dependent parameters and additionally results of probabilistic techniques provide priority process repair, rehabilitation, and replacement related to infrastructures. Ultimately, soft computing techniques especially a variety of Artificial Intelligence (AI) models, were generally flexible to combine with Evolutionary Algorithms (EAs) for various purposes such as optimizing time scheduling, costs on rehabilitation, replacement, repair, pressure fluctuations, and break rates (Clark et al.,1982; Kettler and Goulter, 1985; Goulter et al., 1993; Silinis and Franks, 2007; Berardi et al., 2008; Clair and Sinha, 2012; Tang et al., 2019; Robles-Velasco et al., 2020).

A survey on literature review proved that soft computing models had prosperous performance for evaluation of pipes break rates because such techniques could detect well complicate patterns of pipe break processes. Application of AI models have two main merits over traditional techniques for pipes break rate assessment: (*i*) resulting in non-linear explicit equations when AI techniques are fed by a high volume of data series, and (*ii*) selecting independently effective factors (e.g., diameter of pipe, age of pipe, material type, and loading conditions) for the evaluation of pipe failure. Then, previous investigations demonstrated that there requires a large amount of efforts to provide linear and nonlinear regression equations based on AI techniques to yield more precise break rate of pipes and well-guarded design of WDNs. In this study, three robust AI models, introduced as Gene-expression Programming (GEP), Model Tree (MT), Multivariate Adaptive Regression Spline (MARS), are employed to model pipe break rates associated with urban WDN of Yazd city, Iran. In this way, map of WDN is processed in Geographical Information System (GIS) environment and additionally localization of pipes is surveyed using Global Positioning System (GPS) in order to complete break rates databases for the case study through a comprehensive field investigation. Ultimately, results of AI models are statistically quantified and additionally reliability evaluation is studied.

Literature Review: Overview of Existing Techniques

Since 1980, a board range of attempts have been made to study the break rate under various factors such as age of pipe, various materials of pipe, size of pipe, number of failure, loading conditions, and underground environment (e.g., Shi et al., 2013; Francis et al., 2014; Rezaei et al., 2015; Stephers et al., 2020). Additionally, Jowitt and Xu (1993) proposed an applicable technique of evaluating the effect of different pipe breaks conditions on WDNs. For a specific underground

environment, initial attempts demonstrated that smaller-diameter pipes put more frequently in the failure state than the pipes with larger diameter (Ciottoni, 1983; Kettler and Goulter, 1985). Later, Goulter et al. (1993) proposed a regression-based equations for the break rate prediction of WDNs in terms of time and space for Winnipeg, Canada. Sinske and Zeitsman (2003) developed prosperously a Spatial Decision Support System (SDSS) in order to study analysis of pipe break vulnerability for a municipal WDS in Paarl town, South Africa. Misiunas et al. (2005) investigated pattern of sudden pipe failure by using an improved two-sided cumulative sum (CUSUM) technique as a continuous monitoring approach. They could detect failures in various sizes and opening times. In another research work, Savic et al. (2006) assessed time-dependent break rate of sewer system with the help of a robust soft computing tool, introduced as Evolutionary Polynomial Regression (EPR). They presented evolutionary algorithm-based-relationship for pipe breaks which was successful in engineering interpretation. Berardi et al. (2008) applied EPR to predict pipe deterioration along with presenting symbolic formulae for a water quality zone in UK. Ultimately, the proposed relationship was efficiently used into a Decision Support System for various pipe operations (i.e., rehabilitation and replacement planning). Dridi et al. (2009) proposed a planning methodology with the aim of obtaining the optimal replacement schedule for a WDN. Through their investigation, they considered the cost of pipe failure for repairs and hydraulic performance for minimizing of the pressure deficit. Ultimately, the proposed planning approach has been verified by using two hypothetical WDNs. They found optimal cost for repairs of WDN in the failure state. Furthermore, Yamijala et al. (2009) estimated the likelihood of pipe failures in the future and defined the most influential parameters the likelihood of pipe breaks. They applied Logistic Generalized Linear (LGL) technique in order to obtain reliable prediction of pipe failure for water utilities in pipe for the rehabilitation and maintenance processes. In addition to this, Wang

et al. (2009) employed several multivariate regression techniques to prognosticate break rate of pipes associated with large WDSs located in three municipalities (Moncton, Laval, and Quebec), Canada. From their research, they found that results of regression models helped efficiently who are expert in academic level, municipal engineers, consultants, in better understanding trends related to the break rate of water mains.

Xu et al. (2011a) applied Genetic Programming (GP) to successfully predict the break rate of WDN in Beijing, China during a 19-year period 19 (1987–2005) in comparison with statistical model. Moreover, Xu et al. (2011b) proved superiority of EPR over GP technique in the estimation of pipe failure associated with WDN case study in Xu et al.(2011b) investigations. Xu et al. (2013) investigated on the WDN of Beijing as a case study to predict the pipe breaks using recorded events during 2008-2011 years. They applied GP as a predictive model then a parsimonious strategy of pipe replacement was proposed to define the replacement time.

Moreover, Wang et al. (2013) designed a Data Mining System (DMS) which was created for a water utility in a megacity, China. They found that during the creation of the DDMS, the extremely skew distribution related to the break and non-break pipes is not considered as a restriction. Shi et al. (2013) established relationships for the break factor analysis in a case study, Hong Kong. They found that there was a strong dependency between break rate and influential factors. Toumbou et al. (2014) found that probability distribution of Weibull exponential model was efficient for simulation the time of occurrence of a pipe failure in a small city, Canada. In the case of uncertainty conceptions, Francis et al. (2014) employed Bayesian Belief Networks (BBNs) to create a knowledge-based-model for evaluation of pipe breaks in a WDN, mid-Atlantic United States. From their proposed model, it was found that more investigations are essentially required for better appreciating whether additional risk factors of pipe failure associated with age of pipe, previously-

recorded-failures, pipe material, and diameter of pipe might become the cornerstone of asset management planning.

Aydogdu and Firat (2015) employed fuzzy clustering into topology design of Least Squares Support Vector machine (LS-SVM) technique to estimate break rate in Malatya WDN during a six-year period beginning at 2006, Turkey. Ultimately, they found that results of LS-SVM was more accurate than those obtained by Feed Forward Neural Network (FFNN) and Generalized Regression Neural Network (GRNN) techniques. Ghorbanian et al. (2016) proposed a probabilistic approach to assess pipes break rates occurred in a part of the City of Hamilton WDN in Ontario, Canada. The performance of the proposed technique results demonstrated that the frequency of low-pressure occurrences is significantly marginal whereas a higher minimum pressure criterion would unavoidably augment expected pipe break rates. Wols et al. (2018) used a mechanical technique, introduced as Comsima, to compute the break rate with consideration of stresses and joint rotations in a part of WDN in the Netherlands on the basis of quite a few loading conditions on the pipe (i.e., soil, traffic, water pressure, and differential settlements). Robles-Velasco et al. (2020) applied Logistic Regression (LR) and Support Vector Classification (SVC) to evaluate the pipe breaks in a part of WDN, Spine. The results indicated that the number of unexpected breaks are likely to be largely declined. Approximately 30% of pipe break rates could have been plummeted with substituting merely three percent of the incepted pipes annually, which is a realistic alternative. Ultimately, Yazdekhasti et al. (2020) employed efficiently four Machine Learning (ML) methods, as Classification Tree (CT), Random Forest (RF), LR, and, SVM to predict the water pipe breaks in the mid-Atlantic region during 1999-2018.

Overview of Case Study

The case study of water distribution network was located in the Yazd City, central Iran. This city located from 31° 53′50″ N to 54° 22′4″ E where has 529673 populations. The first region of Yazd's WDN, which is divided into eight zones (Z1-Z8) as Azadshar, Azadegan, Esteghlal, Emamshahr, Enghelab, Shahedieh, Saber Yazdi, and Jomhouri were considered as the case study. Fig.S1 (Supplementary Materials) illustrates eight sub-regions of WDN on a GIS-derived-map. In fact, retrieving sufficient pieces of information about the length of pipe was not possible on the GIS-based-map due to the fact that occurrences of pipes break in the WDN had been recorded on the basis where failure has been observed.

On the other hand, there was no a dedicated identification code for each pipes. In this way, to complete a dataset of pipes break, a handy GPS whose software was installed on the cellphone, was employed to find unknown coordinates of in the WDN case study. In fact, GPS generally has five merits over other traditional land surveying techniques: (*i*) a higher level of accuracy than traditional land surveying methods, (*ii*) computations are performed highly quickly and with a high degree of precision, (*iii*) visibility among stations dose not impose limitation on operation of GPS technology, (*iv*) it can be conveniently carried to collect accurate data, and (*v*) quite a few GPS systems are capable of communicating wireless for delivery of real-time information.

Fig.S2 demonstrates adaptation of GPS with the cellphone. Afterward, localization of pipes which have been deteriorated, was known. The operation of detecting coordinates has been carried out within a 45-day surveying with GPS. To verify GPS-derived-coordinates, a handy GPS manufactured by Garmin International Company was used. More specifically, number of coordinates detected by GPS software on the cellphone were at same as coordinates retrieved by a handy GPS. This indicated that results of GPS software was reliable for continuing the field study.

Water pressure fluctuation is regulated by gauges installed in the WDN of Yazd city, in which water pressure associated with the first region of WDN varies 1.5-2 bar. Moreover, materials of water pipes were made of Asbestos (As) (with nominal diameters: 80,100, and 150 *mm*), Cast Iron (CI) (with nominal diameters: 63, 90, and 110 *mm*), and Polyethylene (Po) (with nominal diameters: 63, 90, and 110 *mm*) which various nominal diameters were dedicated to each typical material of pipes. In the recent years, 4000 pipes breaks of Yazd's WDN have been accumulated from Yazd Water and Wastewater Company (YWWC) which almost 2000 events were considered for analyzing the break rates of water pipe. Generally, in the case study of WDN, 675 pipes with 92.33*km*-long which were damaged, were incepted to detect causes of failure and number of failure. By observing the pipes damaged in the WDN, 827 break cases have been recorded during a-year period whose allowable pressure for operation of the case study WDN were 1.5, 1.7, 1.8, and 2 *bar*. Age of pipes for which were in the damage statute, were computed between 1 and 10 years (*yr*). Moreover, depth of pipes installation were 1, 1.1, and 1.2 *m*. According to the breakage reports of YWWC, there are a board range of factors (i.e., leak, poor materials, excavation operation, pressure fluctuation, settlement, rusty water pipe, heavy vehicles, and inappropriate operation of water pipe) which lead to the pipe failure of Yazd's WDN. In the present field investigation, typical breakages of water pipes were illustrated in Fig.S3.

Effective Factor on the Pipe Break Rate

According to the previous experimental and field investigations, effective factors which play a key role in the pipe failure processes, are generally categorized into three sections: physical, operational, and environmental factors. In fact, physical factors are related to the properties of water pipe: age, material, length, thickness, and depth of pipe installation. The operational factor

is associated with hydraulic conditions of water pipe: number of failure, mean pressure, water velocity, failure type, and pressure fluctuation or time series since last breakage. The last factor is due to environment such as traffic, soil type, soil corrosion, temperature, rainfall, soil resistivity, soil shrinkage swell, and freezing index (*e.g.*, Savic et al., 2006; Berardi et al., 2008; Xu et al., 2011*a*&*b*; Yazdekhasti et al., 2020; Robles-Velasco et al., 2020).

In this study, selection of contributory variables depends on the data availability of WDNs and environmental conditions of case study. All environmental factors, as mentioned above, were excluded from list of variables affecting the failure process due to the fact that the last records of breakage in the Yazd's WDN were not for the reason of traffic flows, chemical properties of soil, and freezing. In accordance with reliable breakage records, physical and operational factors can be considered into account for the evaluation of failure process. Due to limit accessibility of WDN information, length of pipe (L_P), number of breaks (N_B), depth of pipe installation (h_P), allowable pressure in the operation status (P_A), and age of pipe (A_P) were selected as effective variables on the pipe break rate (BR_P). In this study, the values of BR_P associated with inception operation were computed as,

$$BR_P = \frac{N_B}{L_P \times h_P \times A_P \times P_A} \quad (1)$$

For the case study of WDN, BR_P values vary between 0.024 and 77.70. More specifically, information about number of failure occurrence, date of event, type of pipe, and nominal pipe diameter were provided in Table S1 (Supplementary Materials) for the case study of WDN during March 2014-December 2014. To develop soft computing models for prediction of BR_P values, 675 records were considered in which 80 percent of observations were dedicated to the training and additionally the rest of observations were assigned for testing performance. Furthermore, to

develop the practicability of the AI models, 176 observations related to the Jomhouri zone (Z8) are used.

Data-Driven Models

Implementation of Model Tree

The Model Tree, as a one of the widely practicable Data-Mining Techniques, has the high potential of approximating the performance of function (Quinlan, 1992). Development of MT are completed within two steps in order to build tree-like configuration. The first step is dedicated to dividing search space of input variables into sub-divisions. In the second phase, the trees are created by means of datasets trapped in each sub division and a linear regression model is thereafter fitted on each subdivision datasets. In fact, entire search spaces related to the linear models are expressed as a set of If-Then rules. Tree includes a collection of leaves and nodes. Each leaf generally describes linear regression formulation while nodes are representative of If-Then rules associated with input search spaces. The M5 model is one of the most efficient kind of MT, which is used in this study. This linear model might have been simplified by removing many input variables that have marginal contribution on minimizing the value of predicted error. In M5 technique, Greedy Search (GS) is employ to eliminate excessive input vectors, in some cases, all input vectors are neglected and as a results constant values only remain for nodes (Keshtkar and Kisi, 2018).

In this study, pipe length, pipe diameter, and the age of pipe were assigned as input variables and then pipe break rate is expressed by M5 as,

$$\mathrm{BR}_P = C_0 + C_1 \mathrm{h}_P + C_2 \mathrm{L}_P + C_3 \mathrm{P}_A + C_4 \mathrm{A}_P \quad (2)$$

where C_0, C_1, C_2, C_3 and C_4 are weighting coefficients of linear regression-based-equation by M5. The performance of M5, implemented by Weka software, provides 17 linear models along with if-

then rules. Full descriptions of M5-drived-equations were presented in the Table S2. As seen in Table S2, L_P is the first splitting input variable with value of 0.0235 *k*m. More specifically, if L_P is greater than (or equal to) 0.0235*k*m, then A_P is sole variable which play a key role in the prediction of BR_P:

$$\mathrm{BR}_P = 33.174 - 3.913\mathrm{A}_P \quad (3)$$

Table S2 indicates that both L_P and A_P have contributions on creating 16 rules and leading to 16 linear equations, while P_A with splitting value of 1.9 *bar* only produced a linear equation. In fact, it is inferred from Table S2 that P_A has the lowest effects on the prediction of BR_P compared to other input variables.

Implementation of Multivariate Adaptive Regression Spline

MARS, as a sort of nonparametric regression models, was proposed by Friedman (1991). This flexible technique can be applied to estimate continuous numeric variables. The MARS model is capable of producing flexible, accurate, and convenient regression models in order to predict output variables which are continuous and binary. As a major merit, MARS is capable of explaining the sophisticated and non-linear relationships governed among inputs-output vectors of a complicated system. The MARS fundamentally splits the datasets into different regions so that a regression technique to each sub-region is fitted (Yilmaz et al., 2018). The values of breaks among sub-regions are known as "knots", while the term "Basis Function" (BF) is applied to indicate each distinct interval of the independent variables. The basic form of a BF is expressed as, max (0,*x*-*k*) or max(0,*k*-*x*) in which *x* and *k* are the independent variables and a threshold value. The overall formulation of MARS, composed of a combination of BFs with linear trend,

$$\theta(x) = \beta_0 + \sum_{i=1}^{m} \beta_i \cdot BF_i(x) \quad (4)$$

where θ is the expected output vector, β_0 is the constant coefficient, β_i is the coefficient corresponding to the BFs, m is the number of BFs. Development of MARS technique includes forward and backward phases. During the forward stepwise approach, all possible BFs are acquired; thereafter, in the backward stepwise sense, BFs which lead to overfitting, are removed in order to improve accuracy level of the Eq.(4).

In this study, Eq.(4) is developed using forward step and 30 BFs are acquired. Thereafter, four BFs were removed through backward step due to less probability of over parameterization. MATLAB software was employed to implement the MARS model.

$$\mathrm{BR}_P = 3.140 + \sum_{i=1}^{26} \beta_i \cdot BF_i(\mathrm{h}_P, \mathrm{P}, \mathrm{A}_P) \quad (5)$$

Equations given by MARS model were presented in Table S3. As seen in Table S3, it can be said that L_P and A_P are the first influential parameters which were used to split the search space. After that, the pressure in operation is the effective variable which divides the search space into smaller space.

Implementation of Gene-Expression Programming

GEP, as a major development of GP, has been proposed on the basis of population evolutionary theorem (Ferreira, 2006). GEP applies both intrinsic properties of GA and GP in a way that the simple linear chromosomes with fixed length (genome) and parse trees (GP). In GEP, the genetic parameters that are needed to be defined are comparatively the same as ones in the GP. These parameters are in a variety range: terminal set (*i.e.*, four basic algebraic symbols), function set

(known as mathematical functions), type of fitness function, control parameters (e.g., mutation rate, number of population and generation), and termination conditions for GEP development. GEP includes a character string of a specific length. Individuals which are encoded as linear strings of fixed size are ultimately expressed as nonlinear entities of various sizes and configurations introduced as Expression Trees (ETs) (Sarıdemir, 2010). In the process of GEP development, chromosomes are created randomly with certain length for entire individuals; then, the ETs are extracted from the chromosomes and the value of fitness function is assessed for each individual. The best value of fitness function associated with individuals are assigned for carrying out the reproduction stage. The GEP development will continues with new individuals for a certain number of generations until a best solution is acquired.

The GEP technique acquired the most accurate formulation for prediction of pipe break rate. To reach this goal, genetic operations, as presented in Table S4, were acquired by means of the optimal evolution strategy. GeneXproTools5 was used to develop GEP technique. GEP-based-relationship, developed by 150 generations and three genes, was expressed as,

$$\mathrm{BR}_P = 4.281e^{\left(4.612\mathrm{L}_P(\mathrm{L}_P-\mathrm{A}_P)\right)} - 104.5tanh\left(7.922\mathrm{L}_P\sqrt{\mathrm{P}_A+\mathrm{A}_P}\right) + 14.74e^{\mathrm{L}_P(\mathrm{P}_A\times\mathrm{A}_P+0.948)(\mathrm{L}_P-7.922)} + 105.2 \quad (6)$$

Results and Discussion

Evaluation of Accuracy Levels for Data-Driven Approaches

The performance of the AI models are statistically assessed in this part. In this way, Coefficient of Correlation (CC), Root Mean Square Error (RMSE), and Developed Discrepancy Ratio (DDR) are taken into consideration:

$$R = \frac{\sum_{i=1}^{N}\left(BR_{P(Observed)}^{i}-\overline{BR_{P(Observed)}}\right)\left(BR_{P(Predicted)}^{i}-\overline{BR_{P(Predicted)}}\right)}{\sqrt{\sum_{i=1}^{N}\left(BR_{P(Observed)}^{i}-\overline{BR_{P(Observed)}}\right)^{2}}\sqrt{\sum_{i=1}^{N}\left(BR_{P(Predicted)}^{i}-\overline{BR_{P(Predicted)}}\right)^{2}}} \quad (7)$$

$$RMSE = \left[\frac{\sum_{i=1}^{NO}\left(BR_{P(Predicted)}^{i}-BR_{P(Observed)}^{i}\right)^{2}}{NO}\right]^{1/2} \quad (8)$$

$$DDR = \left[\frac{1}{NO}\sum_{i=1}^{NO}\frac{BR_{P(Predicted)}^{i}}{BR_{P(Observed)}^{i}}\right]-1 \quad (9)$$

in which *NO* is the number of observations. The *R*, as a standard benchmark for the evaluation of error values obtained by the predictive techniques, varies between -1 and 1. Values of ±1 indicate the most precise prediction while existing direct and invert relationship between observations and predicted values by the predictive models, respectively. Zero value indicates no correlation between observations and output of predictive techniques. Additionally, the highest rank of accuracy value of DDR is zero. When DDR is greater than zero, the AI models illustrates over estimation. Similarly, the predictive technique provides under estimation as DDR is lower than 1. The statistical results of the AI models associated with training and testing stages were provided

in Table 1. In the training stage, MARS developed Eq.(5) by seven zones of WDN case study with comparatively better performance (R=0.983 and RMSE=0.431) than Eq.(6) extracted from GEP model (R=0.992 and RMSE=0.490). The proposed M5 technique, introduced as 17 if-then rules, could not predict pipe break rate with permissible performance (R=0.755 and RMSE=2.665) compared to the MARS and GEP techniques. In the case of DDR values, MARS (DDR=-0.987) and M5 (DDR=-0.986) had relatively the same efficiency and additionally this superiority over the GEP (DDR=-1.010). In the testing stage, Table 1 indicates satisfying results for both GEP (R=0.971 and RMSE=0.544) and MARS (R=0.981 and RMSE=0.544) models whereas the M5 has no sufficient capability for the estimation of pipe break rate with R=0.888 and RMSE=1.0960. Additionally, DDR values obtained by the GEP (DDR=-0.951) technique proves high capability in the prediction of BR_P values in comparison with MARS (DRR=-1.022) and M5 (DDR=-1.005). Due to the negative values of DDR, all the AI model demonstrated comparatively under prediction of pipes break rates. The performance of the AI models for the prediction of pipes break rate in the training and testing stages in Figs.1-3. As seen in Fig.1a, remarkable quality of MARS performance in the training stage is inferred whereas from Fig.1b, the testing performance indicates overestimation for BR_P=7.67. Fig.2a depicts highly permissible performance for the GEP model in the training stage whereas, in Fig.2b, insignificant under predictions for approximately BR_P=3.5-7 is seen in the testing performance. Fig.3a demonstrates quality of M5 performance in the training stage. From Fig.3a, it is seen that M5 has satisfying performance for $BR_P \leq 10$. M5 illustrates over prediction for BR_P=20-30 whereas Fig.3a depicts significant underestimation only for BR_P=20. What is more, Fig.3b indicates relatively significant under prediction for BR_P=22.3 in the testing stage. Overall, in the terms of a sound comparison, M5 could not provide satisfying performance for a large amount of BR_P (77.7) in the training stage whereas MARS and GEP models

has significant efficiency in this case. In this way, this factor leads to the decreasing the accuracy level of M5 performance in the training phase. Additionally, variations of DDR values versus data samples for the AI models were drawn in Figs.S4-S6. As shown in Fig.S4, almost DDR values acquired by MARS technique [Eq.(5)] were concentrated between 0 and 0.5. This means that MARS model provided relative over predictions for the pipes break rate. As depicted in Fig.S5, almost DDR values obtained by GEP model [Eq.(6)] are around zero level for both training and testing phases. Ultimately, Fig.S6 illustrated that almost DDR values computed by M5 varied between -0.5 and 0. This is inferred that M5 generally under predicted the pipe failure rate.

Generally, explicit equations given by MARS and GEP techniques were statistically more reliable than M5 [Eq.(3)]. Compared to multilinear regression equations by M5, Eqs.(5&6) included more complicate mathematical structure which were obtained by using the physical and operational factors affecting the pipe failure processes. In fact, Eqs.(5&6) with high degree of non-linearity could better understand the complexity of pipes failure rate processes rather than M5 with linear mathematical structure. On the other hand, mathematical structures of Eqs.(5&6) were found to be consistent when compared with available AI models-based-equations from literature. In this research, an explicit equation [Eq.(5)] obtained by MARS model included a set of second order polynomial functions for the pipe break predictions. Non-linear mathematical structure of Eq.(5) was consistent with those existing equations obtained by Xu et al. (2011a). An explicit relationship [Eq.(6)] given by GEP model included exponential inner functions as applied in the explicit equations by Savic et al. (2006), Berardi et al. (2008), Wang et al. (2009), and Xu et al. (2011b).

Comparison of the Present Study with Previous Investigations

In this section, results of AI models were compared with those presented in the literature. The first comparison is related to the Berardi et al. (2008) investigations in which an explicit equation developed by EPR model obtained R=0.926. Although they did not consider depth of pipe installation (h_P), they used both physical and operational factors. In this way, it can be said that Eqs.(5&6) were relatively more accurate than that reported by Berardi et al. (2008). Moreover, Wang et al. (2009) used logarithmic expressions to predict break rate pipe with different materials: Gray Cast Iron (R=0.83), Ductile Iron without lining (R=0.806), Ductile Iron with lining (R=0.845), Polyvinyl Chloride (R=0.888), and Hyperscon (R=0.901). In fact, they did not apply operational factors (e.g., pressure fluctuation, number of failure, depth of pipe installation) to estimate pipe break rate. By the way, equations by Wang et al. (2009) could not address well the complexity of pipe failure processes. Notably, explicit equations by AI models, in this research, are used for all typical materials of water pipes in the Yazd's WDN. AI models-based-relationships [Eqs.(5&6)] were not limited to the typical material rather than equations reported in the Wang et al. (2009) research. Additionally, the results of current study were comparable in terms of accuracy level with those obtained by Xu et al. (2011b). Regardless of operational and environmental factors, they presented explicit formulations by GP (R=0.85) and EPR (R=0.84) models in order to predict the pipe failure rate. Obviously, the non-linearity degree of Eqs.(5&6) is higher than relationships provided by Xu et al. (2011b) and additionally, this issue causes the increase of Eqs.(5&6) accuracy level in comparison with those obtained by Xu et al. (2011b). In a comprehensive comparison, Automatic Learning Bayesian Network (ALBN), designed by Tang et al. (2019), obtained 0.82 accuracy level while considering physical, operational, and environmental factors. In the present investigation, the results of AI models performance were

statistically comparable in terms of complexity of influential factor selection with those reported in Tang et al. (2019) research. Although complexity degree of ALBN model by Tang et al. (2019) (i.e., complete selection of effective factors and intrinsic properties of ALBN) is relatively higher than those obtained by MARS [Eq.(5)] and GEP [Eq.(6)] techniques, the predictions of pipe break rate in the current study were comparatively more precise than Tang et al. (2019) investigation.

Ultimately, the results of AI models-based equations are compared with soft computing models applied in the Robles-Velasco et al. (2020) investigation. Accuracy levels associated with LR and SVC models were 0.794 and 0.796, standing relatively lower precision level than Eqs.(5&6). In fact, it seems that complexity of SVC and LR was not sufficient to provide more accurate predictions of pipe break rate because Robles-Velasco et al. (2020) did not consider important operational factors such as number of pipe failure and depth of pipe installation. On the contrary, the present study included operational factors and provided more accurate predictions rather than those obtained by Robles-Velasco et al. (2020).

Generalization of AI models Results

In this section, the application of the AI models [Eqs.(5&6)] were evaluated to predict the pipes break rates in the eighth zone of Yazd's WDN. In fact, Eqs.(5&6) were fed by the database associated with the Jomhouri zone (Z8). Fig.4 depicted the graphical comparisons of the Eqs.(5&6) predictions to the reported ones. From Fig.4, MARS generally under predicts (DDR=0.0613) the pipes break rate whereas almost the BR_P values predicted by GEP model had overestimation with DDR of 1.237. According to the statistical measures, Eq.(5), obtained by the MARS technique,

has better efficiency (R= 0.878 and RMSE=0.416) in the prediction of pipe break rate in the Z8 of WDN than those yielded by the GEP model (R=0.63 and RMSE=2.436). Comparatively permissible capability of MARS model encourage practicability of MARS in acquiring preliminary pipes failure rate values in field investigations while generalizing the AI models by unseen (unreported) pipe failures datasets.

Probabilistic Analysis of Break Rate Model

The probabilistic technique is basically back on the acquiring negative value (known as violent) for the Limit State Function (*LSF*) in any system. For a given problem, *LSF* is difference between available capacity for output of system and values obtained by mathematical expressions govern on system variables. Therefore, the probability of limit state of violation is described as (Zampieri et al., 2016; Homaei and Najafzadeh, 2020),

$$\mathrm{P}_f = P[G(X) < 0] \quad (10)$$

where $G(X)$ is the Limit State Function (*LSF*). This function assumes a negative or zero value on failure and a positive value for safety of system. Reliability of system is expressed as,

$$\mathrm{R} = 1 - \mathrm{P}_f \quad (11)$$

First, LSF needs to be defined for reliability analysis. In the current study, the LSF is formulated as,

$$LSF = SF - \frac{BR_P^M}{BR_P^C} \quad (12)$$

in which BR_P^C and BR_P^M denote BR_P values related to the capacity of WDN case study and BR_P given by a mathematical expression. Additionally, *SF* is the safety level of BR_P values which can be assigned greater than 1.

In this study, the results of the AI models are used to express LSF. More specifically, the most accurate equation given by AI models was applied. Accordingly, statistical results of Table 1 indicated that MARS-derived-equation had the high potential of expressing BR_P^M. To compute BR_P^C, maximum values of BR_P observed in the seven zones of WDN case study is defined. In reliability analysis, there is a need to determine random variables such as length of pipe (L_P), depth of pipe installation (h_P), allowable pressure in the operation status (P_A), and age of pipe (A_P). Moreover, typical statistical distributions govern on random variables were determined using Kolmogorov-Smirnov (K-S) test. Results of K-S test has been given in Table S5.

According to the Eq.(5), integration of Joint Probability Density Function (JPDF), introduced as LSF, is considered as probability of failure. Values of P_f are computed by using two approximation functions as, First-Order Reliability Method (FORM) and Second-Order Reliability Method (SORM). The FORM and SORM employ approximations of the first and second orders Taylor expansion. In this study, FORM is applied to perform the reliability analysis. The results of reliability analysis have been expressed in terms of reliability index (λ) and probability of failure in Table 2. It is inferred from Table 2 that the increase in the *SF* values from 1 to 2.75 leads to increase of the reliability index values from 4.292 to 8.805. In addition, marginal P_f values, computed for the all the safety factor levels, indicates that break rate predictions by MARS technique has the most level of reliability. For *SF*=2.75, the failure probability obtains zero and the reliability index of 8.805.

Statistical Analysis of AI models

In this section, robustness and sustainability related to the performance of AI techniques were studied by using three probabilistic terms: uncertainty, reliability, and resilience. These statistical indices are useful to conceptually detect variations of break rate of pipes over the time. The first criterion, uncertainty analysis with 95% confidence level (U95) is capable of validating the results of the AI techniques for both training and testing stages. In fact, U95 restrains the uncertainty of BR_P values at a 95% confidence level. Thus, the lower the values of U95, the more precise the BR_P amount. Another robustness index, introduced as the reliability criterion, takes into account the overall consistency of the AI techniques. Reliability criterion is computed on the basis of the random error values from the measurement process. The higher is the number of occurrences for which error amount is lower than a specific value of threshold, the higher reliable the overall consistency of the AI models would be. The final criterion is the resilience analysis considering the potential of BR_P to resists stressors, adapt, and quickly recover from disruptions. Notably, a higher value of resilience measure demonstrates greater values of robustness for the results to noise. The further descriptions about these criteria can be found in literature (Rezaie et al., 2019). Table 3 shows that MARS technique represented the lowest value for U95 (0.201) in the training stage in comparison with other AI models. While GEP (U95=0.346) and M5 (U95=0.346) were at the same level of uncertainty. Additionally, estimations of BR_P given by MARS model were more reliable (0.70) and resilience (0.685), compared to estimations yielded by GEP (0.649 and 0.650) and M5 (0.741 and 0.722). Table 3 indicated that, on the other hand, M5 model was more reliable and resilience in the training stage than performance of GEP model. In the testing stage, the values of uncertainty with 95% confidence level yielded by the GEP model made comparatively more resilience predictions of BR_P (0.733) in comparison to the values provided by MARS technique

(0.667) and M5 (0.591). Furthermore, this trend was seen for values of reliability although break rate values by M5 were lower uncertain (U95=0.399) in the testing stage when compared with GEP (U95=0.412) and MARS (U95=0.465).

Conclusion

In this research, a field investigation has been carried out to report the pipes break rate in the eight zones of Yazd's WDN. First, the key causes of pipes failure were understood and then a datasets was created to formulate the pipe failure rates by MARS, GEP, and M5 MT techniques. Ultimately, a reliability analysis was set on the AI model which had the best performance in the failure rate prediction of WDN. From the current research, following conclusions were drawn:

- Field investigation demonstrated that the common causes of water pipes failures in the case study of WDN included seven reasons: excavation operation, rusty water pipe, leak, pressure fluctuation, settlement, heavy vehicles, and inappropriate operation in the field. In this way, physical characterizations of pipe and its operational factors were considered to predict the pipe failure rate.
- A reliable dataset was provided to model pipes-related-failure rate with the aid of three AI models and it was found that depth of pipe installation, age of pipe, pressure in operation, frequency of breakages were considered to compute the pipes break rate.
- Results of AI models development indicated that MARS and GEP techniques stood at the same level of precision for the WDN calibration (or training stage). M5 MT developed by 17 if-then rules could not provide an estimation of values as well as MARS and GEP models. Overall, statistical results of AI models in the testing phase indicated that

Eqs.(5&6) can be used to model the pipes break rate for the WDN zones whose information had no contribution on training and testing stages. Overall, consistency of mathematical expressions obtained by MARS and GEP models was investigated. In accordance with empirical equations in the literature, it was found that two chief factors including the complexity of AI models-based-formulations and availability of influential parameters play a key role in the improving accuracy level of AI models.

- To develop AI models performance, relationships obtained by MARS and GEP models were used to predict BR_P values for Jomhouri pilot (Z8) of Yazd's WDN. Compared to the GEP technique [Eq.(6)], results demonstrated that MARS had the permissible potential of modeling pipes failure rates when are fed by the fresh data.
- Performance of the reliability analysis indicated that LSF, composed of safety level and Eq.(5), could provide the robust reliability indices for various SF levels. In fact, reliability results along with marginal failure probability values proved the most suitably of MARS technique to generalize the pipe break rates for other zones of WDN case study.
- Eq.(6) given by GEP model provided relatively more resilience and reliable estimations of pipe break rates compared to the MARS and M5 techniques. In addition to this, values of pipe break rate by M5 resulted in lower uncertainty value in comparison to GEP and MARS models.

Ultimately, this research presented structural failure techniques for an individual pipe and the mathematical models on how is practically used in the decision-making and future planning.

Ethical Approval
All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Consent to Participate
Informed consent was obtained from all individual participants included in the study.

Consent to Publish
All the authors give the Publisher the permission of the authors to publish the research work.

Authors Contributions
Yaser Amiri Ardakani; Performing the field study and collecting the data; **Mohammad Najafzadeh;** Formal analysis and investigation, Writing - original draft preparation, Writing - review and editing, Resources, Supervision

Funding
No funds, grants, or other support was received.

Competing Interests
There is no conflict of interest.

Availability of data and materials
The data are not publicly available due to restrictions such their containing information that

Supplemental Materials

Figs.S1-S6 and Tables S1-S5 are the attachment to the manuscript.

Reference

Aydogdu,M., Firat, M. (2015). "Estimation of Failure Rate in Water Distribution Network Using Fuzzy Clustering and LS-SVM Methods." *Water Resour Manage,* 29, 1575–1590.

Berardi, L., Kapelan, Z., Giustolisi, O., and Savic, D. A. (2008). "Development of pipe deterioration models for water distribution systems using EPR." *J. Hydroinf,* 10(2), 113–126.

Clair,A.M.S., Sinha.S. (2012). "State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models!." *Urban Water Journal,* 9(2), 85-112.

Dridi, L., Mailhot, A., Parizeau, M., Villeneuve. J-P. (2009). "Multiobjective Approach for Pipe Replacement Based on Bayesian Inference of Break Model Parameter." *J. Water Resour. Plann. Manage.* 135(5), 344-354.

Francis, R.A., Guikema, S.D., Henneman, L. (2014). “Bayesian Belief Networks for predicting drinking water distribution system pipe breaks.” *Reliab. Eng. Syst. Saf.* 130, 1-11.

Ferreira, C. (2006). “Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence”. Springer.

Goulter, I., Davidson, J., and Jacobs, P. (1993). “Predicting water-Main Breakage Rates.” *J. Water Resour. Plann. Manage.,* 119(4), 419–436.

Ghorbanian, V., Guo, Y., Karney, B.(2016). “Field Data–Based Methodology for Estimating the Expected Pipe Break Rates of Water Distribution Systems.” *J. Water Resour. Plann. Manage.,* 04016040.

Homaei, F., Najafzadeh, M. A. (2020). “Reliability-based probabilistic evaluation of the wave-induced scour depth around marine structure piles.” *Ocean Eng.* 196, 106818.

Jowitt, P. W., and Xu, C. (1993). “Predicting Pipe Failure Effects in Water Distribution Networks.” *J. Water Resour. Plann. Manage.,*119(1),18-31.

Keshtkar, B., Kisi, O. (2018).“RM5Tree: Radial basis M5 model tree for accurate structural reliability analysis.” *Reliab. Eng. Syst. Saf* .180, 49-61.

Misiunas, D., Vítkovský,J., Olsson,G., Simpson, A., Lambert, M. (2005). “Pipeline Break Detection Using Pressure Transient Monitoring.” *J. Water Resour. Plann. Manage.* 131 (4), 316-325.

Quinlan, J.R. (1992). “Learning with Continuous Classes.” In Proceedings of AI’92 (Adams & Sterling, Eds), Singapore: World Scientific, pp. 343-348.

Robles-Velasco, A., Cortes, P., Munuzuri, J., Onieva, L. (2020). “Prediction of pipe failures in water supply networks using logistic regression and support vector classification.” *Reliab. Eng. Syst. Saf.*196, 106754.

Rezaeia,H., Ryanb,B., Stoianovc.I. (2015). “Pipe failure analysis and impact of dynamic hydraulic conditions in water supply networks.” 13th Computer Control for Water Industry Conference, CCWI 2015. Leicester, UK.

Rezaie-Balf, M., Fani-Nowbandegani,S., Samadi, S.Z., Fallah,H., and Alaghmand, S. (2019). “An Ensemble Decomposition-Based Artificial Intelligence Approach for Daily Streamflow Prediction.” *Water*,11, 709.

Sinske, S.A., and Zietsman, H.L. (2004). “Spatial decision support system for pipe-break susceptibility analysis of municipal water distribution systems.” *Water* SA, 30(1), 71-80.

Savic, D. A., Giustolisi, O., Berardi, L., Shepherd, W., Djordjevic, S., Saul, A. (2006). "Modelling sewers failure using evolutionary computing." *Proc. ICE, Wat. Manage.* 159 (2), 111–118.

Sarıdemir, M. (2010). "Genetic programming approach for prediction of compressive strength of concretes containing rice husk ash." *Constr. Build. Mater.* 24, 1911–1919.

Stephens, M., Gong,J., Zhang,C., Marchi,A., Dix, L., Lambert, M.F. (2020). "Leak-Before-Break Main Failure Prevention for Water Distribution Pipes Using Acoustic Smart Water Technologies: Case Study in Adelaide." *J. Water Resour. Plann. Manage.* 146(10), 05020020.

Shi,W-Z., Zhang, A-S., Ho, O-K. (2013). "Spatial analysis of water mains failure clusters and factors: a Hong Kong case study." *Annals of GIS,* 19(2), 89-97.

Silinis, P.G., Franks, S. W. (2007). "Understanding failure rates in cast iron pipes using temporal stratification." *Urban Water Journal,* 4 (1), 1 – 7.

Toumbou, B., Villeneuve, J-P., Beardsell, G., Duchesne, S. (2014). "General Model for Water-Distribution Pipe Breaks: Development, Methodology, and Application to a Small City in Quebec, Canada." *J. Pipeline Syst. Eng. Pract.* 5(1), 04013006.

Tang, K., Parsons, D. J. and Jude, S. (2019). "Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system." *Reliab. Eng. Syst. Saf.* 186, 24–36.

Wols, B., Moerman,A., Horst,P., Laarhoven, K.V.(2018). "Prediction of Pipe Failure in Drinking Water Distribution Networks by Comsima." Proceedings, 2(11), 589; https://doi.org/10.3390/proceedings2110589

Wang, Y., Zayed, T., and Moselhi, O. (2009). "Prediction models for annual break rates of water mains." *J. Perform. Constr. Facil.*, 23(1), 47–54.

Wang,R., Dong,W., Wang,Y., Tang,K., Yao,X. (2013). "Pipe Failure Prediction: A Data Mining Method." IEEE 29th International Conference on Data Engineering (ICDE), 8-12 April, Brisbane, QLD, Australia. 10.1109/ICDE.2013.6544910.

Xu, Q., Chen, Q., Li, W. (2011a). "Application of genetic programming to modeling pipe failures in water distribution systems." *J. Hydroinf.* 13 (3), 419-428.

Xu, Q., Chen, Q., Li, W., Ma, J. (2011b). "Pipe break prediction based on evolutionary data-driven methods with brief recorded data." *Reliab. Eng. Syst. Saf* 96 (8), 942-948.

Xu, Q.,Chen, Q., Ma, J., Blanckaert, K. (2013). "Optimal pipe replacement strategy based on break rate prediction through genetic programming for water distribution network." *J. Hydro-Environ Res.*7(2), 134-140.

Yilmaz, B., Egemen Aras, E., Nacar, S., Kankal,M. (2018). “Estimating suspended sediment load with multivariate adaptive regression spline, teaching-learning based optimization, and artificial bee colony models.” *Sci Total Environ*, 639, 826-840.

Yazdekhasti, S., Vladeanu, G., Daly, C. (2020). “Evaluation of Artificial Intelligence Tool Performance for Predicting Water Pipe Failures.” Pipelines Conference, August 9–12,San Antonio, Texas.

Yamijala,S., Guikema,, S.D., Brumbelow, K. (2009). “Statistical models for the analysis of water distribution system pipe break data.” *Reliab. Eng. Syst. Saf* 94, 282–293.

Wang, Y., Zayed, T., and Moselhi, O. (2009). “Prediction models for annual break rates of water mains.” *J. Perform. Constr. Facil*., 23(1), 47–54.

Zampieri, P., Zanini, M.A., Faleschini, F. (2016). “Influence of damage on the seismic failure analysis of masonry arches.” *Constr Build Mater*. 119, 343-55.

Figures

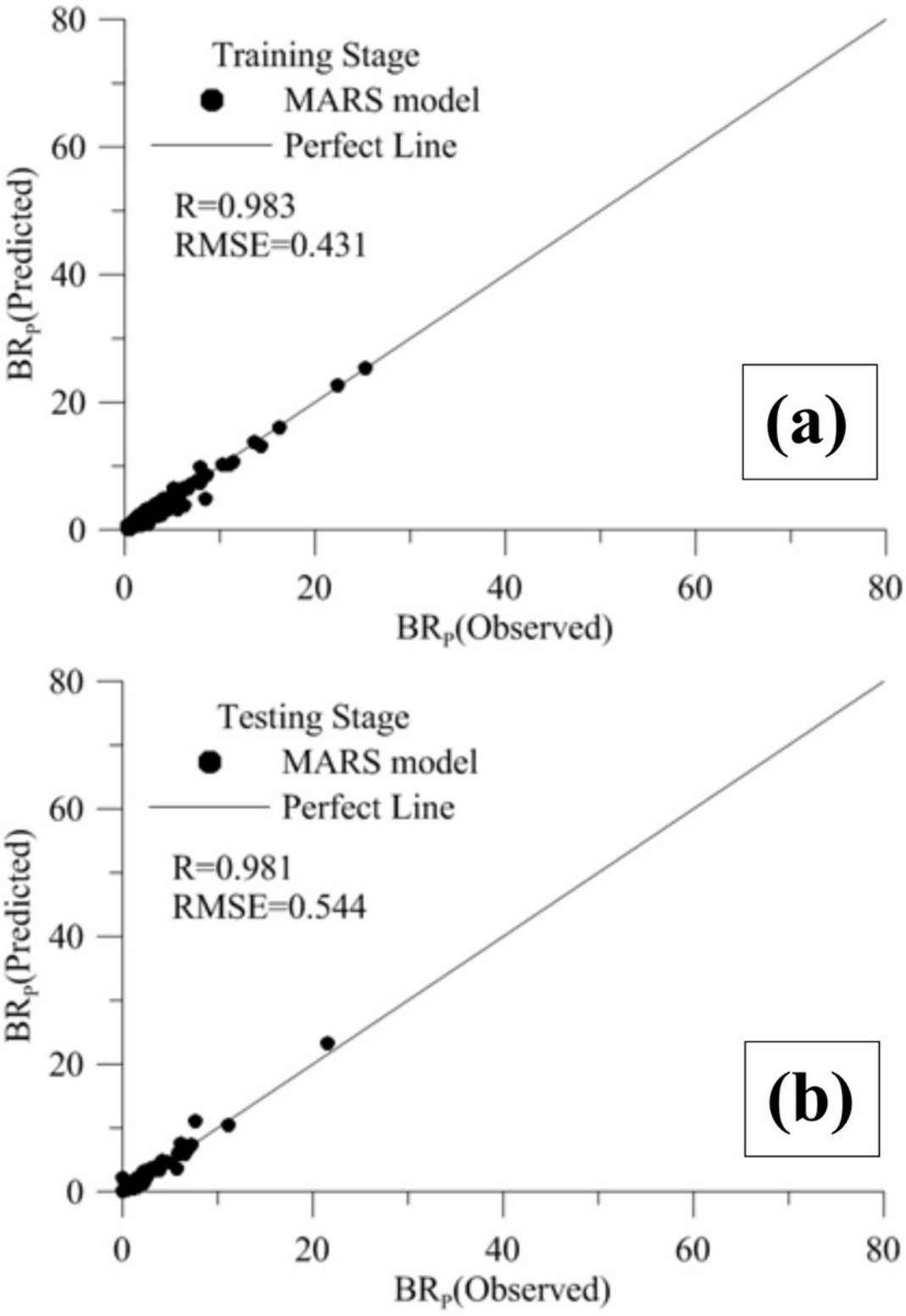

Figure 1

Qualitative performance of MARS for the two development statuses: (a) training stage and (b) testing stage

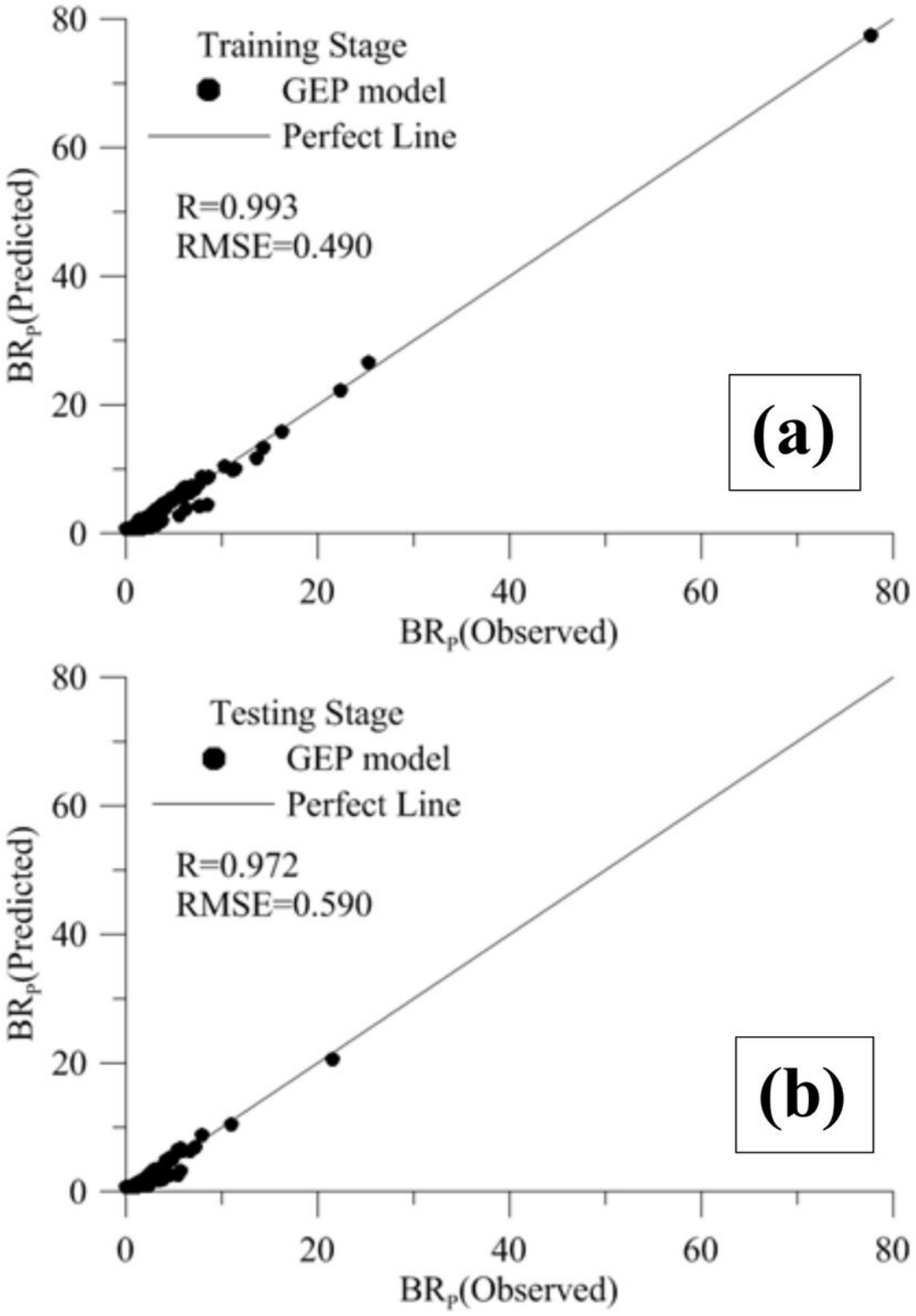

Figure 2

Qualitative performance of GEP for the two development statuses: (a) training stage and (b) testing stage

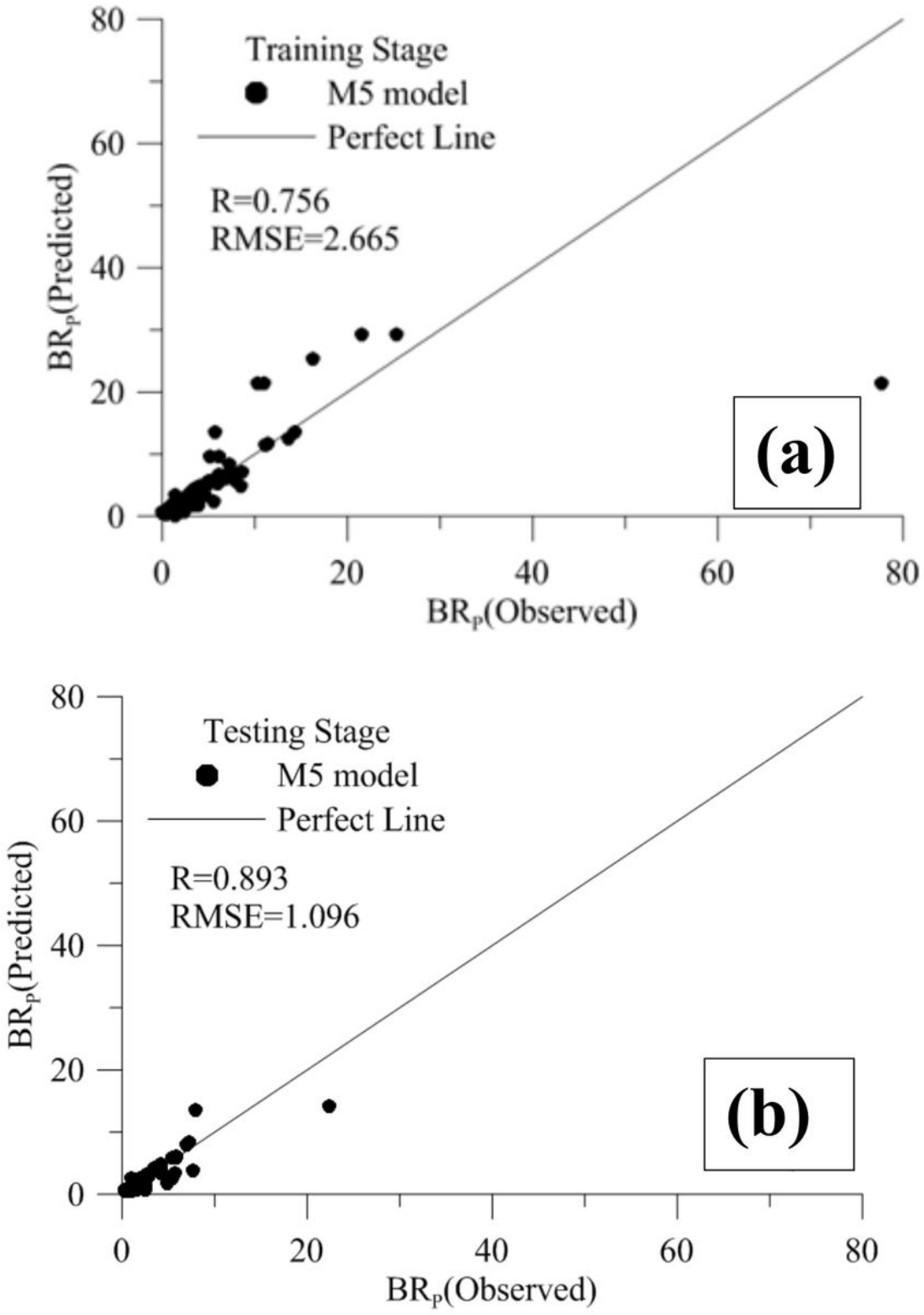

Figure 3

Qualitative performance of M5 for the two development statuses: (a) training stage and (b) testing stage

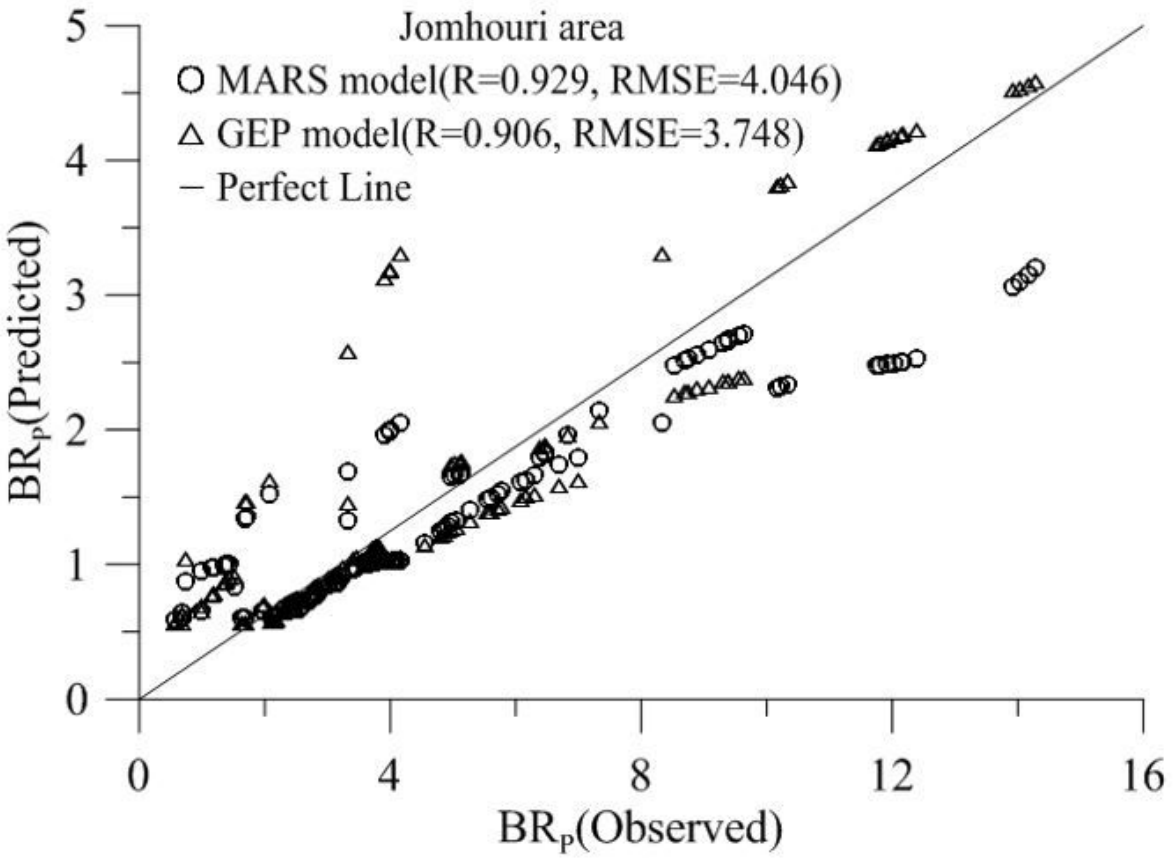

Figure 4

Qualitative performance for generalization of Eqs.(5&6)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx