

Application of black-box models based on artificial intelligence for the prediction of chlorine and TTHMs in the trunk network of Bogotá, Colombia

Laura Enriquez, Laura González and Juan G. Saldarriaga *

Water Distribution and Sewerage Systems Research Center (CIACUA), University of the Andes, Carrera 1 Este No. 19A-40, Bogotá, Colombia

*Corresponding author. E-mail: jsaldarr@uniandes.edu.co

 JGS, 0000-0003-1265-2949

ABSTRACT

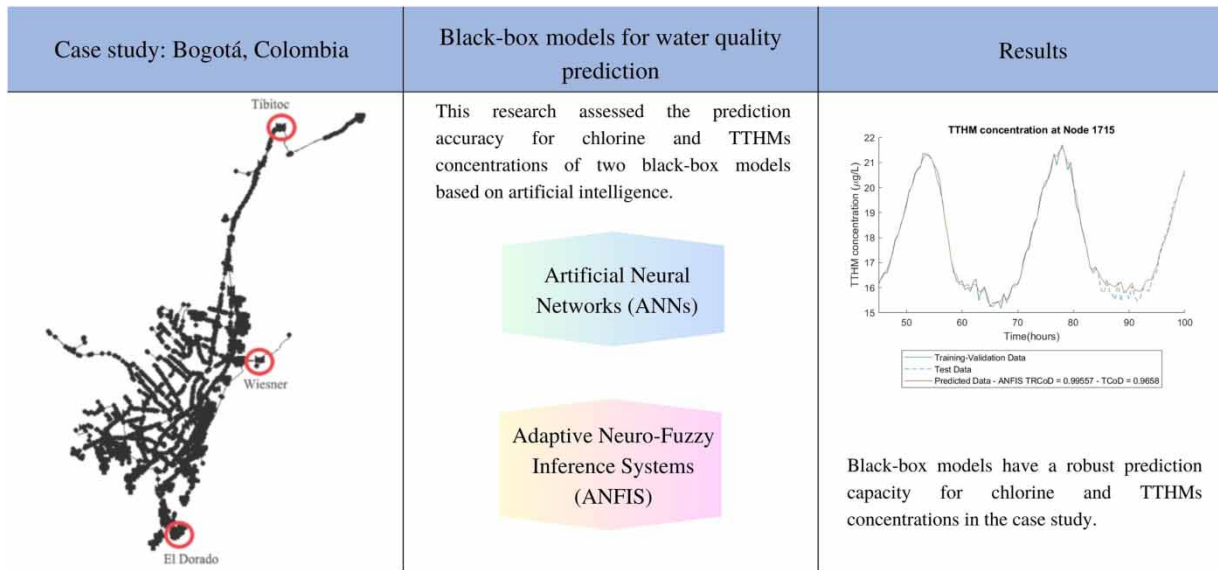
The chlorine and total trihalomethane (TTHM) concentrations are sparsely measured in the trunk network of Bogotá, Colombia, which leads to a high uncertainty level at an operational level. For this reason, this research assessed the prediction accuracy for chlorine and TTHM concentrations of two black-box models based on the following artificial intelligence techniques: artificial neural networks (ANNs) and adaptive neuro-fuzzy inference system (ANFIS) as a modelling alternative. The simulation results of a hydraulic and water quality analysis of the network in EPANET and its multi-species extension EPANET-MSX were used for training the black-box models. Subsequently, the Threat Ensemble Vulnerability Assessment-Sensor Placement Optimization Tool (TEVA-SPOT) and Evolutionary Polynomial Regression-Multi-Objective Genetic Algorithm (EPR-MOGA-XL) were jointly applied to select the most representative input variables and locations for predicting water quality at other points of the network. ANNs and ANFIS were optimized with a multi-objective approach to reach a compromise between training performance and generalization capacity. The ANFIS models had a higher mean Training and Test Nash–Sutcliffe Index (NSI) in contrast with ANNs. In general, the models had a satisfactory mean prediction performance. However, some of them did not achieve suitable Test NSI values, and the prediction accuracy for different operational statuses was limited.

Key words: ANFIS, artificial neural networks, chlorine decay, sampling design, TTHM formation, water supply systems

HIGHLIGHTS

- This is a novel application of a black-box model based on AI to calculate the water quality in water distribution systems.
- The methodology is applied to a large system serving a population of 8 million.
- Black-box models proved to be a promising approach for management applications in water utilities in terms of water quality.
- Novel use of tool linking with EPANES and EPANET-MSX.
- The methodology uses TEVA-SPOT and EPR-MOGA-XL.

GRAPHICAL ABSTRACT



1. INTRODUCTION

The most used disinfection method in water treatment plants (WTPs) is chlorination (WHO 2011). The reaction between chlorine and organic matter causes the formation of disinfection by-products, which mainly consist of trihalomethanes (THMs) that have been related to cancer formation (Mazhar *et al.* 2020). Hence, it is fundamental to monitor and control the concentrations of chlorine and THMs in order to reach a trade-off between both.

However, water quality parameters are only measured at some points of a network, and they may not be available in real-time. For this reason, the concentration of chemical species in a water distribution network (WDN) is usually modelled for operational applications. Although the preferred methods for chlorine modelling include physically based equations, an extensive exploration of data-based or black-box approaches, like artificial neural networks (ANNs), has been performed as well. Conversely, the formation of THMs has mainly been addressed through data-based models, such as partial least squares regression (PLSR), support vector machine regression (SVR), ANNs (Platikanov *et al.* 2012), and adaptive neuro-fuzzy inference system (ANFIS) (Alver *et al.* 2021).

Regarding ANNs, several authors (Rodriguez *et al.* 1997; D'Souza & Kumar 2010; Cuesta Cordoba *et al.* 2014) have found that this method can provide a chlorine prediction in different WDNs with a high level of accuracy as long as comprehensive training is performed. Furthermore, Rodriguez *et al.* (1997) demonstrated that ANNs can be used for both steady and non-steady conditions. Similarly, Milot *et al.* (2002) and Godo-Pla *et al.* (2021) concluded that ANNs are able to achieve a suitable generalization capacity when modelling THMs and, even though multilinear regressions (MLRs) have a better extrapolation ability, ANNs tend to outperform MLR. Nevertheless, the implementation of ANNs poses some challenges, such as the definition of the ANN structure and the time lags between inputs, which are usually estimated by trial and error (D'Souza & Kumar 2010; Wu *et al.* 2011).

The preferred black-box procedure to forecast water quality in WDNs is ANNs. Despite this, ANFIS has proved to be a promising option. In this sense, Hong *et al.* (2012) found that ANFIS accurately predicts chlorine decay in WTPs processes. Similarly, Alver *et al.* (2021) concluded that ANFIS can attain low errors for the prediction of THMs in real WDNs in dry and wet seasons even if the model does not use all the relevant inputs that are known to impact THMs formation. Moreover, ANFIS has been found to outperform ANNs for the prediction of coagulant dosage in WTPs (Bressane *et al.* 2023).

Even though black-box methods have demonstrated having a high forecast capacity to support the prediction of water quality in WDNs, there are still some limitations that should be addressed. First, these techniques have been restricted to small-sized networks of up to 92 nodes (D'Souza & Kumar 2010), where the parameters of the water sources have been defined as inputs. Second, the definition of some key parameters for data-based approaches has been mostly done by trial and error.

Third, despite training the models with different demand scenarios, they have not been validated under different operational plans, which is an element that should be included for a thorough analysis.

Furthermore, there are other black-box techniques based on artificial intelligence that have been used for surface and groundwater quality modelling. The prediction of water quality parameters in surface water has been addressed using support vector machines (SVMs) (Haghiabi *et al.* 2018; Deng *et al.* 2021; Ahmadianfar *et al.* 2022; AIDahoul *et al.* 2022), extreme gradient boosting (XGB), and random forest (RF) algorithms (AIDahoul *et al.* 2022), as well as genetic programming (GP), fuzzy logic (FL), and hybrid methods (Rajaei *et al.* 2020). Regarding groundwater quality, additional approaches such as multilinear regression (MLR), M5P tree (M5P), random subspace (RSS), additive regression (AR), and locally weighted linear regression (LWLR) have been employed (Kouadri *et al.* 2021). However, according to Ighalo *et al.* (2021), the most used artificial intelligence models for water quality assessment in the last decade are ANNs and ANFIS, and several authors have reported accurate results using these methods as described above.

Therefore, this research aims to evaluate ANNs and ANFIS as black-box prediction models in terms of their suitability to predict chlorine and total trihalomethane (TTHM) concentrations in the trunk network of Bogotá, Colombia. This approach extends the application of black-box methods to a bigger network, that has approximately 550 km of pipes with 4,316 junctions, and it further expands the definition of inputs to different nodes that are not labelled as sources. The key challenges of this work and the corresponding approaches are summarized as follows, while the rationale behind each method is explained in each section:

- (a) Black-box methods are proposed as a water quality modelling alternative in the trunk network of Bogotá given that sampling points are sparse.
- (b) The water quality measurements were not available, thus data simulated in the Environmental Protection Agency Network Evaluation Tool (EPANET) and its multi-species extension EPANET-MSX were used to train the black-box models.
- (c) The selection of sampling points was performed with a water security approach using the Threat Ensemble Vulnerability Assessment-Sensor Placement Optimization Tool (TEVA-SPOT).
- (d) The input selection for the black-box models was executed with the Evolutionary Polynomial Regression-Multi-Objective Genetic Algorithm (EPR-MOGA), which is a data-based method.
- (e) Unlike previous studies, the definition of the structure and the parameters of the black-box models was not performed by trial and error. The following approaches were taken instead:
 - (i) ANNs were optimized with a multi-objective genetic algorithm (MOGA) to minimize the input dimension and the number of hidden neurons, while simultaneously optimizing a fitness function.
 - (ii) The rules generation for ANFIS, that is the definition of the number of fuzzy sets and the number of rules, was optimized with a Non-dominated Sorting Genetic Algorithm (NSGA-II).
- (f) The generalization capacity of the ANNs and ANFIS models was tested with data under different operational scenarios.

2. CASE STUDY

This investigation was performed using the trunk network of Bogotá as a case study. The WDN of Bogotá serves a population of approximately 9 million inhabitants, who demand an average daily flow of 16.5 m³/s for 2019 (Empresa de Acueducto y Alcantarillado de Bogotá -E.A.A.B 2018). The main WTPs of the network are Tibitoc, Francisco Wiesner, and El Dorado, with a total treatment capacity of 26.1 m³/s (Empresa de Acueducto y Alcantarillado de Bogotá -E.A.A.B 2018). The location of the WTPs in the hydraulic model of the network, which was provided by Bogotá's water utility, is shown in Figure 1. On the one hand, the Tibitoc and El Dorado WTPs implement a conventional treatment process that includes aeration, coagulation, flocculation, sedimentation, filtration, and disinfection. On the other hand, the Wiesner WTP incorporates all these processes, except for flocculation and sedimentation since its water source contains a low sediment load. Additionally, the traditional disinfection method using chlorine gas is preferred in Wiesner and Tibitoc, whereas El Dorado WTP replaced it for the mixed-oxidant technology (MIOX[®]) since 2012 (Empresa de Acueducto y Alcantarillado de Bogotá -E.A.A.B 2018).

The chlorine and THM concentrations are sparsely measured throughout the system. The chlorine concentration is monitored at the exit of WTPs every 2 h, and at some tanks and nodes inside the system every 3–8 days. Moreover, THMs are only measured biannually in around 70 points of the network. The scarcity of water quality measurements has not allowed the calibration of physically based models. For this reason, black-box methods are proposed as a modelling alternative in this



Figure 1 | Location of WTPs in the trunk network of Bogotá.

work. In this context, the prediction of water quality in the trunk network would allow us to obtain approximate information about the status inside hydraulic DMAs.

Given that the measurements are sparse and that they were not available for this study, data simulated in EPANET and its multi-species extension EPANET-MSX were used to train the black-box models. EPANET is a public domain software that performs hydraulic and water quality analysis in pressurized pipe distribution systems. The EPANET calculation engine is the basis for other commercial software such as WaterCAD-WaterGEMS, MIKE URBAN, and Innovyze's InfoWater. Hence, given its public availability, EPANET was preferred for this research.

The hydraulic model of the trunk network includes 66 tanks; 4,362 pipes; 4,316 junctions; 73 pumps; and 145 valves; it is updated up to November 2014. The model has approximately 550 km of pipes with diameters that range from 38.1 to 3,500 mm, with an average of 689 mm, and the absolute roughness varies from 0.0015 to 0.6 mm, with an average of 0.25 mm. This is coherent with the fact that Concrete Cylinder Pipes (CCPs) are predominant, as well as the average age of the pipes, which is 36 years ([Empresa de Acueducto y Alcantarillado de Bogotá – E.A.A.B 2021](#)). However, it should be emphasized that the accuracy of the model is unknown, and its calibration is beyond the objective of this research.

Regarding other elements of the network, it has more than 20 pumping stations that are mainly located at the east, north-east, and south of the city. The types of valves that are included in the model are flow control, throttle control, pressure breaker, and pressure reducing. Furthermore, the pumps and valves are managed through rule-based controls that depend on the tanks' levels and the clock time, respectively.

Additionally, the model incorporates several 48-h demand patterns that are assigned to different junctions. The water demand corresponds to the behaviour observed in May of 2019 since data of more recent years were not available for this research. According to the water utility, the 2020 pandemic changed the distribution of the demand patterns, thus future work should validate the generalization capacity of the black-box models for these conditions. The data correspond to a low-demand scenario, which is critical in terms of water quality since water age increases when the water demand is lower.

3. METHODOLOGY

Figure 2 shows the general scheme of the methodology used in this work including the optimization methods for each black-box model. The following sections specify the details of each step.

3.1. Hydraulic modelling

The original hydraulic model, provided by the Water Supply and Sewerage Company of Bogotá in EPANET 2.2, had a total duration of 48 h, in which 22 nodes had negative pressures. The duration was increased to stabilize the concentrations of chemical species in the water quality analysis. However, when the time span was set to 100 h, 1,107 nodes had negative pressure issues using a hydraulic time step of 5 min and an accuracy of 0.001. Hence, a 100-h duration was achieved by gradually modifying five attributes of the model. First, the elevation of eight nodes was lowered to less than 0.07% of the initial value. Second, the base demand of two nodes was reduced up to 46% of the original value. Third, the loss coefficients of two valves were decreased, and the settings of two Flow Control Valves (FCVs), located upstream of a tank, were increased. Fourth, the initial level of 10 tanks was risen. Fifth, 16 rule-based controls were modified to synchronize the demand of six tanks with the behaviour of the valves that supplied them with water.

3.2. Water quality modelling

The chlorine decay was represented using the first-order models for bulk and wall reactions, as shown in Equations (1) and (2), respectively, where k_b is the first-order decay rate coefficient [1/T], C_{Cl} is the residual bulk chlorine concentration, D is the pipe diameter [L], $k_{w,1}$ is the wall decay coefficient for first-order kinetics [L/T], and k_f is the mass transfer coefficient [L/T]. In regard to TTHM formation, a linear model was adopted, which is represented by Equation (3), where TTHM is the bulk total THM concentration [$\mu\text{g/L}$] and Y is the yield parameter [μg of THMs formed/mg of Cl_2 consumed] (Boccelli *et al.* 2003). In this sense, the simplest models were chosen because the monitoring data of chlorine concentrations were not available, and the TTHM measurements were sparse.

$$\left(\frac{dC_{Cl}}{dt}\right)_{\text{bulk}} = -k_b C_{Cl} \quad (1)$$

$$\left(\frac{dC_{Cl}}{dt}\right)_{\text{wall}} = -\frac{4}{D} \frac{k_{w,1} \cdot k_f}{(k_{w,1} + k_f)} C_{Cl} \quad (2)$$

$$\frac{dC_{\text{TTHM}}}{dt} = -Y \frac{dC_{Cl}}{dt} \quad (3)$$

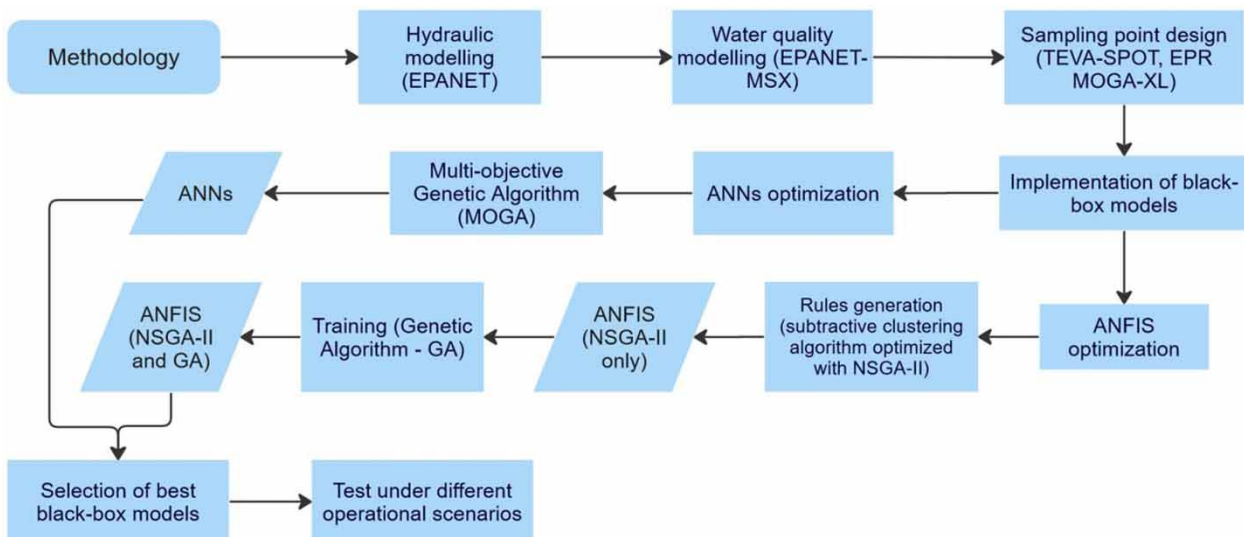


Figure 2 | General scheme of the methodology.

This research applied the MATLAB EPANET Toolkit developed by [Eliades et al. \(2016\)](#), which is an open-source software that integrates EPANET 2.2 hydraulic analysis and EPANET-MSX 1.1 water quality capabilities. The chlorine and TTHM concentrations were modelled using a multi-species model in EPANET-MSX with a 5-min time step, as shown in Equations (4) and (5), respectively. In addition, a conservative tracer with a concentration of 1.0 mg/L was assigned at Tibitoc and El Dorado WTPs to consider the different mixing proportions of water sources throughout the network. A tracing analysis in EPANET revealed that the sum of the three WTP tracers would not always sum up to one due to the storage effect of tanks. Given that the greatest contribution of water is provided by the Wiesner WTP, the proportion of water from this WTP was estimated by one minus the tracers of the other two WTPs, to ensure that the tracers would sum up to one. The concentration of each tracer in every pipe was used to calculate a weighted average value for k_b and Y on a pipe basis, which is displayed in Equations (6) and (7), as suggested by [Shang et al. \(2011\)](#):

$$\left(\frac{dC_{Cl_i}}{dt}\right)_{total} = -k_{b_i}C_{Cl_i} - \frac{4}{D_i} \frac{k_{w_i} \cdot k_{f_i}}{(k_{w_i} + k_{f_i})} C_{Cl_i} \tag{4}$$

$$\frac{dC_{THM_i}}{dt} = Y_i k_{b_i} C_{Cl_i} + \frac{4}{D_i} \frac{Y_i \cdot k_{w_i} \cdot k_{f_i}}{(k_{w_i} + k_{f_i})} C_{Cl_i} \tag{5}$$

$$k_{b_i} = TT_i k_{b_t} + (1 - TT_i - TD_i) k_{b_w} + TD_i k_{b_d} \tag{6}$$

$$Y_i = TT_i Y_t + (1 - TT_i - TD_i) Y_w + TD_i Y_d \tag{7}$$

where k_{b_i} is the first-order bulk coefficient for the pipe i ; Y_i is the THM yield coefficient for the pipe i ; TT_i and TD_i are the concentrations of the Tibitoc and El Dorado conservative tracers for the pipe i ; k_{b_t} , k_{b_w} and k_{b_d} are the first-order bulk coefficients for Tibitoc, Wiesner and El Dorado, respectively, and Y_t , Y_w , and Y_d are the THMs yield coefficients for Tibitoc, Wiesner, and El Dorado, respectively.

Given that the chlorine measurements in the network were not available, the parameters of the chlorine decay model were assumed considering that [Díaz & Saldarriaga \(2015\)](#) and [Saldarriaga et al. \(2016\)](#) have previously calibrated chlorine models in the trunk network of Bogotá using data of 2008. The bulk chlorine first-order coefficients (k_b) reported by [Díaz & Saldarriaga \(2015\)](#) and [Saldarriaga et al. \(2016\)](#) were adopted for this research, namely 0.203 days^{-1} , 0.163 days^{-1} , and 0.436 days^{-1} for the Tibitoc, Wiesner, and El Dorado WTPs, respectively. These values are in the range defined by [Nuckols et al. \(2001\)](#), and they are thus considered to be valid.

In regard to wall decay coefficients, the pipes of the network were grouped in 23 sets according to three criteria: diameter, absolute roughness and average flow velocity during the first 24 h, based on the observations of [Vasconcelos et al. \(1997\)](#), [Dini & Tabesh \(2017\)](#), and [Monteiro et al. \(2020\)](#). Subsequently, two linear functions to compute the wall decay coefficient were defined, having the mean pipe diameter of each group as an independent variable. The bounds of each function are shown in [Table 1](#), where a limit of 300 mm was set based on the range reported by [Vasconcelos et al. \(1997\)](#).

Regarding the parameters of the TTHM model, the yield coefficients (Y) for each WTP were calibrated through the Monte-Carlo Analysis Tool (MCAT) v.2 created by [Lees & Wagener \(2000\)](#). The range for this coefficient was defined from 31.9 to $50.6 \mu\text{g}/\text{mg}$ according to [Boccelli et al. \(2003\)](#). The TTHM concentrations in the network were available from half-yearly measurements performed at 59 sampling points from 2010 to the first half of 2021. The data were provided by the Water Supply and Sewerage Company of Bogotá. The objective function included three terms that computed the difference between the measured mean, standard deviation and skewness and the modelled values. Although a sensitivity analysis of the weight

Table 1 | Lower and upper bounds for calculating the first-order wall coefficient

Mean diameter (mm)	Lower bound		Upper bound	
	D (mm)	k_w (m/day)	D (mm)	k_w (m/day)
≤300	93.65	0.01 (Saldarriaga et al. 2016)	300	0.71 (Nuckols et al. 2001)
>300	300	0.71 (Nuckols et al. 2001)	3,500	1.52 (Vasconcelos et al. 1997)

of each term was performed, the tests led to the same results: 33.32, 32.12, and 36.82 $\mu\text{g}/\text{mg}$ for the Wiesner, Tibitoc, and El Dorado WTPs, respectively.

Moreover, the calibration process involved the definition of boundary conditions based on previous research (CIACUA 2009). The three reservoirs of the model were set as the sources of chlorine with a mean concentration of 1.56, 1.47, and 1.66 mg/L for Tibitoc, Wiesner, and El Dorado WTPs, respectively. A different 24-h time pattern with a 30-min time step was defined for each WTP considering the data of March, May, and June of 2008.

In addition, the initialization of the chlorine and TTHM concentrations in every node was done by setting a warm-up period based on the longest travel time of the system (Kang & Lansey 2010). For this purpose, a water age analysis was executed using EPANET 2.2, but the water age values would not stabilize when the duration was increased. Therefore, given that the duration of the simulation could not be enlarged indefinitely without raising negative pressure issues, the warm-up period was estimated by verifying the time required for the chemical species to stabilize in periodic patterns, especially at the tanks. In this sense, a 45-h warm-up period was chosen for the case study to achieve an approximate stabilization. Furthermore, the simulation was run using the final concentrations of the previous model for the current run several times until the initial and final concentrations matched.

3.3. Sampling point design

The selection of the inputs for the black-box models was divided in two stages: (a) sampling points and (b) the most representative input variables and points. The manual sampling locations for water quality parameters were not available, and the network does not have online sensors. Hence, the location of water quality sensors had to be determined. The optimization of the placement of water quality sensors has focused on getting information about the state of the water provided to consumers and security concerns (Krause *et al.* 2008; Rathi & Gupta 2014), with the objectives of minimizing the impact or the time to detection of contamination events with tools like TEVA-SPOT (Schal *et al.* 2013), or maximizing the demand coverage based on simulations with tools like Chama (Klise *et al.* 2017) or topological features (Giudicianni *et al.* 2020). The optimization algorithms implemented include genetic algorithms (GAs), mixed-integer programming, the greedy algorithm, NSGA-II, among others (Rathi & Gupta 2014). Conversely, few authors have addressed the optimization of sensor placement for the calibration of water quality models. For example, Wu & Roshani (2014) optimized sensor placement maximizing the total pipe length covered by sensors. Hence, the priority of water utilities has concentrated on water security to optimize the placement of water quality sensors. For this reason, this research adopted this approach using TEVA-SPOT, that has been validated by several applications (Hart *et al.* 2008; Storey *et al.* 2011; Khorshidi *et al.* 2018). In this way, the locations of inputs and the outputs were defined for the design of the black-box models.

Subsequently, the most significant inputs for describing the chlorine and TTHMs at the output sampling points were determined. The main approaches that have been taken for input selection in black-box methods are the Principal Component Analysis (D'ouza & Kumar 2010), the Akaike Information Criterion (AIC) (Wu *et al.* 2011), parallel factors analysis (Peleato *et al.* 2018), best subset regression analysis, and forward stepwise filtering (Kouadri *et al.* 2021), among others. The alternatives that imply the simultaneous optimization of the inputs and the training of the black-box models, such as best subset regression analysis and forward stepwise filtering, were not feasible given the high dimensionality of the problem. EPR-MOGA was adopted for this purpose since it provided two substantial advantages: (a) it allows to select the most significant inputs with a multi-objective algorithm that considers both parsimony and accuracy, and (b) it is a data-based method, that may be classified as a grey-box approach, which allows to explore the possible polynomial relationship between the input and output variables. The prediction accuracy of the EPR-MOGA technique is not presented in this work since the main focus of this research is the implementation of black-box models.

In the first place, TEVA-SPOT is a software that evaluates the effects of contamination incidents and optimizes the locations of water quality sensors; accordingly, it was developed by the EPA with collaborators from the Sandia National Laboratories, Argonne National Laboratory and the University of Cincinnati (Schal *et al.* 2013). This research adopted the settings implemented by Schal *et al.* (2013) for the sensor placement using TEVA-SPOT. The contamination scenario consisted of the injection of a chemical at a rate of 800 mg/min from 0 to 2 h at every non-zero demand node, assuming no decay. Subsequently, the sensor design was optimized using the mean time to detection with a response time of 15 min as the objective function of the optimization heuristic named Greedy Randomized Adaptive Sampling Process (GRASP), which is the only algorithm available in TEVA-SPOT-GUI (U.S EPA National Homeland Security Research Center and

Sandia National Laboratories 2015). The algorithm included sensor set sizes of 0, 20, 30, 40, 50, 60, and 70 considering non-zero demand nodes as feasible locations, which correspond to 322 nodes in the case study.

In this way, as shown in Table 2, the nodes selected in the 20-sensor design were defined as the sampling points since they would be the priority if the location was based on a restricted budget. The input variables included pressure, chlorine, and TTHMs, since these parameters are already measured by the water utility, although it is neither continuous nor in real time. Therefore, the candidate inputs for each output were equal to 60. On the other hand, the output locations were chosen from the 30-, 40- and 50-sensor designs, that have a secondary level of priority, considering two restrictions: (a) there must be at least one output node for each zone defined by the Bogotá's water utility and (b) the maximum number of output locations that can be selected is 15 nodes. Hence, given that the output variables consisted of chlorine and TTHM concentrations, the number of outputs was equal to 30.

In the second place, the most significant inputs for each output were selected with the software EPR-MOGA-XL, which finds the symbolic regression that best fits observed data using a multi-objective optimization genetic algorithm (MOGA). The objectives of the optimization process include the maximization of the model accuracy and parsimony, as well as the minimization of the number of inputs (Giustolisi & Savic 2006, 2009). The definition of the inputs with EPR-MOGA-XL was done in two steps without testing data since it caused an out of memory error due to the automatic generation of additional plots. First, the candidate inputs were divided into three groups, each of 20 variables, and all of them were tested to optimize a symbolic regression for every output variable. Second, one regression was selected for each test, which means that three regressions were chosen for every output variable, based on a low number of inputs as a priority criterion, and the fitness metrics as a secondary criterion. Subsequently, the inputs of the three selected models were tested to generate a single regression for each output variable. The maximum number of input variables was set to 9 due to the limited number of training samples, based on the literature review performed by Salleh *et al.* (2017) and preliminary tests, which in turn avoids overfitting.

In regard to the settings of the EPR-MOGA-XL software, the expression structure that was selected is presented in Equation (8). The equation includes an exponential function because it is similar to the first-order model that was selected for the decay of bulk chlorine, which is in turn related to the formation of TTHMs. The maximum number of terms was set to 5, and the maximum number of past input values considered for the model was 3. Additionally, the candidate exponents set was $[-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2]$:

$$Y = \exp \left(a_0 + \sum_{j=1}^m a_j \cdot (\mathbf{X}_1)^{ES(j,1)} \cdot \dots \cdot (\mathbf{X}_k)^{ES(j,k)} \right) \quad (8)$$

where Y is the least squares estimate of the output value, a_j is a constant parameter for the j th term, a_0 is an optional bias, m is the number of terms in the expression, \mathbf{X}_k is the k th vector of inputs, k is the number of input variables and \mathbf{ES} is a matrix of exponents ($m \times k$).

3.4. Black-box models

The black-box models based on artificial intelligence that were implemented in this research were the following: ANNs with the software ANN-MOGA-XL (Giustolisi & Simeone 2006) and ANFIS. The models were designed with a Multiple Input Single Output (MISO) approach, which implies that 30 models were created: 15 for chlorine and 15 for TTHMs. The available data from the water quality simulation were sampled every 30 min from 45 to 100 h, considering the warm-up period, which resulted in 111 records. The data were divided in two sets: 70% of the samples were used in the training and validation

Table 2 | Input and output variables of the models

Parameter	Inputs	Outputs
Number of points	20	15
Variables	Pressure Chlorine TTHMs	– Chlorine TTHMs

set and 30%, in the test set, based on [AlDahoul et al. \(2022\)](#). Other authors have proposed a 75–25% distribution ([Cuesta Cordoba et al. 2014](#)) or a 80–20% division, but a higher percentage was given to the test set in this work for representativeness due to the low number of records.

In addition, the best model for each output variable was tested under two different operational conditions. Regarding Scenario A, a random deviation $N(0, \sigma)$ was added to the demand patterns of each node, where $\sigma = 0.05\mu/3.27$ and μ is the mean demand multiplier, that replicates the process proposed by [Kang & Lansey \(2010\)](#). The duration of this simulation was only 80 h due to math operation errors related to EPANET-MSX. In regard to Scenario B, the status of a strategic pipe was set to closed, which led to increasing the initial level of six tanks and changing the rule-based control of a valve to avoid negative pressures. The duration of this scenario was the same as the original model, namely 100 h. The accuracy of both scenarios was assessed with the water quality results from 45 h on.

Concerning ANNs, the software ANN-MOGA-XL uses a MOGA to select the single-hidden-layer ANNs that lie on the Pareto front considering the minimization of three objective functions: fitness function computed on a validation set of data, the input dimension, and the number of hidden neurons ([Giustolisi & Simeone 2006](#)). The architecture and the performance of ANNs have been optimized using single-objective algorithms such as artificial bee colony (ABC), GAs, evolutionary strategy (ES) ([Gupta & Raza 2019](#); [Abdolrasol et al. 2021](#)), particle swarm optimization (PSO) and stochastic gradient descent deep learning (SGD) ([Hai et al. 2022](#)), among others. In contrast, multi-objective approaches have not been explored thoroughly, although they are necessary due to the tendency to overfitting and the curse of dimensionality ([Giustolisi & Simeone 2006](#)). For this reason, the MOGA approach was adopted for the ANNs design.

In this research four tests were performed using different kernel functions, as displayed in [Table 3](#). On the first test, ANNs were trained for all the output variables using the hyperbolic tangent as a kernel function, as recommended by [Berardi et al. \(2011\)](#). On the other tests, ANNs were trained only for 10 output variables, which had not achieved a suitable Test Nash–Sutcliffe Index (NSI) using ANFIS nor ANNs with a hyperbolic tangent function (Test 1). Regarding the settings, based on preliminary tests, the maximum number of hidden neurons was set to 5, the maximum number of past input values considered for the model was 2 and the number of generations was 1,000.

Regarding ANFIS, it is an adaptive network with ANN topology together with FL rules ([Yetilmezsoy 2019](#)). One of the most used ANFIS is the Takagi–Sugeno–Kang system, that contains if-then rules in which the premise refers to a fuzzy set, and the consequent is a linear combination of membership values ([Jang 1993](#)). In this sense, the adjustment of the ANFIS parameters was developed with a two-step optimization process in MATLAB[®]: first, the rules generation was done with NSGA-II; second, the training, that is the tuning of the premise and consequent parameters of each rule, was completed with GAs. Unlike ANNs, the ANFIS optimization did not include past input values for the model; only the current values of the inputs were used.

The rules generation was done with the subtractive clustering algorithm, that determines the number of fuzzy sets for each input and the number of rules based on the selection of cluster centres ([Alata et al. 2013](#)). The subtractive clustering algorithm depends on four parameters: the acceptance ratio (AR), the rejection ratio (RR), the cluster influence range (CIR) and the squash factor. The first three parameters range from 0 to 1, and the last one is a positive scalar for which the default MATLAB[®] value of 1.25 was adopted.

The number of fuzzy sets and the number of rules for ANFIS are usually defined by trial and error ([Zio & Bazzo 2011](#)), although some authors have proposed single-objective optimization algorithms for this aim, such as GAs ([Alata et al. 2013](#)), differential evolution and PSO ([Ahmadianfar et al. 2022](#)), among others. However, it is fundamental to reach a compromise between the fitness calculated on the training dataset and the generalization capacity of this method. For this reason,

Table 3 | Characteristics of the test performed with ANN-MOGA-XL

Test	Kernel function	Output variables
1	Hyperbolic tangent	30
2	Sigmoid	10
3	Quadratic	10
4	Linear	10

a multi-objective approach is necessary to avoid overfitting. Some of the multi-objective optimization algorithms that are used for different purposes include Pareto Archived Evolution Strategy (PAES), Pareto Envelop based Selection Algorithm-II (PESA-II), NSGA-II, multi-objective PSO, multi-objective ant colony optimization, among others (Sharma & Kumar 2022). The NSGA-II algorithm was chosen since it is one of the most popular optimization algorithms due to its fast and efficient convergence, as well as the diversity of solutions it is able to explore (Sharma & Kumar 2022).

In this way, the first step of the optimization process consisted of using NSGA-II to optimize the acceptance ratio, the RR and the CIR with a single constraint: the acceptance ratio must be greater than the RR. The algorithm minimized three objective functions: training error, test error and test-to-training error difference, in conformity with the process described in Mathworks (2021b). The errors were calculated as 1-NSI. The Constrained NSGA-II code written by Baskar *et al.* (2015) was modified and adopted for this research. After preliminary tests, the parameters of the algorithm were defined as 20 for the distribution index for crossover, 100 for the distribution index for mutation, one-third for the probability of mutation, based on the values proposed by Deb *et al.* (2002). Additionally, 150 was set for the number of generations and 50 for the population size considering two runs, relying on the default values of MATLAB[®] (Mathworks 2021a).

Subsequently, the optimal models for each output variable obtained from the NSGA-II process were trained using GAs. This process consists of adjusting the premise parameters, related to membership functions, and the consequent parameters, associated to the defuzzy linear function. Other training approaches that have been proposed include ant colony optimization extended to continuous domain (ACOr), PSO and differential evolution (Karami *et al.* 2022); GAs were adopted since they tend to have a better generalization capacity (Azad *et al.* 2019). The objective function was the training RMSE weighed against the increase in the validation RMSE for each iteration. Moreover, a *k*-fold cross validation procedure was used to avoid overfitting (Hadjisolomou *et al.* 2021). The values adopted for the parameters of the algorithm were 100 for the number of generations, 20 for the number of stall generations, 2% for the validation tolerance and five sets for the *k*-fold cross validation with a validation window of 3. These values are based on the application described by Mathworks (2021b) and the analysis of the results with different values of the parameters. Furthermore, a fitness limit in terms of RMSE was assigned to avoid overfitting.

4. RESULTS AND DISCUSSION

Regarding sampling design, EPR-MOGA-XL was successfully applied to select the most significant inputs to describe the behaviour of the chlorine and TTHM concentrations at the output nodes. For example, Figure 3 shows the locations of the sampling points or inputs, represented by blue stars, for the TTHM concentration at node 1,005, symbolized by a green star. Figure 3 illustrates that the inputs are not necessarily located near the output, and that they may be placed downstream or upstream. Moreover, a preliminary analysis showed that the inputs and the outputs share similar tracer percentages for each WTP of the network.

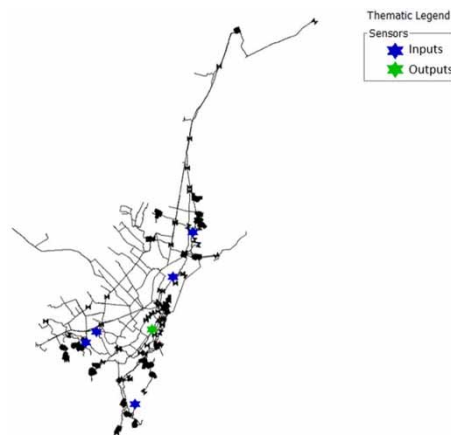


Figure 3 | Sensor locations for TTHMs at node 1005. Please refer to the online version of this paper to see this figure in colour: <http://dx.doi.org/10.2166/hydro.2023.028>.

Additionally, it was found that chlorine and TTHM inputs were more frequently selected for chlorine and TTHM outputs, respectively. Therefore, the variables that should be measured at the sampling points are directly related to the type of output variable that the model intends to predict. On the other hand, pressure was the least frequently selected variable, although this value measured at a few sampling points proved to be essential for the prediction of chlorine and TTHMs throughout the network.

The best ANNs and ANFIS models in terms of the Training NSI (TrNSI) and Test NSI (TNSI) with a one-step-ahead prediction are displayed in Tables 4 and 5. The Model column refers to the name of each output, that has the initial of the variable (T for TTHMs and C for chlorine), followed by a hyphen and the node ID. The Model ID column specifies the test conditions in which the best results were achieved: NSGA-II refers to ANFIS in which only this optimization method was used, while GA stands for an NSGA-II rules generation followed by a GA training. The most chosen test for ANFIS was NSGA-II, which implies that using this algorithm is enough to ensure an acceptable performance, and that GA training tends to cause overfitting. Regarding ANNs, the most frequently selected test was Test 1, which was done using hyperbolic tangent as a transfer function, that confirms the suitability of this function for prediction purposes.

It should be noted that five TTHMs and five chlorine models did not present a satisfactory TNSI using neither ANFIS nor ANNs considering an NSI above 0.7 as a criterion. Given that ANFIS tended to provide a better performance than ANNs and that GA training was prone to generate overfitting, a test only applying NSGA-II doubling the quantity of generations (300 generations) was done for these output variables. The Model ID for this test is NSGA-II (*). This approach was not effective as demonstrated by the low values of the TNSI.

The #inputs column refers to the number of input variables of the model without considering past input values, whereas the #inputs with lag column accounts for past inputs that have a time lag with respect to the output variable, with a maximum of two time-steps, as mentioned above for ANNs. Conversely, the #inputs EPR presents the number of input variables that were originally selected using EPR-MOGA. Supplementary material, Tables S1 and S2 show the inputs for each black-box model presented in Tables 4 and 5. It should be noticed that the number of inputs in these three columns is the same for ANFIS models, but for ANNs the effective number of inputs is lower than the original EPR inputs since ANN-MOGA further reduces the input dimension. Moreover, Tables 4 and 5 show that a lower number of inputs is related with a higher TNSI, thus ANN-MOGA has an advantage that could be included within the ANFIS algorithm.

Table 4 | Best models for TTHM outputs

Model	Best model	Training NSI	Test NSI	Model ID	# Inputs	# Inputs with lag	# Inputs EPR
T-1005	ANFIS	0.7083	-0.0567	NSGA-II	6	6	6
T-10745	ANFIS	0.9711	0.9702	GA	6	6	6
T-1191	ANN	0.7320	0.7223	ANN 3	5	10	6
T-1257	ANN	0.9776	0.9696	ANN 1	3	4	7
T-1295	ANN	0.7014	0.3933	ANN 3	7	9	9
T-1351	ANFIS	0.9603	0.8869	GA	6	6	6
T-1497	ANFIS	0.8941	0.3439	NSGA-II	8	8	8
T-1562	ANFIS	0.9568	0.8459	NSGA-II	8	8	8
T-1715	ANFIS	0.9956	0.9658	NSGA-II	5	5	5
T-190	ANFIS	0.7401	0.5332	NSGA-II	9	9	9
T-2312	ANFIS	1.0000	0.1366	NSGA-II(*)	9	9	9
T-2375	ANN	0.9799	0.9467	ANN 1	3	4	6
T-1028	ANN	0.9685	0.9334	ANN 1	2	2	6
T-3213	ANFIS	0.9668	0.8632	NSGA-II	6	6	6
T-920	ANFIS	0.9982	0.9549	GA	8	8	8
Mean	-	0.9034	0.6939	-	6.07	6.67	7.00

Table 5 | Best models for chlorine outputs

Model	Best model	Training NSI	Test NSI	Model ID	# Inputs	# Inputs with lag	# Inputs EPR
C-1005	ANFIS	0.7244	0.4272	NSGA-II	7	7	7
C-10745	ANN	0.9905	0.9648	ANN 1	3	3	5
C-1191	ANFIS	0.9911	0.7988	GA	6	6	6
C-1257	ANN	0.9704	0.9357	ANN 1	4	6	6
C-1295	ANN	0.8468	0.3661	ANN 2	8	14	9
C-1351	ANN	0.9765	0.9306	ANN 1	6	7	7
C-1497	ANFIS	0.7943	0.7484	GA	6	6	6
C-1562	ANN	0.8157	0.6907	ANN 4	2	2	4
C-1715	ANFIS	0.9905	0.9479	NSGA-II	7	7	7
C-190	ANFIS	1.0000	0.2872	NSGA-II	7	7	7
C-2312	ANFIS	1.0000	0.2836	NSGA-II(*)	8	8	8
C-2375	ANFIS	0.9410	0.9268	NSGA-II	4	4	4
C-1028	ANFIS	1.0000	0.7852	NSGA-II	8	8	8
C-3213	ANN	0.9758	0.8576	ANN 1	3	4	7
C-920	ANFIS	0.9648	0.9212	GA	6	6	6
Mean	–	0.9321	0.7248	–	5.67	6.33	6.47

The most frequently selected model was ANFIS since it has a better performance in terms of both Training and TNSI. Indeed, 10 out of 15 TTHM output variables had the best results using ANFIS, and nine out of 15 chlorine variables generated the best fitness values applying ANFIS. Additionally, the mean TrNSI and the mean TNSI were 25.72 and 23.20% higher for ANFIS models compared to ANNs, respectively. Nonetheless, 60% of the output variables have a relative change less or equal than 5.32 and 3.26% in the TrNSI and the TNSI, respectively, for ANFIS models in contrast to ANNs.

The TrNSI was greater than 0.7 for all the output variables, thus the training efficiency of the models was considered suitable. Particularly, T-920, T-2312, and C-2312 did not achieve an acceptable training performance when using ANNs, but they did when using ANFIS. Additionally, the test fitness of T-920, C-1191 and C-1497 fulfilled acceptable values with ANFIS, while they did not with ANNs.

Figure 4 displays the performance of the best model in terms of the TrNSI and the TNSI, that was found with ANFIS for the TTHM concentration at node 1715. Figure 5 exhibits the data predicted by the worst model in terms of TrNSI and TNSI versus the simulated concentrations of TTHMs at node 1005. Despite the TrNSI is above 0.7, the TNSI is below zero, which implies that the mean is a better predictor of the output than the model. According to McCuen *et al.* (2006), low values of the NSI may be caused by significant bias, outliers or low sample size. Moreover, negative values take place when measured parameters approach the average value (ASCE 1993), which is the case for the test data of T-1005, as demonstrated in Figure 5.

In this sense, as mentioned above five TTHMs and five chlorine models did not achieve a suitable TNSI. Particularly, the test fitness in four nodes was low for both chlorine and TTHM models: 1005, 1295, 190, and 2312. In addition, the models for T-1497 and C-1562 had low TNSI values. The corresponding plots that show the contrast between the simulated and the predicted data for these models are presented in the supplementary material. The predicted data show visual similarity and lie within the range of percentage error that is required to monitor water quality in a WDN, although the TNSI is low. The data were not found to have considerable bias nor outliers; thus, the low TNSI values may have been caused by the small sample size.

The test performance may be improved by using a longer simulation time so that the concentrations stabilize in periodic patterns, which ensures a similar behaviour between training and test data. This is particularly relevant for node 190, T-1005 and T-1497, as illustrated in the supplementary material. Moreover, node 2312 exhibits overfitting as demonstrated by a high TrNSI and low TNSI, which may be due to a high number of inputs. In addition, C-1562 and the models for node 1295 display a time-offset bias and a poor test performance. The reason for this may be the lack of relevant inputs to describe the tendency of the outputs, as well as the variation of the determining input variables over time.

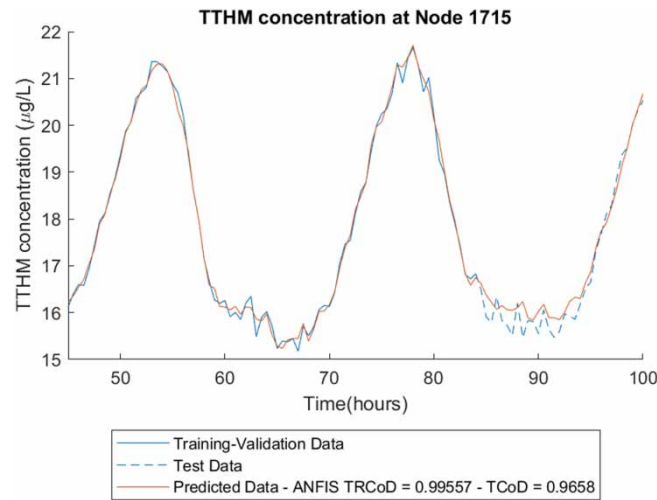


Figure 4 | TTHM concentrations at node 1715.

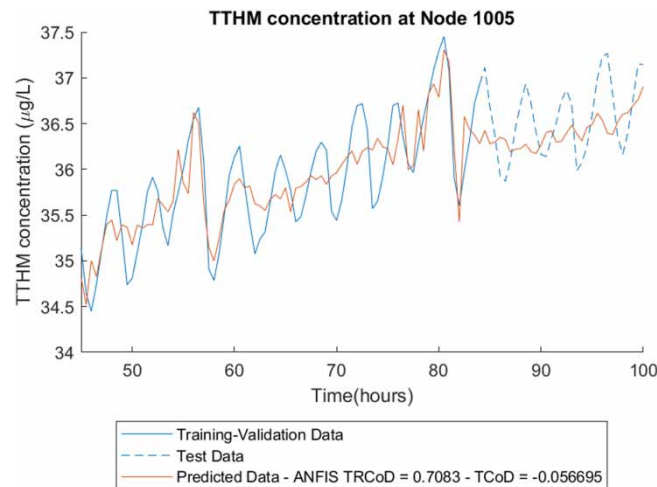


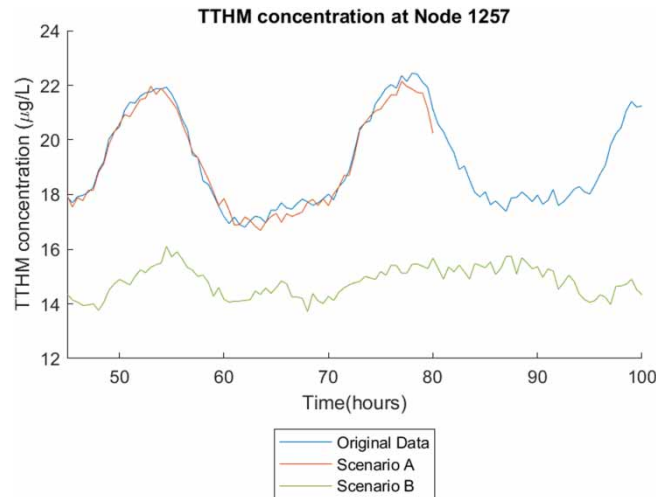
Figure 5 | TTHM concentrations at node 1005.

Regarding the tests under different operational conditions, [Table 6](#) summarizes the results of both Scenarios A and B compared with the fitness obtained with the original data. The mean test performance of the models for Scenario A is above 0.7 for both TTHMs and chlorine. In fact, the mean TNSI for Scenario A of the TTHM models is greater than the mean for the original TNSI. In addition, in the case of chlorine models, the TNSI is slightly lower than the original value, but it is still above 0.7. Hence, although the TNSI reduction was significant for eight output variables that already had a poor test performance in the original scenario, in general it is considered that the models have a satisfactory test fitness under Scenario A.

On the contrary, the mean TNSI for Scenario B was below zero for both TTHM and chlorine output variables, thus the test performance of the models was not suitable. Particularly, there are 14 variables for which the TNSI under Scenario B is negative, which indicates that the mean is a better predictor than the model for all the simulation time. Indeed, [Figure 6](#) illustrates that the operational conditions under Scenario B cause significant changes in comparison with the original data or Scenario A in the case of node 1257. However, nine models exhibited a higher level of similarity between the time series of the Original Scenario and Scenario B while having a lower TNSI. This fact demonstrates that the NSI has several limitations that should be further studied.

Table 6 | Results for the tests under different operational conditions

Model	Mean values			
	Training NSI	Test NSI	A – Test NSI	B – Test NSI
TTHMs	0.9034	0.6939	0.7234	-0.7178
Chlorine	0.9321	0.7248	0.7234	-8.2125

**Figure 6** | Results under different operational conditions: TTHM concentrations at node 1257.

5. CONCLUSIONS

Regarding the sampling design, the results showed that proximity is not the main criterion to define the most significant input locations for the prediction of chlorine and TTHMs, but that a node having a similar tracer percentage for the water sources of the network in comparison with the output node has a greater possibility to be selected. In addition, it was found that the most frequently selected input variables are of the same type as the output variable that the model seeks to predict, while only the pressure at a few nodes was identified as influential.

Regarding the black-box models, it was found that an ANN with a hyperbolic tangent kernel function and an ANFIS with optimized subtractive clustering parameters using NSGA-II allow for a sufficient prediction estimation in most cases. Although ANFIS outperformed ANNs considering the average indicators of the models on both training and generalization capacity, the relative change between both is low for most of the variables. Furthermore, the advantage of the ANN-MOGA software over ANFIS is that it further reduces the input dimension, which in fact enhances the test fitness. Hence, both methods offer satisfactory prediction accuracy, and the results are not conclusive regarding which conditions are better to apply one or another.

In this way, by choosing the best models between both approaches, the mean training performance was satisfactory for both chlorine and TTHMs considering 0.7 as a lower bound for the NSI values. Despite the mean TNSI for chlorine variables was above 0.7, the mean TNSI for TTHMs was below 0.7, as well as the test fitness for 10 output variables. There are five possible reasons for this: (a) differences between the tendency of the training and the test data, (b) overfitting, (c) a lack of relevant inputs to fully describe the behaviour of the outputs, (d) the determining inputs variation over time, (e) the limitations of the NSI in terms of outliers and bias. The recommendations for each issue are the following: (a) to increase the simulation duration and the warm-up period, (b) to review the stopping criteria and reduce the number of inputs, (c) to include other hydraulic and water quality variables, (d) to verify if varying inputs improve the performance of the models, and (e) to explore other indicators besides this index.

In addition, the models were tested under two different operational conditions. In general, the models achieved satisfactory fitness values under Scenario A. On the contrary, the models' performance under Scenario B were poor due to the substantial

changes of this condition with respect to the Original Scenario. Consequently, it is suggested incorporating different operational status and chlorine doses in the training data or even include the operational status as an input, since the models are limited by the range of the variables in the training data.

In summary, this research aimed to identify if black-box models based on artificial intelligence with an appropriate sampling design are an effective water quality modelling alternative as a proof-of-concept study. Based on the results, it can be concluded that in general black-box models have a robust prediction capacity for chlorine and TTHM concentrations in the case study. Although in this research the generalization capability was limited to particular operational conditions and the sampling design was static over time, black-box models based on artificial intelligence proved to be a promising approach for management applications in water utilities in terms of water quality.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the help and the data of the trunk network of Bogotá provided by Dr Ivonne Navarro, Dr Mauricio Jiménez, and the Water Supply and Sewerage Company of Bogotá. Moreover, the authors wish to extend their special thanks to Dr Orazio Giustolisi and Dr Luigi Berardi for providing the licenses for EPR-MOGA-XL and ANN-MOGA-XL, which were used for this research.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

- Abdolrasol, M. G. M., Suhail Hussain, S. M., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef, S. & Milad, A. 2021 **Artificial neural networks based optimization techniques: a review**. *Electronics (Switzerland)*. <https://doi.org/10.3390/electronics10212689>.
- Ahmadianfar, I., Shirvani-Hosseini, S., He, J., Samadi-Koucheksaraee, A. & Yaseen, Z. M. 2022 **An improved adaptive neuro fuzzy inference system model using conjoined metaheuristic algorithms for electrical conductivity prediction**. *Scientific Reports* **12** (1). <https://doi.org/10.1038/s41598-022-08875-w>.
- Alata, M., Molhim, M. & Ramini, A. 2013 **Using GA for optimization of the fuzzy C-means clustering algorithm**. *Research Journal of Applied Sciences, Engineering and Technology* **5** (3). <https://doi.org/10.19026/rjaset.5.5011>.
- AIDahoul, N., Ahmed, A. N., Allawi, M. F., Sherif, M., Sefelnasr, A., Chau, K. w. & El-Shafie, A. 2022 **A comparison of machine learning models for suspended sediment load classification**. *Engineering Applications of Computational Fluid Mechanics* **16** (1). <https://doi.org/10.1080/19942060.2022.2073565>.
- Alver, A., Baştürk, E. & Kılıç, A. 2021 **Development of adaptive neuro-fuzzy inference system model for predict trihalomethane formation potential in distribution network simulation test**. *Environmental Science and Pollution Research*. <https://doi.org/10.1007/s11356-020-11801-0>.
- ASCE 1993 **Criteria for evaluation of watershed models**, ASCE task committee on definition of criteria for evaluation of watershed models of the watershed management, irrigation, and drainage division. *Journal of Irrigation and Drainage Engineering* **119** (3).
- Azad, A., Karami, H., Farzin, S., Mousavi, S. F. & Kisi, O. 2019 **Modeling river water quality parameters using modified adaptive neuro fuzzy inference system**. *Water Science and Engineering* **12** (1). <https://doi.org/10.1016/j.wse.2018.11.001>.
- Baskar, S., Selvaraj, T. & Varshini, P. R. 2015 **MATLAB Code for Constrained NSGA II**. [Online]. Available from: <https://www.mathworks.com/matlabcentral/fileexchange/49806-matlab-code-for-constrained-nsga-ii-dr-s-baskar-s-tamilselvi-and-p-r-varshini> (accessed 5 October 2021).
- Berardi, L., Laucelli, D. & Giustolisi, O. 2011 **ANN MOGA – XL User Guide**.
- Boccelli, D. L., Tryby, M. E., Uber, J. G. & Summers, R. S. 2003 **A reactive species model for chlorine decay and THM formation under rechlorination conditions**. *Water Research*. [https://doi.org/10.1016/S0043-1354\(03\)00067-8](https://doi.org/10.1016/S0043-1354(03)00067-8).
- Bressane, A., Goulart, A. P. G., Melo, C. P., Gomes, I. G., Loureiro, A. I. S., Negri, R. G., Moruzzi, R., Reis, A. G. d., Formiga, J. K. S., da Silva, G. H. R. & Thomé, R. F. 2023 **A non-Hybrid data-driven fuzzy inference system for coagulant dosage in drinking water treatment plant: machine-learning for accurate real-time prediction**. *Water* **15** (6). <https://doi.org/10.3390/w15061126>.
- CIACUA 2009 **Informe sobre la modelación de calidad de la Red Matriz de Bogotá (Report: Water Quality Modelling in the Trunk Network of Bogotá)**. Bogotá.

- Cuesta Cordoba, G. A., Tuhovčák, L. & Tauš, M. 2014 Using artificial neural network models to assess water quality in water distribution networks. *Procedia Engineering*. <https://doi.org/10.1016/j.proeng.2014.02.045>.
- Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. 2002 A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **6** (2). <https://doi.org/10.1109/4235.996017>.
- Deng, T., Chau, K. W. & Duan, H. F. 2021 Machine learning based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management* **284**. <https://doi.org/10.1016/j.jenvman.2021.112051>.
- Díaz, Ó. R. & Saldarriaga, J. G. 2015 *Calibración de modelos de calidad del agua en redes de distribución*. [Online]. Universidad de los Andes. Available from: <https://repositorio.uniandes.edu.co/bitstream/handle/1992/13382/u722222.pdf?sequence=1&isAllowed=y>.
- Dini, M. & Tabesh, M. 2017 Water distribution network quality model calibration: A case study-Ahar. *Water Science and Technology: Water Supply* **17** (3), 759–770. <https://doi.org/10.2166/ws.2016.166>.
- D'Souza, C. D. & Kumar, M. S. M. 2010 Comparison of ANN models for predicting water quality in distribution systems. *Journal/American Water Works Association*. <https://doi.org/10.1002/j.1551-8833.2010.tb10152.x>.
- Eliades, D. G., Kyriakou, M., Vrachimis, S. G. & Polycarpou, M. M. 2016 EPANET-MATLAB Toolkit : An Open-Source Software for Interfacing EPANET with MATLAB. In: *Computing and Control for the Water Industry CCWI 2016*.
- Empresa de Acueducto y Alcantarillado de Bogotá – E.A.A.B. 2018 *Manual de operación del sistema matriz acueducto (Operations Manual of the Trunk Network)*. Bogotá.
- Empresa de Acueducto y Alcantarillado de Bogotá – E.A.A.B. 2021 *Geographic Information System EAAAB*. [Online]. Available from: <https://www.acueducto.com.co/wassigue1/VisorBaseEAB/> (accessed 6 February 2021).
- Giudicianni, C., Herrera, M., Di Nardo, A., Greco, R., Creaco, E. & Scala, A. 2020 Topological placement of quality sensors in water-distribution networks without the recourse to hydraulic modeling. *Journal of Water Resources Planning and Management* **146** (6). [https://doi.org/10.1061/\(asce\)wr.1943-5452.0001210](https://doi.org/10.1061/(asce)wr.1943-5452.0001210).
- Giustolisi, O. & Savic, D. A. 2006 A symbolic data-driven technique based on evolutionary polynomial regression. *Journal of Hydroinformatics* **8** (3). <https://doi.org/10.2166/hydro.2006.020b>.
- Giustolisi, O. & Savic, D. A. 2009 Advances in data-driven analyses and modelling using EPR-MOGA. *Journal of Hydroinformatics* **11** (3–4). <https://doi.org/10.2166/hydro.2009.017>.
- Giustolisi, O. & Simeone, V. 2006 Optimal design of artificial neural networks by a multi-objective strategy: Groundwater level predictions. *Hydrological Sciences Journal*. <https://doi.org/10.1623/hysj.51.3.502>.
- Godó-Pla, L., Emiliano, P., Poch, M., Valero, F. & Monclús, H. 2021 Benchmarking empirical models for THMs formation in drinking water systems: an application for decision support in Barcelona, Spain. *Science of the Total Environment* **763**. <https://doi.org/10.1016/j.scitotenv.2020.144197>.
- Gupta, T. K. & Raza, K. 2019 Optimization of ANN architecture: a review on nature-inspired techniques. In: *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*. <https://doi.org/10.1016/B978-0-12-816086-2.00007-2>.
- Hadjisolomou, E., Stefanidis, K., Herodotou, H., Michaelides, M., Papatheodorou, G. & Papastergiadou, E. 2021 Modelling freshwater eutrophication with limited limnological data using artificial neural networks. *Water (Switzerland)* **13** (11). <https://doi.org/10.3390/w13111590>.
- Haghiabi, A. H., Nasrolahi, A. H. & Parsaie, A. 2018 Water quality prediction using machine learning methods. *Water Quality Research Journal* **53** (1). <https://doi.org/10.2166/wqrj.2018.025>.
- Hai, T., Li, H., Band, S. S., Shadkani, S., Samadianfard, S., Hashemi, S., Chau, K. & Mousavi, A. 2022 Comparison of the efficacy of particle swarm optimization and stochastic gradient descent algorithms on multi-layer perceptron model to estimate longitudinal dispersion coefficients in natural streams. *Engineering Applications of Computational Fluid Mechanics* **16** (1).
- Hart, W. E., Berry, J. W., Boman, E. G., Murray, R., Phillips, C. A., Riesen, L. A. & Watson, J. P. 2008 The TEVA-SPOT toolkit for drinking water contaminant warning system design. In: *World Environmental and Water Resources Congress 2008: Ahupua'a – Proceedings of the World Environmental and Water Resources Congress 2008*. [https://doi.org/10.1061/40976\(316\)513](https://doi.org/10.1061/40976(316)513).
- Hong, S. T., Lee, A. K., Lee, H. H., Park, N. S. & Lee, S. H. 2012 Application of neuro-fuzzy PID controller for effective postchlorination in water treatment plant. *Desalination and Water Treatment*. <https://doi.org/10.1080/19443994.2012.696810>.
- Ighalo, J. O., Adeniyi, A. G. & Marques, G. 2021 *Artificial intelligence for surface water quality monitoring and assessment: A systematic literature analysis*. *Modeling Earth Systems and Environment*. <https://doi.org/10.1007/s40808-020-01041-z>.
- Jang, J. S. R. 1993 ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man and Cybernetics* **23** (3), 665–685.
- Kang, D. & Lansey, K. 2010 Real-time optimal valve operation and booster disinfection for water quality in water distribution systems. *Journal of Water Resources Planning and Management* **136** (4). [https://doi.org/10.1061/\(asce\)wr.1943-5452.0000056](https://doi.org/10.1061/(asce)wr.1943-5452.0000056).
- Karami, H., DadrasAjirlou, Y., Jun, C., Bateni, S. M., Band, S. S., Mosavi, A., Moslehpour, M. & Chau, K. W. 2022 A novel approach for estimation of sediment load in dam reservoir with hybrid intelligent algorithms. *Frontiers in Environmental Science* **10**. <https://doi.org/10.3389/fenvs.2022.821079>.
- Khorshidi, M. S., Nikoo, M. R. & Sadegh, M. 2018 Optimal and objective placement of sensors in water distribution systems using information theory. *Water Research* **143**. <https://doi.org/10.1016/j.watres.2018.06.050>.
- Klise, K. A., Nicholson, B. & Laird, C. D. 2017 *Sensor Placement Optimization Using Chama*, Sandia Report SAND2017-11472.
- Kouadri, S., Elbeltagi, A., Islam, A. R. M. T. & Kateb, S. 2021 Performance of machine learning methods in predicting water quality index based on irregular data set: Application on Illizi region (Algerian southeast). *Applied Water Science* **11** (12). <https://doi.org/10.1007/s13201-021-01528-9>.

- Krause, A., Leskovec, J., Guestrin, C., VanBriesen, J. & Faloutsos, C. 2008 Efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management* **134** (6). [https://doi.org/10.1061/\(asce\)0733-9496\(2008\)134:6\(516\)](https://doi.org/10.1061/(asce)0733-9496(2008)134:6(516)).
- Lees, M. & Wagener, T. 2000 *Monte-Carlo Analysis Tool (MCAT) V.2*.
- Mathworks 2021a *Genetic Algorithm Options*. [Online]. Available from: <https://www.mathworks.com/help/gads/genetic-algorithm-options.html> (accessed 10 September 2021).
- Mathworks 2021b *Optimize FIS Parameters with k-Fold Cross-Validation*. [Online]. Available from: <https://www.mathworks.com/help/fuzzy/fis-parameter-optimization-with-k-fold-cross-validation.html>.
- Mazhar, M. A., Khan, N. A., Ahmed, S., Khan, A. H., Hussain, A., Rahisuddin, Changani, F., Yousefi, M., Ahmadi, S. & Vambol, V. 2020 Chlorination disinfection by-products in municipal drinking water – A review. *Journal of Cleaner Production*. <https://doi.org/10.1016/j.jclepro.2020.123159>.
- McCuen, R. H., Knight, Z. & Cutter, A. G. 2006 Evaluation of the Nash–Sutcliffe efficiency index. *Journal of Hydrologic Engineering* **11** (6). [https://doi.org/10.1061/\(asce\)1084-0699\(2006\)11:6\(597\)](https://doi.org/10.1061/(asce)1084-0699(2006)11:6(597)).
- Milot, J., Rodriguez, M. J. & Sérodes, J. B. 2002 Contribution of neural networks for modeling trihalomethanes occurrence in drinking water. *Journal of Water Resources Planning and Management*. [https://doi.org/10.1061/\(asce\)0733-9496\(2002\)128:5\(370\)](https://doi.org/10.1061/(asce)0733-9496(2002)128:5(370)).
- Monteiro, L., Carneiro, J. & Covas, D. I. C. 2020 Modelling chlorine wall decay in a full-scale water supply system. *Urban Water Journal [Online]* **17** (8), 754–762. <https://doi.org/10.1080/1573062X.2020.1804595>.
- Nuckols, J. R., Rossman, L. W. & Singer, P. S. 2001 *Development of Exposure Assessment Methods for THM and HAA in Water Distribution Systems*. Denver.
- Peleato, N. M., Legge, R. L. & Andrews, R. C. 2018 Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products. *Water Research* **136**. <https://doi.org/10.1016/j.watres.2018.02.052>.
- Platikanov, S., Martín, J. & Tauler, R. 2012 Linear and non-linear chemometric modeling of THM formation in Barcelona's water treatment plant. *Science of the Total Environment* **432**. <https://doi.org/10.1016/j.scitotenv.2012.05.097>.
- Rajaei, T., Khani, S. & Ravansalar, M. 2020 Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review. *Chemometrics and Intelligent Laboratory Systems*. <https://doi.org/10.1016/j.chemolab.2020.103978>.
- Rathi, S. & Gupta, R. 2014 Sensor placement methods for contamination detection in water distribution networks: A review. *Procedia Engineering*. <https://doi.org/10.1016/j.proeng.2014.11.175>.
- Rodriguez, M. J., West, J. R., Powell, J. & Sérodes, J. B. 1997 Application of two approaches to model chlorine residuals in Severn Trent Water LTD (STW) distribution systems. *Water Science and Technology*. [https://doi.org/10.1016/S0273-1223\(97\)00489-7](https://doi.org/10.1016/S0273-1223(97)00489-7).
- Saldarriaga, J., Díaz, O. R. & Bohórquez, J. 2016 Water quality model calibration in potable water distribution systems. In: *Proceedings. World Environmental and Water Resources Congress 2016*. <https://doi.org/doi:10.1061/9780784479865.057>.
- Salleh, M. N. M., Talpur, N. & Hussain, K. 2017 Adaptive neuro-fuzzy inference system: Overview, strengths, limitations, and solutions. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-61845-6_52.
- Schal, S., Lothes, A., Bryson, L. S. & Ormsbee, L. 2013 Water quality sensor placement guidance using TEVA-SPOT. In: *Proceedings. World Environmental and Water Resources Congress 2013*. <https://doi.org/doi:10.1061/9780784412947.099>.
- Shang, F., Uber, J. & Rossman, L. 2011 *EPANET Multispecies Extension – User's Manual*.
- Sharma, S. & Kumar, V. 2022 A comprehensive review on multi-objective optimization techniques: past, present and future. *Archives of Computational Methods in Engineering* **29**, 5605–5633.
- Storey, M. V., van der Gaag, B. & Burns, B. P. 2011 Advances in on-line drinking water quality monitoring and early warning systems. *Water Research* **45** (2). <https://doi.org/10.1016/j.watres.2010.08.049>.
- U.S EPA National Homeland Security Research Center and Sandia National Laboratories 2015 *Water Security Toolkit User Manual: Version 1.3*.
- Vasconcelos, J. J., Rossman, L. A., Grayman, W. M., Boulos, P. F. & Clark, R. M. 1997 Kinetics of chlorine decay. *Journal/American Water Works Association* **89** (7). <https://doi.org/10.1002/j.1551-8833.1997.tb08259.x>.
- WHO 2011 *Guidelines for Drinking-Water Quality FOURTH EDITION WHO Library Cataloguing-in-Publication Data*. [online]. World Health Organisation. Available from: <http://www.who.int>.
- Wu, Z. Y. & Roshani, E. 2014 Sensor placement optimization for water quality model calibration. In: *World Environmental and Water Resources Congress 2014: Water Without Borders – Proceedings of the 2014 World Environmental and Water Resources Congress*. <https://doi.org/10.1061/9780784413548.057>.
- Wu, W., Dandy, G. C. & Maier, H. R. 2011 Application of artificial neural networks to forecasting water quality in a chloraminated water distribution system. In: *MODSIM 2011 - 19th International Congress on Modelling and Simulation – Sustaining Our Future: Understanding and Living with Uncertainty*. <https://doi.org/10.36334/modsim.2011.c1.wu>.
- Yetilmezsoy, K. 2019 Applications of soft computing methods in environmental engineering. In: *Handbook of Environmental Materials Management*. https://doi.org/10.1007/978-3-319-73645-7_149.
- Zio, E. & Bazzo, R. 2011 A clustering procedure for reducing the number of representative solutions in the Pareto Front of multiobjective optimization problems. *European Journal of Operational Research* **210** (3). <https://doi.org/10.1016/j.ejor.2010.10.021>.