Water Practice & Technology

PUBLISHING

© 2024 The Authors

Water Practice & Technology Vol 00 No 0, 1 doi: 10.2166/wpt.2024.231

Space–time prediction of residual chlorine in a water distribution network using artificial intelligence and the EPANET hydraulic model

Julius Caesar Kwio-Tamale* and Charles Onyutha

Department of Civil and Environmental Engineering, Kyambogo University, P.O. Box 1 Kyambogo, Kampala, Uganda
*Corresponding author. E-mail: tamalekwiojc@gmail.com

CO, 0000-0002-0652-3828

ABSTRACT

Insufficient knowledge of physical models and difficulty in fitting statistical models impair the choice of models to regulate residual chlorine in water distribution. This paper compared the performance of physical and statistical models in predicting residual chlorine concentrations in drinking water distribution. Drinking water was sampled from the downstream 128 water points water pipeline. Online chlorine concentrations were determined at water draw-off points. EPANET, the physical model, was used because of its efficiency in tracking dissolved chemicals. Statistical models used were regression, decision tree, random forest and artificial neural network. In the whole distribution network, the artificial neural network performed at R^2 of 94%, multi-linear regression (62%), random forest (55%), decision tree (41%), and EPANET (24%). However, EPANET yielded improved performance with R^2 above 70% when separately applied to individual sub-distribution networks; hence, is recommended for secondary chlorination in small distribution networks. For modelling large distribution networks, statistical models, especially an artificial neural network, are recommended. However, such cases still need support from confirmatory systems of interpretable parametric or hydraulic models that can achieve good performance with $R^2 \geq 80\%$. Water utilities can use these results to deploy model(s) for managing residual chlorine within safe limits of residual chlorine concentration in water distribution practice.

Key words: artificial intelligence, EPANET, residual chlorine decay, water quality modelling

HIGHLIGHTS

- Drinking water was sampled from 128 points downstream from a water treatment plant.
- R^2 for artificial neural network: 0.94, random forest: 0.44 and EPANET: 0.24.
- RMSE for EPANET was 0.58 mg/l and for statistical models ranged from 0.04 to 0.18 mg/l.
- An artificial neural network is recommended for predicting residual chlorine.
- Interpretable parametric models should supplement the use of an artificial neural network.

1. INTRODUCTION

Chlorine decay is the decrease in free residual chlorine concentration in drinking water as it passes from a water treatment plant to downstream points of the drinking water distribution system (World Health Organization 2017). Chlorine is the most common disinfectant in drinking water treatment because of its efficacy against pathogenic micro-organisms, low cost, ease of application, monitoring and extended disinfectant durability compared with other disinfectants (Hyunjun & Sanghyun 2017). In 2014, the World Health Organization recommended a minimum residual chlorine concentration of 0.2–5 mg/l at water consumption points to safeguard public health from microbial secondary contamination in the treated water supply (World Health Organization 2017). During water-borne disease outbreaks and emergencies, minimum residual chlorine is increased to 1.0 mg/l at tap stands and 2.0 mg/l at water delivery trucks (Rajasingham *et al.* 2020). Residual chlorine rotects against subsidiary pathogen intrusion in drinking water distribution networks (Kim *et al.* 2014).

Physical (process-driven or knowledge-driven) models (Bowden *et al.* 2019; Gibbs *et al.* 2019; Tiruneh *et al.* 2019a; Vuta & Dumitran 2019) and statistical (data-driven) models (Bowden *et al.* 2019; Gibbs *et al.* 2019; Tiruneh *et al.* 2019a) are used to simulate free residual chlorine concentrations in drinking water distribution systems. Physical models use underlying physical processes that transform inputs into outputs (Louppe 2014). The inputs in this case would be residual chlorine decay parameters and output as final residual chlorine concentration at water consumption draw-off points. However, the current knowledge of chemical kinetic reactions as the commonest physical model for residual chlorine decay in water is insufficient (Hyunjun & Sanghyun 2017; Bowden *et al.* 2019). Numerical methods are used to solve the differential equations that represent the physical laws that govern chemical reactions of residual chlorine decay in water (Vuta & Dumitran 2019). However, calculus and algebra are not as reliable as process models (Louppe 2014) because the underlying relationships of residual chlorine decay in treated water depend on several unknown and complex factors (Karadirek *et al.* 2015; Ricca *et al.* 2019). Consequently, process models take time to solve (Louppe 2014). Besides, simplifications and assumptions on model structure, one-dimensional flow are limitations (Karadirek *et al.* 2015). Simplifications and assumptions in physical models may be inaccurate, rendering process models unsuitable. Physical models are also deterministic(Steurer *et al.* 2019). This suggests that they are free from random variations in water quality modelling. EPANET hydraulic model, because of its popularity (Torretta *et al.* 2019), availability (Environmental Protection Agency 2018) and Lagrangian tracking of water chemical constituents (Kumar *et al.* 2015; Ricca *et al.* 2019), is the process model that is most commonly used to model water quality including residual chlorine decay (Karadirek *et al.* 2015; Tiruneh *et al.* 2019a).

In contrast, statistical models are data-driven (Loucks & van Beek 2017). Statistical models that have been used for residual chlorine decay modelling include linear regression models (Kim *et al.* 2014; Loucks & van Beek 2017; Oladipupo *et al.* 2019), multi-layer perceptron artificial neural networks (Gibbs *et al.* 2019; Ricca *et al.* 2019), GRNN (Generalized Regression Neural Network) models (Hyunjun & Sanghyun 2017), MLR (multiple linear regression) models (Hyunjun & Sanghyun 2017; Azad *et al.* 2019), artificial neural networks (Loucks & van Beek 2017; Azad *et al.* 2019; Oladipupo *et al.* 2019), time series neuro-fuzzy and adaptive neuro-fuzzy inference systems (ANFIS) networks and directed graph tree-based models like decision trees and random forests (Louppe 2014). However, statistical models are prone to underfitting or overfitting in prediction models (Bowden *et al.* 2019). This causes a model error by deviating from ground truths.

Therefore, regulating residual chlorine within this safe limit (0.2–5.0 mg/l) is still a big challenge (Vuta & Dumitran 2019; Wu & Dorea 2020). Despite such modelling challenges in determining water quality, we need models to use in predicting water quality. This is especially necessary and mandatory in long water distribution pipelines to guarantee public health and high environmental performance, as argued by (Torretta *et al.* 2019). Therefore, there is a need to compare the performance of physical and statistical models in predicting residual chlorine decay in water distribution systems like the 90 Km-long gravity drinking water distribution system in this study. Accordingly, the aim of this paper was to compare how physical and statistical models perform in predicting residual chlorine in drinking water distribution systems. Water utilities can use these results to choose and deploy model(s) for managing residual chlorine concentrations to remain within safe limits throughout the water distribution network.

2. METHODS AND MATERIALS

2.1. Study area

This research was conducted on the Lirima Gravity Flow Scheme located in Manafwa and Namisinde districts in the Mount Elgon region in Eastern Uganda. This gravity scheme is owned and operated by the National Water and Sewerage Corporation (NWSC) of Uganda which is a government parastatal. Figure 1 shows the water transmission main from the NWSC Lirima water treatment plant.

The GPS (Geographical Positioning System) coordinates of the water source and treatment plant of this gravity flow scheme is 36 N (Latitude), 0657122 (Northing), 0098196 (Easting) at an altitude of 1,812 m above sea level. The scheme starts just inside Uganda at the Uganda-Kenya border and it traverses 90 km in the hinterland in the study area.

2.2. Data sample size and data collection strategy

Morning and afternoon runs were conducted each day on particular distribution mains. On each run, data werecollected at sampling points at approximate intervals of 800–1,000 m. The strategy for this spacing interval was to ensure notable observations in residual chlorine concentrations in light of the low initial chlorine dosage at the water treatment plant and how fast it may decay in this system. This water sampling interval was based on the low initial chlorine dose of 0.73–1.00 mg/l at the water treatment plant to minimize chlorination cost and also

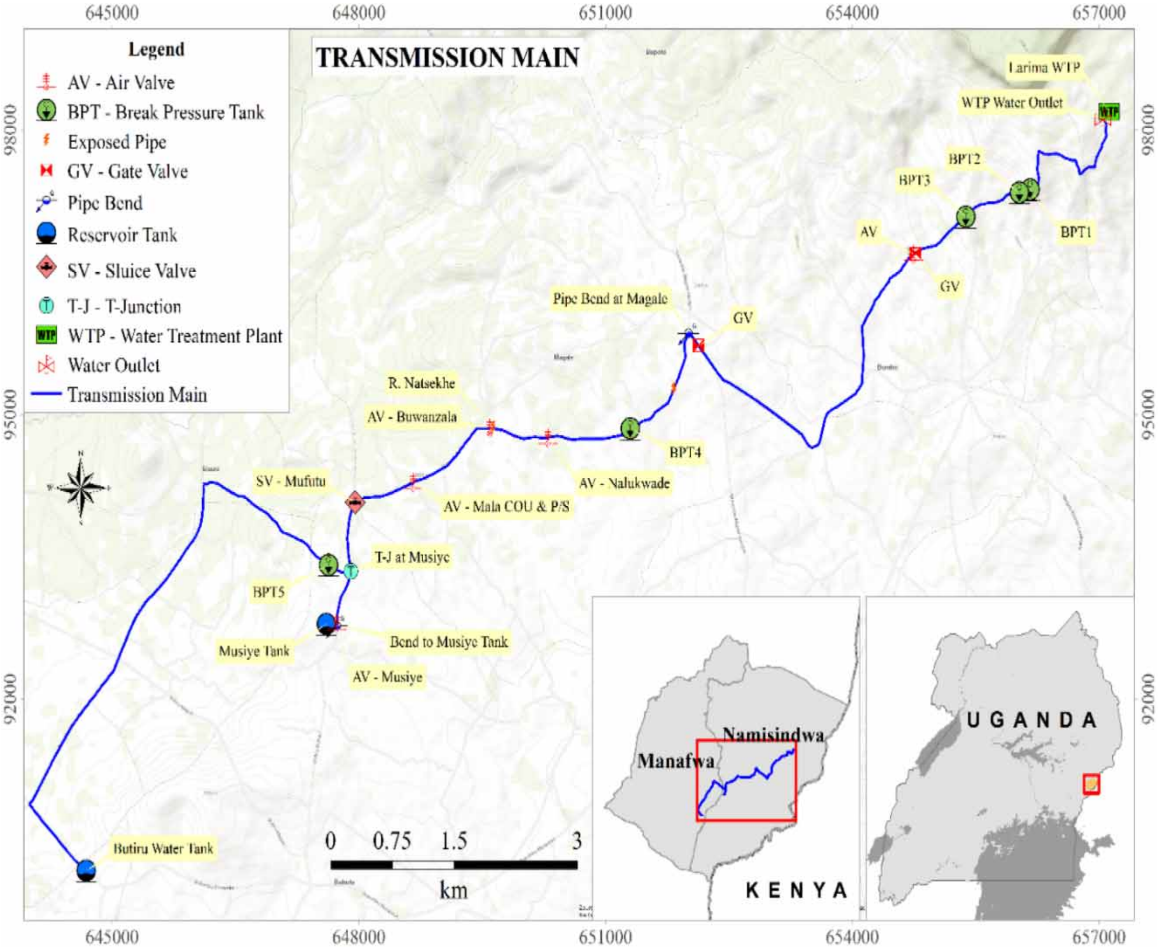

Figure 1 | Location of the Lirima gravity water distribution scheme in eastern Uganda.

minimize the formation of carcinogenic DBPs (disinfection by-products) that are associated with high chlorine dosage. Data collection was replicated on different days to simulate variations in study data. Replication of data on different days was also a strategy to increase the sample size of the study data. Three to four data points per sampling point were collected in the dry months of February–March 2021 totalling overall 128 datasets.

2.3. Data collection instruments and testing procedure

Water was sampled at clear water reservoir and break-pressure tank outlets, wash-outs and nearest functional yard taps that were on direct supply lines from water distribution and transmission mains. The yard water taps from which water was sampled were those that were very close to distribution mains within off-sets of less than 5 m, as shown in Figure 2. Horizontal distances and altitudes of these physical infrastructure components i.e. clear water reservoirs, break-pressure tank outlets, wash-outs and nearest functional yard taps were captured using GARMIN GPSMAP64s hand-held GPI (Geographic positioning instrument). Internal pipe diameters (pipe bores) were measured using steel tape measures directly at break-pressure outlets when the outlets were empty. The GPS coordinates were used to track the hydraulic paths and gradients of water transmission and distribution pipelines for hydraulic modelling in EPANET. It was assumed that water quality parameters at yard taps close to distribution networks would not have varied significantly from the water in the nearby distribution lines. Therefore, water in yard taps was considered to be practically representative of water quality parameter values. Online tests of residual chlorine, turbidity, temperature and electrical conductivity were done on water samples drawn from each water sample point mentioned above. Standard 1-litre bottles were used to draw water from break-pressure outlets, wash-outs and yard taps. Within seconds of sampling water in standard 1-litre bottles, a multi-functional Lovibond MD 600 digital metre was used to measure water quality parameters of residual chlorine in

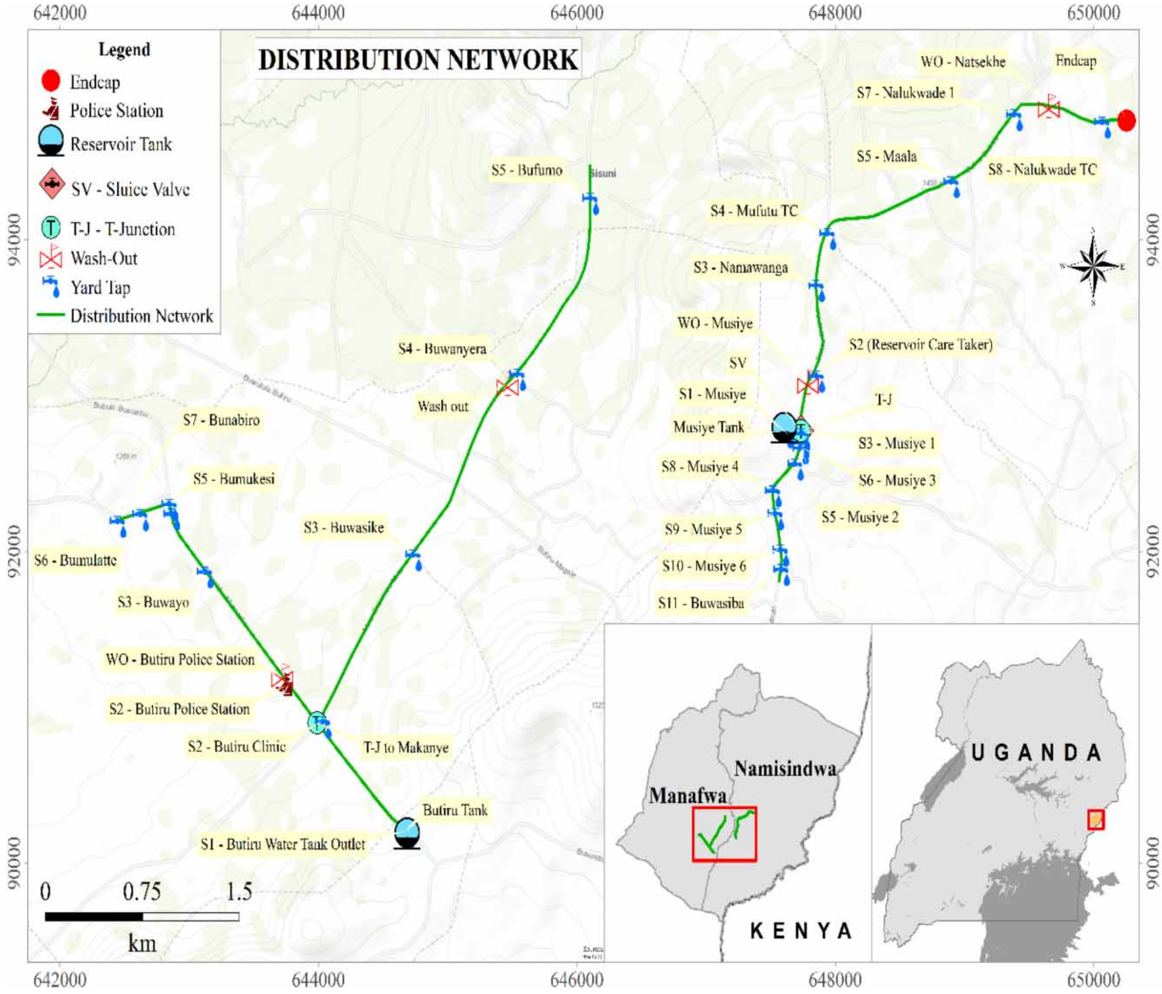

Figure 2 | Water sampling points (yard taps) on the Lirima gravity scheme.

the range of 0–6 mg/l and turbidity (*NTU*), A pH and conductivity 901 digital metre was used to measure online temperature (°C), pH and electrical conductivity (μS/cm). Test results of these water quality parameters were recorded in pre-prepared notebooks designed to record water qualities at geo-referenced positions in the water distribution system.

2.4. Model performance evaluation

For model evaluation, this study applied R^2, RMSE (root mean squared error) and MAE. The RMSE is a monotonic square root transformation of MSE (mean squared error) as shown in Equation (1).

$$\text{RMSE} = \sqrt{1}/N \sum_{n=1}^{N} (po - pp)^2 \tag{1}$$

where $N = sample\ size;\ po = observed\ outcome;\ pp = predicted\ outcome.$

Finally, the models were ranked based on interpretability, penalty and reward metrics.

2.5. Data analysis

EPANET 2.0 was used to develop a hydrologic model from the GPS coordinates picked from a water treatment plant, water outlets of break-pressure tanks, online wash-outs and other sections of transmission and distribution lines. EPANET 2.0 was further used to develop a process model of residual chlorine decay from upstream to downstream points within water transmission and distribution lines The influence of both water quality and water system parameters was investigated using triangulated methods of (1) decision tree analysis importance score, (2) random forest ensemble importance score, (3) principal component analysis equinox rotated matrix loadings and communalities and (4) *p*-values and standardized beta coefficients of independent variables in backward elimination in ordinary least squares (OLS) regression models. Python tree-based modules of decision trees and random forests were used to feature the importance of both physical and water quality parameters. Python and IBM SPSS software were used to analyse the correlation between physical and water quality parameters with residual chlorine and also regression analysis of these parameters with residual chlorine.

3. RESULTS AND DISCUSSION

This section presents the results of physical and statistical models in predicting residual chlorine decay in drinking water distribution systems. The physical process model used was the EPANET hydraulic analysis model. Statistical models used were decision tree, random forest, Lasso and ridge regularization models, ordinary least square multivariate regression model and supervised machine learning artificial neural network (ANN).

3.1. Descriptive statistics for physical and water quality parameters

Table 1 (descriptive statistics) shows how physical and water quality parameters are related to residual chlorine in the water distribution system.

Table 1 shows that the mean residual chlorine of 0.14 mg/l was below the lower limit of 0.2–0.5 mg/l specified by WHO (2014). The pH that ranged from 6.71 to 7.83 was within an acceptable range of 6.5–8.5 of US EAS 12 (Universal Standards of East African Standard 12) and Uganda National Bureau of Standards (UNBS, 2014).

Table 1 | Descriptive statistics for physical and water quality parameters in water distribution networks

Water quality and physical parameters	Count	Mean	Std	Min	25th percentile	Median	75th percentile	Max
Residual free chlorine (mg/l)	128	0.14	0.07	0.00	0.09	0.14	0.19	0.37
Distance (Km)	128	2.50	2.2	0.01	0.67	1.71	4.60	7.50
travel time (min)	128	46.13	42.63	5.00	15.00	30.00	65.00	190.00
Diameter (mm)	128	108.28	51.92	50.00	80.00	100.00	100.00	250.00
Turbidity (NTU)	128	0.96	0.77	0.00	0.75	1.07	1.07	5.00
Electrical Conductivity (μS/cm)	128	70.01	2.53	65.40	68.38	70.01	70.03	78.50
pH	128	7.53	0.17	6.71	7.48	7.53	7.60	7.83
Temperature (°C)	128	23.98	1.06	20.10	23.59	23.98	24.31	27.05
Pressure (Bar)	128	2.00	1.08	0.00	1.73	2.00	2.00	6.00
Velocity (m/s)	128	0.04	0.02	0.001	0.02	0.04	0.05	0.10

3.1.1. Effect of pressure on the performance of EPANET

The traditional design of most water distribution systems uses the DDA (Demand Driven Analysis) model because it assumes that nodal demands are known and are independent of nodal pressures (Ricca *et al.* 2019). However, this study used the PDA (Pressure Driven Analysis) model to calibrate the EPANET hydraulic model because nodal demand depends on nodal pressure (Oladipupo *et al.* 2019). This approach was used to avoid errors in the quantities of water delivered at each water sampling point. The quantity of online residual chlorine determined at each sampling point depends on quantities of water at that sampling point which is the medium through which residual chlorine is carried. Pressure is a hydraulic transient that influences significantly residual chlorine decay in drinking water distribution systems (Kim *et al.* 2014; Rajasingham *et al.* 2020). This is especially so under unsteady flow conditions (Goyal & Patel 2015). The importance of nodal demand pressure, which is one of the aims of drinking water distribution (Environmental Protection Agency 2018; Mentes *et al.* 2020), was demonstrated when residual chlorine decay in water distribution responded well to PDA (Pressure Demand Analysis) model compared to DDA model in extended EPANET simulation analysis (Melkumova & Shatskikh 2017).

3.2. Physical and water quality parameters

Table 2 (correlation matrix) shows how physical and water quality parameters are related to residual chlorine in a water distribution network.

Table 2 | Correlation matrix of chlorine decay with physical and water quality parameters

Parameters	RC	IC	dist	tt	dia	tur	EC	pH	temp	pre	vel
RC (mg/l)	1										
IC (mg/l)	0.69	1									
Dist (Km)	−0.11	0.30	1								
tt (min)	−0.08	0.31	0.71	1							
Dia (mm)	−0.09	0.20	0.63	0.52	1						
tur (NTU)	−0.02	−0.07	−0.03	0.05	0.08	1					
EC (μSiem^{-1})	−0.20	−0.05	−0.05	−0.06	−0.10	−0.29	1				
pH	0.15	0.11	0.14	0.17	0.12	0.03	−0.11	1			
temp (°C)	−0.21	−0.21	−0.05	−0.002	0.05	−0.10	0.21	−0.06	1		
pres (Bar)	0.03	0.19	0.31	0.17	0.02	−0.11	−0.15	0.17	0.23	1	
vel (m/s)	−0.17	−0.07	−0.03	−0.004	−0.06	−0.01	−0.15	0.20	−0.001	0.32	1

RC = residual free chlorine, IC = initial chlorine, dist = distance, tur = turbidity, EC = electrical conductivity, temp = temperature, pre = pressure, vel = velocity, tt = travel time, dia = diameter.
Source: (Kwio-Tamale 2022).

Table 2 shows that turbidity, temperature and pH were less influential. This result is consistent with the findings of (Bowden *et al.* 2019). The exclusion of turbidity as a predictor in modelling residual chlorine decay is a good measure as it is inappropriate because it does not account for dissolved chlorine reactants (Wu & Dorea 2020).

Velocity was also non-influential. This was probably because of the low velocities. Although high velocities as hydraulic transients influence residual chlorine, low values have little or no impact on residual chlorine decay (Kim *et al.* 2014). The velocity in the main pipes as measured at water demand nodes during this study varied from 0.001 to 0.10 m/s with a standard deviation of 0.02 m/s (Kwio-Tamale 2022). Pearson's correlation coefficient was very low at $r = 0.17$ at a 5% level of significance. High velocities are known to induce fast residual chlorine decay rates (Kim *et al.* 2014). Therefore, the minimal and insignificant influence of velocity on residual chlorine decay in this study was most probably because of low velocities. This agrees with a reported velocity of 0.56 m/s which had no influence on residual chlorine decay in an experimental study by (Kim *et al.* 2014).

The pressure at demand nodes during this study varied from a minimum of zero bars to a 75th percentile (third quartile) value of 2 bars with a mean of 2 bars, standard deviation of 1.08 bars and very low Pearson's correlation coefficient of $r = 0.03$ at 5% level of significance (Kwio-Tamale 2022).

3.3. EPANET hydraulic model

The architecture, structure and functionality of EPANET depend on physical and water quality parameters. The relevant physical parameters include pipe size, pipe roughness and pipe length. Equally, the relevant water quality parameters include initial chlorine concentration. Hydraulic parameters were velocity and pressure. The traditional pseudo-first-order single reactant model was used instead of the parallel double reactant model. The decay equation for the first-order single reactant kinetic model is as in Equation (2):

$$C_t = C_o.\exp(-K_b t) \tag{2}$$

where

C_o = initial concentration of chlorine, (mg/l)
C_t = concentration of chlorine at time t, (mg/l)
K_b = bulk reaction coefficient of chlorine, (hr^{-1} or day^{-1})
t = time (Hr. or Day)

The traditional pseudo-first-order single reactant model (Tiruneh *et al.* 2019b) was chosen because of (1) its popularity and wide use (Tiruneh *et al.* 2019b), (2) its simplicity and reasonable accuracy in representing chlorine decay in water systems and (3) performance benefit of higher orders over first-order model being marginal (Goyal & Patel 2015). The reason for choosing a single reactant model is the very low ratio of fast to slow reactants in the order of (10:1,000) (Tiruneh *et al.* 2019b) which is only 1%. In real practice, a mixture of slow and fast chemical reactants occurs in water distribution systems starting with faster and less concentrated organic reactants followed by slower more concentrated inorganic reactants (Tiruneh *et al.* 2019b). Therefore, each biochemical reactant group varies from another group in both reaction rates and order of kinetic reactions. This reality supports the current debate on insufficient knowledge about process modelling for residual chlorine decay (Bowden *et al.* 2019; Tiruneh *et al.* 2019b).

3.3.1. Performance of EPANET in predicting space–time decay of residual chlorine

The EPANET model was calibrated by measured flows and cross-validated by chlorine concentrations at demand nodes. The purpose of this was to ensure that the calibrated EPANET model carries chlorine residuals. Table 3 shows EPANET hydraulic model calibration and validation for three different water zones in the Lirima distribution system.

Table 3 | EPANET hydraulic model calibration and validation by water flows at demand nodes

Demand model			**Chlorine model**		
	Calibration statistics			**Cross-validation statistics**	
Calibration data	**Mean error (m^3/d)**	***R^2***	**Test data**	**Mean error (mg/l)**	***R^2***
(a) Musiye-Nalukwade water distribution zone					
Day1_Run3 (Evening) Tuesday 9th Feb 2021	1.749	0.997	Day8_Run1 (Morning) Wed 24th Feb 2021	0.168	0.722
(b) Butiru-Manyeke water distribution zone.					
Day5_Run1 (Afternoon) Tuesday 23rd Feb 2021	0.0001	0.999	Day2_Run1 (Afternoon) Wed 10th Feb 2021	0.002	0.997
(c) Butiru-Vermiculite water distribution zone.					
Day3_Run2 (Afternoon) Thursday 11th Feb 2021	0.653	0.999	Day7_Run1 (Morning) Tue 23rd Feb 2021	0.773	0.937

Table 3 shows that the range of R^2 from 0.72 to 0.98 and mean error from 0.002 to 0.773 mg/l for all three water zones in calibrating and validating the EPANET model was satisfactory. Table 4 presents the performance metrics of EPANET in modelling residual chlorine decay in three different water zones and combined water zones of the Lirima water distribution network. The R^2 registered for the combined water zone with the highest sampling points was 0.24.

Table 4 | Performance of EPANET in modelling free chlorine decay in different water zones

			Performance metrics	
Item	**Water zone**	**Sample points**	**RMSE (mg/l)**	**MAE (mg/l)**
1	Musiye	91	0.43	0.63
2	Vermiculite	45	0.87	1.02
3	Manyeke	10	0.00	0.00
4	Combined	146	0.58	1.02

RMSE = Root Mean Square Error, MAE = Maximum Absolute Error.

The average performance error some scholars argue to be the acceptable range for good EPANET performance is 0.1–0.2 mg/l (Fisher *et al.* 2011). Although Table 3 shows that the EPANET hydraulic model was well calibrated and validated, it did not perform well enough in simulating chlorine residual, as shown in Table 4.

Therefore, Table 4 shows that the size of the datasets used clearly influenced the analysis results. Small datasets performed poorly.

3.3.2. Effect of data size on the performance of EPANET

This poor performance is attributed to limited datasets (data points). The performance of EPANET shows that the complexity in hydraulic modelling of water quality increases with the increasing density and size of the water distribution network. Each of the parameters, such as the bulk reaction coefficient of a hydraulic model, can be different for the separate sub-distribution networks or water zones. In such cases, the use of a single value of the same parameter when the various sub-distribution networks are combined most probably reduces EPANET's hydraulic model performance. Again, in such cases, statistical models expectedly perform better than physical models given that they eliminate the need for focusing on the details of the chlorine decay process or mechanisms.

An important note is that water was sampled on average at 1-Km intervals to ensure notably observable changes in chlorine residual concentrations. This was because the average initial residual chlorine concentrations at the water treatment plant were low ranging from 0.70 to 1.0 mg/l with a mean dosage of 0.76 mg/l. Apart from limiting the upper limit of residual chlorine range to 1.0 mg/l possibly to minimize operational costs, it could have also been to safeguard against the formation of harmful carcinogenic DBPs (Disinfection By-Products) (Goyal & Patel 2015). This is because excess chlorination above the WHO-recommended minimum of 0.2 mg/l gives rise to DBPs (Disinfection By-Products) (Tiruneh *et al.* 2019a). Besides, limiting terminal residual chlorine to under 1.5 mg/l enhances water acceptability (Rajasingham *et al.* 2020). Although the data augmentation technique was used to increase on data size, the wide spacing of water sampling points resulted in few data points. A repeated sampling at the same sampling points on different days and times (mornings, afternoons and evenings), as shown in Table 3(a)–3(c), ensured that the varied hydraulic parameters (pressures, velocities) and varied water quality parameters (turbidities, electric conductivities, temperatures, pH, etc.) were captured to increase data size and variability in data points. This was inappropriate for physical (process) modelling like EPANET hydraulic modelling which requires large data (Tiruneh *et al.* 2019a).

3.4. Statistical models

Dimensionality control by the reduction of less important explanatory variables resulted in better model performance, as shown by a better MSE as low as 0.02 (Bowden *et al.* 2019). This translates into an interpretable monotonic square root transformation of MSE to become RMSE = 0.14 mg/l.

3.4.1. Decision tree regressor model

The model in Table 5 presents the training and test performance of the decision tree regressor model.

Table 5 | Performance metrics of the decision tree model for free chlorine decay

		Performance metrics		
Item	**Dataset**	**RMSE (mg/l)**	**MAE (mg/l)**	R^2
1	Training	0.03	0.01	0.41
2	Test	0.03	0.01	0.41

RMSE = root mean square error, MAE = maximum absolute error.

3.4.2. Random forest regressor model

The model in Table 6 presents the training and test performance of the random forest regressor model.

Table 6 | Performance metrics of the random forest model for free chlorine decay

		Performance metrics		
Item	**Dataset**	**RMSE (mg/l)**	**MAE (mg/l)**	R^2
1	Training	0.02	0.02	0.41
2	Test	0.05	0.04	0.55

RMSE, root mean square error, MAE, maximum absolute error.

The decision tree regressor model in Table 5 is overfitted compared to the random forest regressor model, as shown in Table 6. Since the random forest is an ensemble of decision trees as base models to minimize individual decision tree overfitting, the random forest model performed expectedly better. This conforms to domain knowledge of random forest's better performance than that of decision trees.

3.4.3. Regression models

The results for two MLR models for residual chlorine decay in water distribution are presented. These are results from (1) the OLS (Ordinary Least Squares) MLR model and (2) the PCA (Principal Component Analysis) MLR model. The model in Table 7 presents the OLS MLR model based on statistically significant variables that are largely independent of each other, as supported by their respective low VIFs (Variable Inflationary Factors).

Table 7 | Ordinary least squares linear regression model for statistically significant variables

Model	**R**	**R Squared**	**Adjusted R Squared**	**Std. error of estimate**		
(a) Model summary						
2	0.79[a]	0.63	0.62	0.045		
a. Predictors: (1) Constant, (2) initial chlorine, (3) distance, (4) EC Durbin-Watson = 1.26						
	Unstandardized coefficients		Standardized coefficients		Collinearity statistics	
(b) Model coefficients						
Predictors	B	Std. Error	Beta	t	Tolerance	VIF
Constant	0.42	0.112		3.69		
initial chlorine	0.55	0.040	0.80	13.76	0.000	0.906
Distance	−0.01	0.002	−0.37	−6.25	0.000	0.907
EC	−0.01	0.002	−0.18	−3.18	0.002	0.995
Dependent Variable: final chlorine EC = electrical conductivity VIF = Variable Inflationary Factor						

Multicollinearity was statistically insignificant ($p > 0.05$). The final residual chlorine concentration in the water distribution network, based on statistically independent and statistically significant water quality and water distribution system parameters, is described by Equation (3).

$$fc = 0.415 + 0.548ic - 0.012d - 0.005ec \qquad (3)$$

where fc = *Final free chlorine (mg/l),* ic = *Initial free chlorine (mg/l),* d = *Distance (Km) and* ec = *Electrical conductivity (µS).*

Table 8 compares two regression models of (1) OLS and (2) PCA (principal component analysis) based on the three most statistically significant predictors of (1) initial chlorine, (2) electrical conductivity and (3) distance that explain over 90% variability in residual chlorine decay in water distribution.

Table 8 agrees with Bowden *et al.* (2019) whose regression model found that initial chlorine together with pipe diameter and temperature accounted for 73%–87% of residual chlorine decay in water distribution.

3.4.4. ANN model

The performance of the ANN for residual chlorine in gravity water distribution is summarized in Figure 3. The outputs and targets in Figure 3 are the predicted and actual residual chlorine concentrations, respectively. The R-score values were 0.93, 0.68 and 0.94 for training, validation and test datasets, respectively.

The superior performance of ANN over MLR in this study was consistent with earlier findings of (Bowden *et al.* 2019). This further demonstrates the general superiority of non-parametric models over parametric models.

3.5. Comparison of physical and statistical models

Table 9 ranks models based on the performance metrics. More than one performance metric was used to benefit from the unique advantages of each metric in order to capture various aspects of model performance, as advanced by (Oladipupo *et al.* 2019). The ranking also considered how interpretable a model is (Oladipupo

Table 8 | Performance of linear and principal component regression models

		Regression models		
Item	Model characteristic statistics	OLS linear regression model	Principal component regression model	Deviation from each other
1.0	**Model summary**			
1.1	Pearson's coefficient, *r*	0.793	0.788	0.005
1.2	R^2	0.623	0.620	0.003
1.3	Adjusted R^2	0.619	0.611	0.008
1.4	Standard error of model estimate (mg/l)	0.0453	0.0456	0.0003
2.0	**ANOVA statistics**			
2.1	F-score	69.250	67.480	1.770
2.2	Model significance (*p*-value)	0.000	0.000	0.000
3.0	**Collinearity statistics (VIF)**			
3.1	Electrical conductivity	1.005	1.000	0.005
3.2	Distance	1.103	1.000	0.103
3.3	Initial chlorine (Chlorine dose)	1.004	1.000	0.004

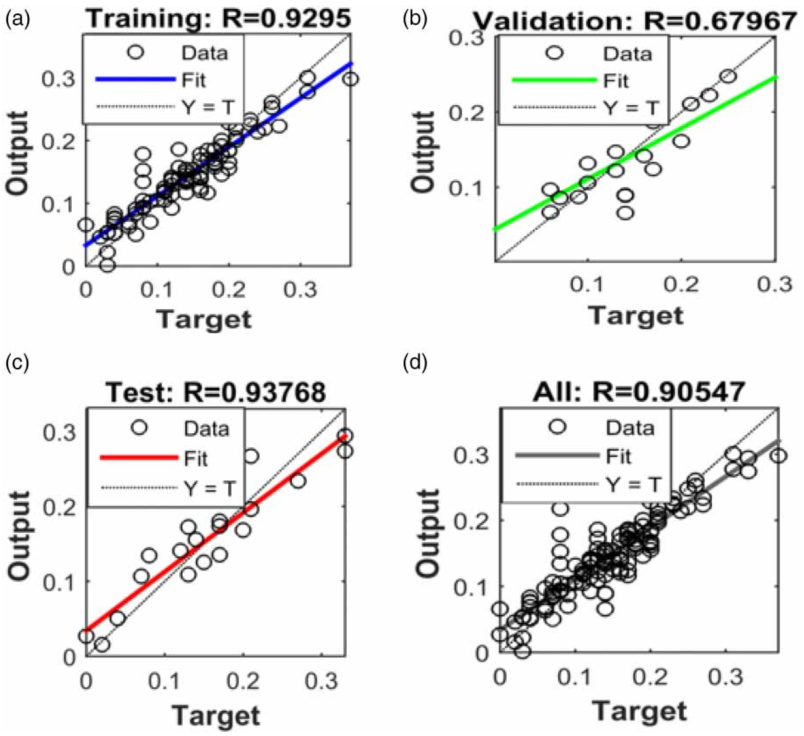

Figure 3 | Training and validation performance of the artificial neural network model for residual chlorine. Y = output: predicted residual chlorine (mg/l); T = target: actual residual chlorine (mg/l); All = whole (unsplit) dataset.

et al. 2019). The statistics for the multi-linear regression model were those contained in Table 7 (the linear regression model for statistically significant parameters of chlorine decay in the water distribution network).

Table 9 shows that the ANN performed best because of having the highest R^2 of 0.94, least RMSE of 0.04 mg/l and second lowest MAE of 0.05 mg/l (just above the lowest MAE of 0.04 mg/l). Besides, the test R^2 of ANN of 0.94 as in Figure 3(c) was more than its training R-score of 0.93, as shown in Figure 3(a). This suggests that the ANN model did not overfit. However, ANNs are difficult to interpret. The contribution of each parameter in predicting a target variable like final chlorine is important. This is not the case with ANN because it is a 'blackbox'

Table 9 | Comparison of models in predicting free residual chlorine

	The goodness-of-fit statistics		Performance accuracy		
Model	**Adjusted R^2**	**Std/error (mg/l)**	**RMSE (mg/l)**	**MAE (mg/l)**	**Rank of model**
1. EPANET	0.24		0.43	0.63	7
2. Linear regression	0.62	0.045	0.18		2
3. Lasso regression			0.06		6
4. Ridge regression			0.05		5
5. Decision tree	0.41		0.05	0.04	4
6. Random forest	0.55		0.05	0.04	3
7. Artificial neural network	0.94		0.04	0.05	1

RMSE, root mean square error; MAE, maximum absolute error.

unlike interpretable hedonic parametric models like linear regression. The **SHAP** (**SH**apley **A**ddtive ex**P**lainer) utility function that has been developed to 'whitewash' blackbox model outputs by explaining the importance of parameters still has a major weakness. SHAP explainer only presents feature importance values as in tree-based (decision tree and random forest) and ANN regressor models. However, the desirable analytical relationship in terms of algebraic expressions between each blackbox model input variable and blackbox model output variable is lacking. This is a major weakness and therefore, a research gap. In this regard, hedonic parametric linear regression models that express the relationship of each independent variable to the dependent variable by individualized distinct beta coefficients are better. Although decision trees and random forest regressors performed well, they too are non-interpretable non-parametric models.

4. CONCLUSION

The values of RMSE for EPANET, linear regression, lasso regression, ridge regression, decision tree, random forest and ANN were 0.43, 0.18, 0.06, 0.05, 0.05, 0.05 and 0.04 mg/l, respectively. The corresponding R^2 values for ANN, multi-linear regression, EPANET, random forest and decision tree were 0.94, 0.62, 0.24, 0.44 and 0.41, respectively. These results illustrate that the ANN performed better than EPANET. Therefore, this result supports the use of a supervised machine learning (deep learning) ANN as a primary tool for predicting residual chlorine decay in water distribution systems. However, non-parametric models like artificial neural networks that perform well in model prediction are poor in model interpretation. These results illustrate that statistical models, especially the ANN performed better than EPANET especially when the density and size of the water distribution network are large. However, since mechanistic models predict better than empirical models, EPANET, which is a mechanistic (physical) model, should be used in small water zones for secondary chlorination because of more uniform parameters like temperature, hence a more reliable decay constant in an exponential kinetic model in smaller water distribution networks. For empirical (statistical) models, a dual approach of (1) model performance metric and (2) model interpretability is necessary for choosing and deploying prediction models. Such a dual approach in empirical (statistical) prediction modelling will enable non-parametric artificial neural networks as models with the best model performance metric but poor interpretability to be enhanced with parametric regression models that have good model interpretability but inferior model performance relative to artificial neural networks. The study was limited to EPANET as the only process-based model because of its popularity in tracking dissolved chemical constituents like residual chlorine in a Lagrangian manner. However, other water quality modelling tools like WaterCAD, WaterGEMs, OpenFlows, Mike, Pipe, Civil Designer, etc. exist in the water industry. Therefore, future research is necessary to compare these other process models against the six statistical models used in predicting residual chlorine decay in drinking water distribution systems. The study was also limited to the dry period of February–March during which data were collected. Therefore, a study during the wet season is recommended to compare how residual chlorine decay occurs in wet spells.

FUNDING

This study was financed through an award to the first author in the phase of the third competitive research grant from Kyambogo University, Uganda, under the support of the Government of Uganda.

AUTHOR CONTRIBUTIONS:

J.C.K.-T. contributed to conceptualization, formal analysis, investigation, resources, data curation, writing- original draft, review & editing and visualization. C.O. contributed to conceptualization, formal analysis, investigation, resources, data curation, supervision – review & editing and visualization.

ACKNOWLEDGEMENTS

We are grateful to the National Water and Sewerage Corporation for their kind permission and support staff in conducting this study on their Lirima Gravity Flow Scheme. The authors acknowledge that this paper was partially based on the first author's MSc dissertation 'Comparison of physical and statistical models in predicting space–time decay of residual chlorine in water distribution system' submitted in 2022 for the award of MSc in Water and Sanitation Engineering of Kyambogo University.

DATA AVAILABILITY STATEMENT

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

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Azad, A., Karami, H., Farzin, S., Mousavi, S. F. & Kisi, O. (2019) Modeling river quality parameters using modified adaptive neuro fuzzy inference system, *Water, Science and Engineering*, **12** (1), 45–54. https://doi.org/10.1016/j.wse.2018.11.001.

Bowden, G. J., Nixon, J. B., Dandy, G. C., Maier, H. R. & Holmes, M. (2019) Forecasting chlorine residuals in a water distribution system using generalized neural network. https://doi.org/10.1.1.519.8859.pdf.

Environmental Protection Agency. (2018) *Online Water Quality Monitoring Systems in Distribution Systems: For Water Quality Surveillance and Response Systems*. MC 140. Cincinnati, USA: Office of Water.

Fisher, I., Kastl, G. C. & Sathasivan, A. (2011) Evaluation of suitable chlorine bulk-decay models for water distribution systems, *Water Research*, **45** (2011), 4896–4908. https://doi.org/10.1016/j.watres.2011.06.032.

Goyal, R. P. & Patel, H. M. (2015) Analysis of residual chlorine in simple drinking water distribution system with intermittent water supply, *Applied Water Science*, **5**, 311–319. https://doi.org/10.1007/s13201-014-0193-7.

Hyunjun, K. & Sanghyun, K. (2017) Evaluation of chlorine decay models under transient conditions in a water distribution system, *Current Science*, **3** (8), 522–537. https://doi.org/10.2166/hydro.2017.082.

Karadirek, I., Kara, S., Muhammetoglu, A., Muhammetoglu, H. & Soyupak, S. (2015) Management of chlorine dosing rates in urban water distribution networks using online continuous monitoring and modelling, *Urban Water Journal*, **19** (4), 1–15. htpps://doi/10.1080/1573062X.2014.992916.

Kim, S., Kim, H. & Koo, J. (2014) 'Prediction of chlorine in various hydraulic conditions from a pilot-scale water distribution system', *12th International conference on computing and control for the water industry*, Vol. 70, 2014, pp. 934–942. Available at: https://doi.org/10.1016/j.proeng.2014.02.104

Kumar, A., Kumar, K., Bharanidharan, B., Neha, M., Eshita, D., Singh, M., Thakur, V., Sharma, S. & Malhotra, N. (2015) Design of water distribution system using EPANET, *International Journal of Advanced Research*, **3** (9), 789–812. Available at: https://www.journalijar.com.

Kwio-Tamale, J. C. (2022) *Comparison of Performance of Physical and Statistical Models in Predicting Residual Chlorine Decay in Drinking Water Distribution*. Master of Science, Water and Sanitation Engineering Dissertation, Kyambogo University.

Loucks, D. P. & van Beek, E. (2017) *Water Resource Systems Management and Planning: An Introduction to Methods, Models and Applications*, 2nd edn. Cham, Switzerland: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-44234-1.

Louppe, G. (2014) *Understanding Random Forests-From Theory to Practice*. PhD Thesis, Department of Electrical Engineering & Computer Science, Faculty of Applied Sciences, University of Liège. https;//arxiv.org

Melkumova, L. E. & Shatskikh, S. Y. (2017) 'Comparing ridge and lasso estimators for data analysis', *3rd Iinternational conference on information technology and nanotechnology*. Samara, Russia, April 2017, pp. 25–27.

Mentes, A., Galiatsatou, P., Spyrou, D., Samara, A. & Stournara, P. (2020) Hydraulic simulation and analysis of an urban centre's aqueducts using scemario analysis for network operations: The case of Thessaloniki City in Greece, *Water*, **12**, 1627. https://doi.org/10.3390/w12061627.

Oladipupo, B., Adnan, M. A., Hamam, Y., Page, P. R., Adedeji, K. B. & Piller, O. (2019) Solving management problems in water distribution networks: A survey of approaches and mathematical models, *Water 2019 MDPI*, **11**, 562. https://doi.org/10.3390/w11030562.

Rajasingham, A., Harvey, B., Taye, Y., Kamwaga, S., Martinsen, A., Sirad, M., Aden, M., Gallagher, K. & Thomas, H. (2020) Improved chlorination and rapid water quality assessment in response to an outbreak of acute watery diarrhea in Somali region, Ethiopia, *Journal of Water, Sanitation, Hygiene and Development*, **10** (3), 596–602. https://doi.org/10.2166/washdev.2020.146.

Ricca, H., Aravinthan, V. & Mahinthakumar, G. (2019) Modelling chloramine decay in full-scale drinking water supply systems, *Water Environment Federation*, 441–454. https://doi.org/10.1002/wer.1046.

Steurer, M., Hill, R. J. & Pfeifer, N. (2019) Metrics for evaluating the performance of machine learning based automated valuation models, *Journal of Property Research*, **38** (2), 99–129. doi:10.1080/09599916.2020.1858937.

Tiruneh, A. T., Debessai, T. Y., Bwembya, G. C. & Nkambule, S. J. (2019b) A mathematical model for variable chlorine decay rates in water distribution systems, *Hindawi Modelling and Simulation in Engineering*, **2019**. https://doi.org/10.1155/2019/5863905.

Tiruneh, A. T., Debessai, T. Y., Bwembya, G. C., Nkambule, S. J. & Zwane, L. (2019a) Variable chlorine decay rate modelling of the Matsapha town water network using EPANET program, *Journal of Water Resource and Protection*, **11**, 37–52. https://doi.org/10.4236/jwarp.2019.111003.

Torretta, V., Tolkou, A. K., Katsoyoyiannis, I. A., Katsoyoyiannis, A., Trulli, E., Magaril, E. & Rada, E. C. (2019) Consumption of free chlorine in an aqueduct with low protection: case study of the new aqueduct Simbrivio-Castelli (NASC), Italy, *Water 2018 MDPI*, **10** (127). https://doi.org/10.3390/w10020127. Available at: www.mdpi.com/journal/water.

Vuta, L. & Dumitran, G. E. (2019) Some aspects regarding chlorine decay in water distribution networks, *Aerul si Compaonente ale Mediului*, 253–259. Available at: htpps://www.researchgate.net>2284….

World Health Organization (2017) *Principles and Practices of Drinking-Water Chlorination: A Guide to Strengthening Chlorination Practices in Small-to Medium Sized Water Supplies*. Available at: https://creativecommons.org/licenses/by-nc-sa/3.0/igo.

Wu, H. & Dorea, C. C. (2020) Towards a predictive model for initial chlorine dose in humanitarian emergencies, *Water*, **12**, 1506. https://doi.org/10.3390/w12051506.

First received 23 March 2024; accepted in revised form 20 August 2024. Available online 10 September 2024