Proceedings Paper:
Kyritsakas, G., Speight, V. and Boxall, J. orcid.org/0000-0002-4681-6895 (2023) A data-driven model for the prediction of chlorine losses in water distribution trunk mains. In: IOP Conference Series: Earth and Environmental Science. 14th International Conference on Hydroinformatics, 04-08 Jul 2022, Bucharest, Romania. IOP Publishing , 012048.

https://doi.org/10.1088/1755-1315/1136/1/012048

A data-driven model for the prediction of chlorine losses in water distribution trunk mains

A data-driven model for the prediction of chlorine losses in water distribution trunk mains

Grigorios Kyritsakas[1], Vanessa Speight[1] and Joby Boxall[1]

[1] Department of Civil and Structural Engineering, The University of Sheffield, Mappin Street, Sheffield S1 3JD, UK

gkyritsakas1@sheffield.ac.uk

Abstract. A data-driven model that uses 4 different machine learning (ML) algorithms (Feed forward artificial neural networks (ANN), Nonlinear autoregressive exogeneous (NARX) ANN, support vector machine and Random Forest) was designed for the prediction of chlorine loss events in water distribution trunk mains. The model, firstly, identifies past chlorine loss events and their associate flow or temperature events. Then, the detected past flow events and their associate past chlorine loss events are used to train the ML algorithms. The model was tested in 3 trunk mains of the same drinking water distribution system with similar diameter but with different characteristics, using each time a different combination of parameters (flow (input) - past chlorine losses (output) or flow, temperature, and chlorine (input) - past chlorine losses (output)) and machine learning algorithms. Results indicate that the model could predict a future chlorine loss event with a period between 2 to 10 hours depending on the parameter and ML algorithm used and the trunk mains' hydraulic characteristics.

1. Introduction

Chlorine (Cl) is the most common disinfectant used to minimise the bacteriological activity in the drinking water exiting the treatment processes and prevent bacterial regrowth in the water distribution systems (WDS). Chlorine concentration in the water exiting the works is particularly high; however, due to chlorine reactions in the bulk water and in the pipe walls, it decreases during the water travel to the customer taps through the WDS [1]. Therefore, it is required by the water utilities to monitor the chlorine residual and, if possible, to accurately predict the chlorine behaviour in the WDS.

Traditional approaches in predicting chlorine decay in the WDS are made using a combination of processed based models and hydraulic models. Generally, these models produce accurate results; however they require a large number of accurate data and a good understanding of the physical chemical and hydraulic characteristics of the WDS for their implementation. In addition, these models require a determination and a continuous recalibration of a number of parameters (e.g. chlorine decay constant) in order to accurately simulate the chlorine behaviour. This process could be complicated and sometimes it requires a vast amount of computational time for its completion.

Data-driven models that apply machine learning (ML) algorithms for understanding the trends over the collected data and predict future behaviour could be an alternative solution for the prediction of chlorine decay in the WDS. In the past, data-driven models have been applied in various projects in the water sector for analysing the quality of the source water [2], improving the water treatment processes [3] and predicting turbidity events in trunk mains [4, 5]. As regards the chlorine decay

prediction, Gibbs et al. presented a promising data-driven model using artificial neural networks (ANNs) to predict chlorine concentrations in customer taps [6].

In this paper, the model that Kazemi et al. [4] developed for the prediction of turbidity events in the water distribution trunk mains is redeveloped and adapted to predict chlorine loss events. More specifically, the model, presented here, is a regression model that predicts chlorine loss events, in terms of periods with low chlorine concentrations, at the end of the water distribution trunk mains up to several hours ahead. The model initially detects past chlorine loss events in the dataset and then it uses these events in combination with past water quality data related to these events (temperature, flow and chlorine concentration) as inputs in 4 ML algorithms (Feed Forward ANN (FF-ANN), non-linear autoregressive exogeneous ANN (NARX-ANN), random forest (RF) and support vector machines (SVM)) for the prediction of future chlorine loss events. The model was tested in 3 real world trunk mains of the same WDS and the prediction performance of the various ML algorithms was compared.

2. Site details and available datasets

The datasets used in this work were created for a case study on the impact of hydraulic interventions for managing the discolouration risk in the trunk mains [7]. The study area (figure 1) is a WDS located in the north of the UK consisting of 3 different trunk mains (TM-1, TM-2, TM-3) fed with water from the same water treatment plant. Chlorine concentrations were monitored with a frequency of 15 minutes at the plant outlet and at the outlet of each one of the trunk mains. Water temperature was also measured at the end of each trunk main, and flow was measured at the start of each trunk main with the same frequency. The main pipe characteristics for all the trunk mains are similar (table 1), however, during the study period, their hydraulic characteristics differed for the purposes of the case study [7].

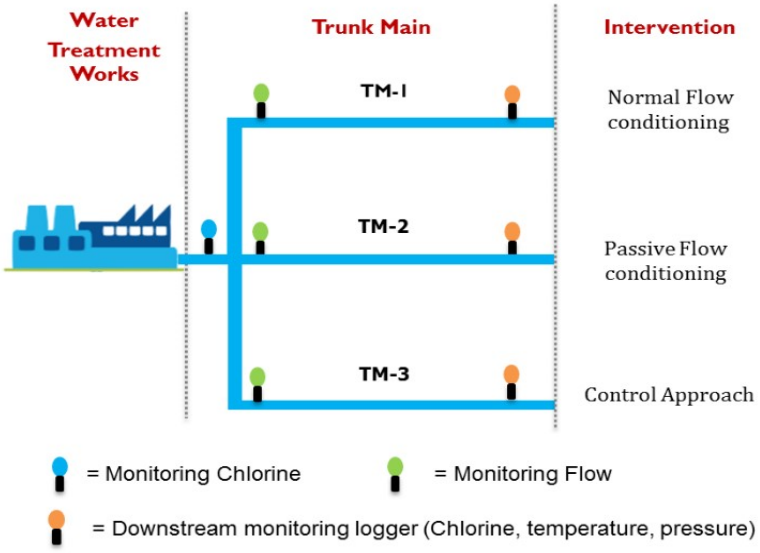

Figure 1. Schematic of the DWDS trunk mains [7].

Table 1. Water distribution trunk mains characteristics.

Trunk main	Mean internal diameter [mm]	Pipe material	Velocity [m/s]	Length from WTW outlet to downstream logger [km]
TM-1	304.8	Unlined CI (25% lined)	0.6	6.4
TM-2	406.8	Unlined CI	0.3	5.6
TM-3	304.8	Unlined CI	0.4	5.9

For our work, we assumed that there is no leakage in the trunk mains, and therefore for each trunk main and for each time step there were no changes in the monitoring flow during the travel of the water through them. Overall, three different datasets, one for each water distribution trunk main, are collected. Each dataset consists of 15 minutes time chlorine, flow and temperature data for a period of 7 months. During this period, there are some months with no available chlorine or temperature data that are excluded from the analysis.

3. Methodology

3.1. Data preparation

3.1.1. Removing data anomalies and outliers. In the collected datasets there were several outliers, either spikes or zero and negative values that could influence the methodology's results. These were identified and removed using a gradient algorithm created in MATLAB (version 2019b). More specifically, the algorithm followed these steps:

- It identified and replaced the zero and the negative chlorine values with missing values;
- It calculated the gradient between a data point and its previous one and if this was found greater than 0.4 mg/l it was replaced with a missing value;
- If the consecutive missing values were more than 12 (3 hours without measurement, the points were ignored, otherwise these were filled using the spline interpolation.

3.1.2. Smoothing the data. The last data preparation step was to smooth the data for removing the noise that could influence the training of the data-driven methodology. For the smoothing of the dataset the cubic spline function was used in MATLAB (version 2019b). The smoothing length was set equal to $2\Delta t$ (30 minutes) due to the increased noise in the Chlorine dataset. Figure 2 shows the original and the smoothed Cl, flow, and temperature data in a small part of the TM-3 dataset.

3.2. Chlorine loss event detection

A model was designed to capture and extract the chlorine loss events from the dataset. These events are identified as the periods of the dataset where unusually low chlorine concentrations are measured. The model extracted the chlorine loss events by following these steps:

- The event peaks were identified as the local minimum Chlorine values;
- For each peak the event starting and finishing points were detected using the gradient of the timeseries before and after the event. The model traced the data up to certain hours before the event to find a gradient that surpassed the gradient threshold of 0.02mg/l per 15 min (0.02mg/l per timestep). If no gradient higher than the threshold was found the peak was ignored. If a gradient higher than the threshold was found, the chlorine event was detected. The detected event consisted of equal datapoints prior the peak and after the peak depending on the defined forecasting period. For example, if the forecasting period was set as 4 hours ahead, the event consisted of datapoints 2 hours before and 2 hours after the peak;
- A "base value" was set as equal to the pre-event chlorine concentration in the dataset and baseline that connected all the base values was generated;
- The event magnitude was calculated as the difference between the peak and the base value of each event. The final events were then selected as those that exceeded the magnitude threshold of -0.15mg/l.

3.3. Flow and temperature event detections

Once the final chlorine events were detected a similar model was used to detect their associate flow or temperature events. These events were detected by searching the dataset over a period of certain hours before the detected chlorine event. This detection model identified the associate flow or temperature events by following 5 steps. The first four were same steps as the ones that the chlorine loss event

detection model was using. The magnitude threshold in the fourth step was set equal to 10l/s ($0.01m^3/s$) when the flow dataset was used and to 0.3°C when the temperature dataset was used. The final step in this model was to remove all the Chlorine loss events that no flow event or no temperature event was found.

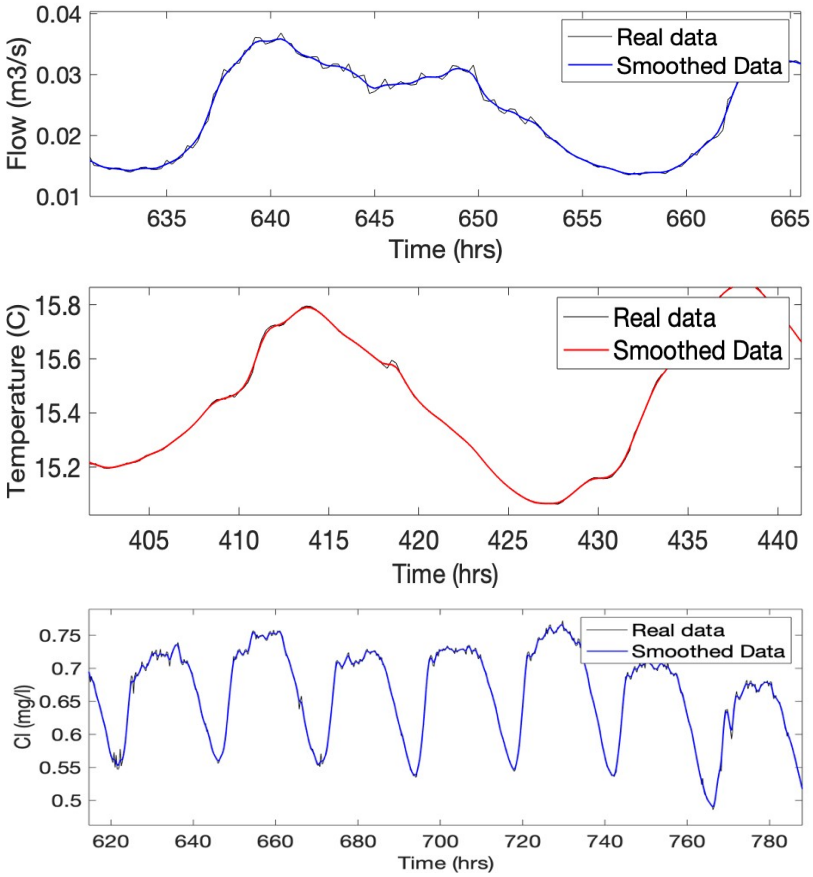

Figure 2. Original and smoothed flow (top), temperature (middle) & chlorine (bottom) data in TM-3.

The model was applied in all three trunk mains and on average there were 30 chlorine loss events associated with high flow events per trunk main in comparison with 12 chlorine loss events associated with temperature events. Therefore, the events used in the predictive model were chlorine loss events as outputs and their associate flow events as inputs. Figure 3 shows two detected chlorine events with their associated high flow events for TM-1.

3.4. Predictive model

A multistep data-driven prediction model was constructed in MATLAB (version 2019b) for the prediction of future chlorine loss events. In the predictive model, 4 ML algorithms (RF, SVM, FF-ANN and NARX-ANN) were selected for the chlorine loss forecasting and a comparison between their performances was made.

3.4.1. ML algorithms. The selected ML algorithms tested here were as follows:

Random Forest (RF) [8]: RF is an ensemble decision tree algorithm that uses the outputs of large number of equal weak decision trees for making predictions. Each weak tree makes its splitting decision using randomly a small part of training dataset and part of the available input parameters. The final RF prediction is equal to the mean of the predictions made by the week trees. In this predictive

model the number of weak trees was set equal to 1000 and the minimum number observations per tree were set to 5.

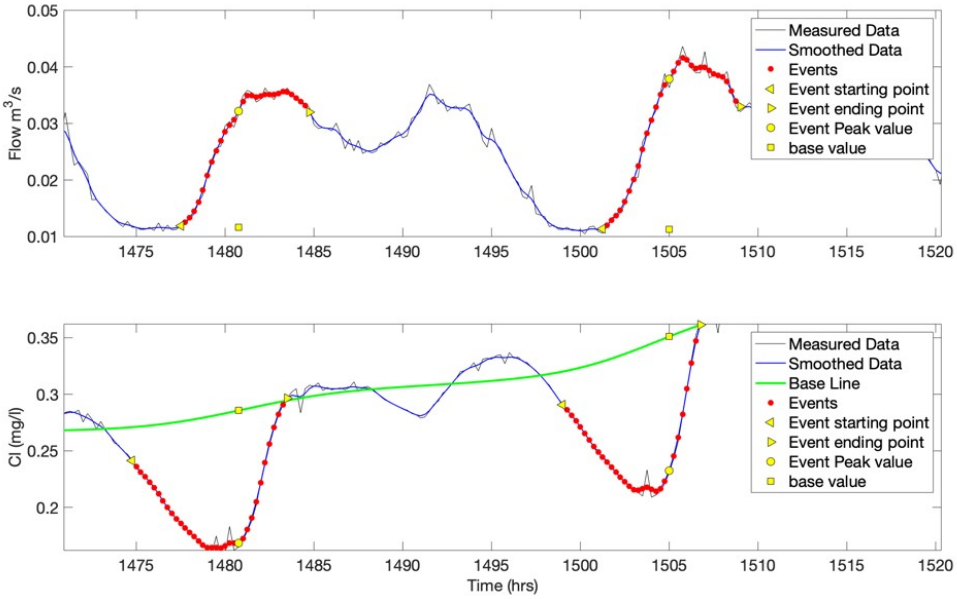

Figure 3. Detected chlorine and their associated flow events in TM-1.

Support vector machine (SVM) [9]: SVMs is trained to find the maximum margins between the boundaries that separate the input dataset and thus reducing the generalisation error. In this predictive model, the 5-Fold cross-validation was used for the data calibration and the Gaussian kernel for its training.

FF-ANN [9]: ANNs were constructed to simulate the human brain function. The FF-ANN is an ANN where the information is moving only forward from the input layer to the output layer through the hidden layers. The FF-ANN used here had one hidden layer with 14 neurons and it was using the regularization backpropagation function for training.

NARX-ANN [10]: NARX-ANN is an ANN that uses the non-linear version of the autoregressive exogenous model. There are two versions of the algorithm architecture, the open and closed loop. The former uses the past and presents inputs and the past targets for predicting the future targets and the latter uses past and present inputs and past targets and predicted past targets by the model for the prediction of future targets. The NARX-ANN selected here was using a hidden layer with 10 neuros and the input and feedback delays were set to 3 empirically.

3.4.2. Predictive model input variables. The model initially collected the past chlorine loss events and their associated flow events. Then, either the flow measurements or the flow with the temperature and the chlorine measurements during the flow event period were used as inputs in the predictive model and the chlorine loss events were used as targets (outputs).

The model is trained using a few past events combinations (10-20 events) to predict a future chlorine loss event given new unseen input data (flow or flow, chlorine, and temperature) during the current flow event. To remove the seasonality effect in the dataset, instead of the absolute chlorine values, the predictive model uses the magnitude of the event i.e. the difference between each data point of the event and the its base value. A simplified schematic of the predictive model that used 10 past chlorine events and all the inputs in their associated flow events is presented in figure 4.

3.4.3. Performance metrics. Three performance metrics were selected for quantifying the performance of the predictive model and justify which of the ML models was better, the mean absolute error

(MAE), the root mean squared error (RSME) and the Nash-Sutcliffe Model Efficiency Coefficient (NSE). MAE is used as an indication of the overall agreement between predicted and targeted values, RMSE is used as a metric for the identification of high errors in the prediction and NSE is used for the comparison of the predictive skills of the model with the mean of the observed data.

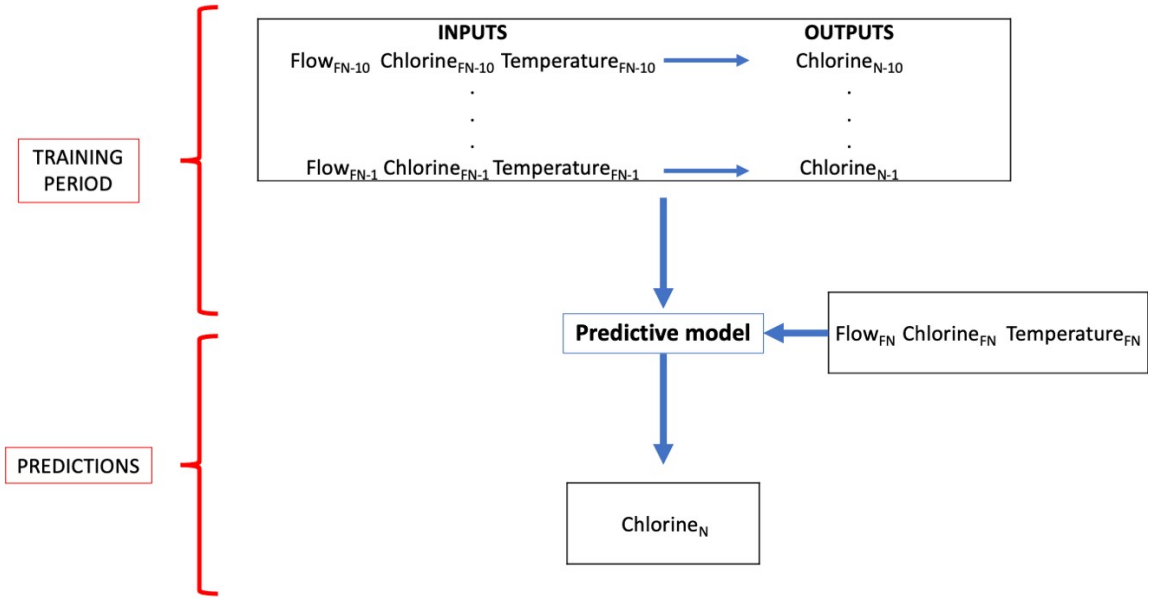

Figure 4. Schematic of the predictive model for the prediction of the Nth Chlorine event using for training the data from 10 past chlorine events (N-10…N-1) as outputs and all the inputs (chlorine, temperature, flow) from their associated flow events (FN-10…FN-1).

4. Results and discussion

The model was tested using 8 different combinations of ML algorithms and input variables (NARX-flow, NARX- flow/temperature/chlorine, FF-flow, FF - flow/temperature/chlorine, SVM-flow, SVM-flow/temperature/chlorine, RF-flow, RF-flow/temperature/chlorine) in each trunk main and for different forecasting periods (4-10 hours). In each trunk main different combinations of inputs and outputs were used for training and testing. Overall, the model managed to have accurate predictions with a forecasting horizon of up to 8 hours in TM-1 and up to 10 hours in TM-2 and TM-3. In table 2, the average performance results from all the model's tests in all 3 trunk mains are presented.

The model performance metrics as presented in table 2, indicate that the RF algorithm is the best out of all the tested ML algorithms and the NARX - ANN is the worst one. More specifically, NARX outputs produced only negative NSE results when tested in all three trunk mains while the RF, using flow, chlorine, and temperature as input variables, produced the highest NSE average and lowest MAE and RMSE average in TM-2 and TM-1. In addition, in one of the test events in TM-3, RF produced the highest NSE (0.92) of all simulations. This result disagrees with the finding that Kazemi et al found using the NARX algorithm with timeseries data in their work [4]. However, in their case their aim was to predict turbidity timeseries events instead of chlorine events.

Figure 5 shows a 9-hour event in TM-2 with 10 hours ahead forecasting period and figure 6 shows a 12-hour event in TM-3 with 10 hours ahead forecasting period. Both figures 5 and 6 shows that RF, FF and SVM follow the measured data changing trend in comparison with NARX which, especially in TM-3, produced completely different to the measured data outputs.

Regarding the other models, SVM with only the flow parameter used as input was the second-best predictive model in TM-1 and TM-2 with an NSE of 0.3 and 0.425 respectively. However, in TM-3 the SVM with all input parameters outperformed the SVM with flow input.

Table 2. Performance of the model using different ML algorithm-input combinations.

Trunk Main	ML Method	Prediction horizon	Parameters used	MAE	RMSE	NSE	Highest NSE
TM-1	RF	8	cl flow temperature	0.0205	0.02645	0.543	0.579
			flow	0.0275	0.03475	0.280	0.386
	SVM		cl flow temperature	0.02625	0.03375	0.274	0.54
			flow	0.02425	0.03125	0.300	0.5
	FF		cl flow temperature	0.02895	0.03075	0.296	0.65
			flow	0.03025	0.04025	-0.002	0.473
	NARX		cl flow temperature	0.2545	0.35175	-2.878	-1.62
			flow	0.07825	0.12355	-1.225	-1.09
TM-2	RF	10	cl flow temperature	0.0245	0.031	0.448	0.796
			flow	0.028	0.0365	0.32	0.327
	SVM		cl flow temperature	0.04	0.046	-0.1475	0.378
			flow	0.0245	0.033	0.425	0.591
	FF		cl flow temperature	0.0425	0.0465	0.403	0.74
			flow	0.0275	0.0355	0.3535	0.449
	NARX		cl flow temperature	0.145	0.196	-1.85	-1.8
			cl flow	0.1285	0.1925	-0.55	0.33
TM-3	RF	10	cl flow temperature	0.028	0.036	0.530	0.920
			flow	0.0385	0.048	0.308	0.569
	SVM		cl flow temperature	0.0273	0.035	0.587	0.644
			flow	0.033	0.046	0.395	0.440
	FF		cl flow temperature	0.119	0.138	-0.410	0.780
			flow	0.0395	0.0485	0.313	0.320
	NARX		cl flow temperature	0.164	0.1895	-1.865	-1.800
			flow	0.3985	0.4615	-1.350	-1.300

FF also produced good results, but the algorithm had completely different predictions in each TM depending on the inputs used. In TM-1 the FF algorithm when all the inputs were used, outperformed the one that used the flow as input in contrast with the other 2 TMs where the opposite occurred. Because all 3 trunk mains had different hydraulic characteristics during the study period, this finding indicates that both SVM and FF algorithms are influenced by the noise in the data due to the different hydraulic characteristics of the 3 trunk mains. In contrast, the predictive model that used the RF algorithm with either the flow or all the parameters as inputs, produced good outputs in all three trunk mains which indicates its ability to absorb the noise in the data. Moreover, in all the trunk mains the RF with all the outputs outperformed the RF where only the flow was used as input parameter which

potentially indicates that the more the relevant parameters are given to RF the better the results would be. Overall, this investigation demonstrated the potential of this model in predicting chlorine events in this dataset with a forecasting horizon of up to 10 hours ahead. However, further work is required in other water distribution networks with larger available datasets that cover a period of at least 2 years. This investigation will aim to examine how the model behaves in periods with different hydraulic conditions and its adaptation to seasonality changes.

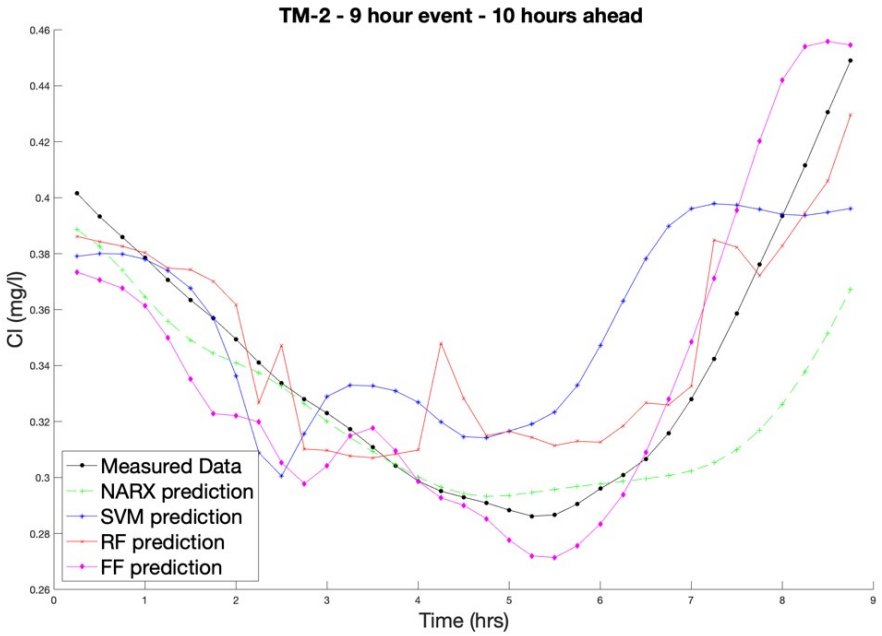

Figure 5. Chlorine predicted vs Measured data of a chlorine loss event in TM-2.

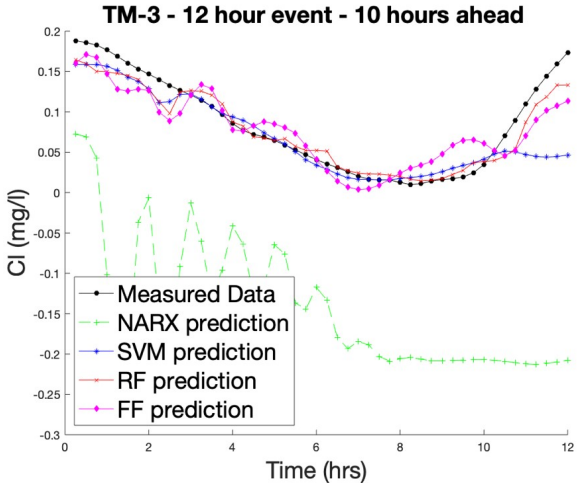

Figure 6. Chlorine predicted vs Measured data of a chlorine loss event in TM-3.

5. Conclusions

A data-driven methodology for the forecasting of a future chlorine loss event at the end of a water distribution trunk main is presented. This methodology, firstly, identifies past chlorine loss events and their associated flow events. Then, it imports either the flow, or the flow, the temperature, and the chlorine measurements during the flow event as inputs and the chlorine events as outputs in a predictive model that uses one of the following ML techniques: NARX, FF, SVM and RF. The

methodology was tested in a WDS that consists of 3 trunk mains (TM-1, TM-2, TM-3). The aim was to investigate the predictive ability of the model when different input parameters (flow or flow, temperature chlorine) - ML algorithms (NARX, SVM, FF, RF) combinations were used.

For the quantification of the model's performance, 3 main performance metrics were used, MAE, RMSE and NSE. The main conclusions are as follows:

- The predictive model managed to predict accurately a future chlorine loss event with a period of 8 hours ahead in TM-1, 10 hours ahead in TM-2 and 10 hours ahead in TM-3;
- The performance metrics (MAE, RMSE, NSE) indicated that best overall algorithm was the one that was using the RF algorithm with all the parameters as inputs;
- NARX was the worst algorithm in all three datasets for this predictive model;
- FF and SVM models had acceptable performances but both models appeared to be influenced by the different noises in the 3 datasets caused by the different hydraulic conditions in each trunk main;
- This predictive model has the potential of becoming an accurate supporting tool for supporting water utilities decision making regarding proactive intervention. However, as the available data for this work were taken for a small period (overall less than 5 months of data were available if we include the spikes and the missing data), further research is required using larger datasets - with at least 2 year of data - to investigate the ability of the model to adapt in the seasonality changes and different hydraulic conditions.

Acknowledgments
The authors would like to thank the Engineering and Physical Sciences Research Council (EPSRC), the UKRI and Scottish Water for providing the financial support through STREAM IDC and PODDS for the data provision and permission to use it in this work. For the purpose of open access, the authors have applied a creative commons attribution (CC BY) license to any author accepted manuscript versions arising.

References

[1] Parsons SA and Jefferson B 2006 *Introduction to Potable Water Treatment Processes* (Blackwell Publishing Ltd) pp 1–179 doi:10.1002/9781444305470
[2] Liu P 2019 Analysis and prediction of water quality using LSTM deep neural networks in IoT environment *Sustainability* **11**(7) 2058 https://doi.org/10.3390/su11072058
[3] Li L, Rong S, Wang R and Yu S 2021 Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review *Chem. Eng. J.* **405** 126673 https://doi.org/10.1016/j.cej.2020.126673
[4] Kazemi E, Mounce S, Husband S and Boxall J 2018 Predicting turbidity in water distribution trunk mains using nonlinear autoregressive exogenous artificial neural networks *Proc. 13th Int. Conf. Hydroinformatics* (1-4 July, Palermo, Italy)
[5] Meyers G, Kapelan Z and Keedwell E 2017 Short-term forecasting of turbidity in trunk main networks *Water Res.* **124** 67–76 doi: 10.1016/j.watres.2017.07.035
[6] Gibbs M et al 2006 Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods *Math. Comput. Model.* **44**(5-6) 485–498 https://doi.org/10.1016/j.mcm.2006.01.007
[7] Sunny I, Husband S, Moore G, Drake N, Mckenzie K and Boxall J 2017 Discolouration risk managememnt and chlorine decay *Proc. 15th Int. Conf. CCWI* (5-7 Sept., Sheffield, UK)
[8] Breiman L 2001 Random forests *Machine Learning* **45** 5–32 https://doi.org/10.1023/A:1010933404324
[9] Hastie T, Tibshirani R and Friedman J 2008 *The Elements of Statistical Learning* 2nd edition (New York: Springer)
[10] Lin T, Horne BG, Tino P and Giles CL 1996 Learning long-term dependencies in NARX recurrent neural networks *IEEE T. Neural Networ.* **7**(6) 1329-51 doi: 10.1109/72.548162