Implementation of an environmental decision support system for controlling the pre-oxidation step at a full-scale drinking water treatment plant

Lluís Godo-Pla, Pere Emiliano, Santiago González, Manel Poch, Fernando Valero and Hèctor Monclús

ABSTRACT

Drinking water treatment plants (DWTPs) face changes in raw water quality, and treatment needs to be adjusted to produce the best water quality at the minimum environmental cost. An environmental decision support system (EDSS) was developed for aiding DWTP operators in choosing the adequate permanganate dosing rate in the pre-oxidation step. To this end, multiple linear regression (MLR) and multi-layer perceptron (MLP) models are compared for choosing the best predictive model. Besides, a case-based reasoning (CBR) model was approached to provide the user with a distribution of solutions given similar operating conditions in the past. The predictive model consisted of an MLP and has been validated against historical data with sufficient good accuracy for the utility needs ($R^2 = 0.76$ and RSE = 0.13 mg·L^{-1}). The integration of the predictive and the CBR models in an EDSS gives the user an augmented decision-making capacity of the process and has great potential for both assisting experienced users and for training new personnel in deciding the operational set-point of the process.

Key words | case-based reasoning, drinking water, EDSS, modelling, multi-layer perceptron, oxidation

Lluís Godo-Pla
Manel Poch
Hèctor Monclús (corresponding author)
LEQUIA, Institute of the Environment,
University of Girona,
E-17003, Girona, Catalonia,
Spain
E-mail: *hector.monclus@udg.edu*

Lluís Godo-Pla
Pere Emiliano
Santiago González
Fernando Valero
Ens d'Abastament d'Aigua Ter-Llobregat (ATL),
Sant Martí de l'Erm, 30. E-08970 Sant Joan Despí,
Barcelona,
Spain

INTRODUCTION

Drinking water treatment plants (DWTPs) consist of a series of operation units that provide physical and chemical barriers against chemicals and pathogens that occur in raw water in order to supply potable water to citizens. The correct management of these facilities becomes challenging when the source water presents high variability in terms of quantity and quality.

Pre-oxidation with potassium permanganate is the first chemical barrier at many utilities. This chemical is dosed at the beginning of the treatment to oxidise a wide range of compounds for their subsequent removal by treatment processes. It has also some advantages in comparison with other alternatives like chlorine, because it does not generate trihalomethanes (THM), a hazardous disinfection by-product (DBP) of chlorination. This fact becomes very important in utilities that present high THM formation potential since it has allowed the chlorine dosing to be moved to the end of the treatment process, where THM formation is highly reduced. Permanganate is applied for oxidising a wide range of compounds, including: iron and manganese, algal-derived compounds, taste and odour compounds, DBP precursors and for control of microorganisms in the intake structures or treatment basins (World Health Organization 2004; Hu *et al.* 2018).

DWTP treatment managers adjust the permanganate dosing rate according to a multi-parametric evaluation that includes kinetic and inlet quality parameters. At this point, an optimal dosage is the one that maximises the oxidation of a wide group of compounds in raw water (and therefore improving the subsequent treatment unit operations) but does not surpass a certain manganese residual concentration in water. An overdose of permanganate is easily detectable through visual inspection, since it gives water a pink colour. The development of an advanced control tool can help treatment plant managers and operators to deal with this multi-parametric challenge, especially in

doi: 10.2166/wst.2020.142

Mediterranean-like areas where surface water can have strong variations in quantity and quality through the year.

Environmental decision support systems (EDSSs) were designed to cope with this kind of challenge because of their ability to integrate different kinds of mathematical models and expert knowledge (Poch *et al.* 2004). EDSSs are generally structured by a data acquisition level, where all data is gathered and processed (e.g. from a DWTP database); a control level, where mathematical models are fed with data and provide an output for the response variable; and supervisor level, which contains expert rules or mathematical models that evaluate the answer given by the control level. In a previous study, Godo-Pla *et al.* (2019) developed a multi-layer perceptron (MLP), a simple type of artificial neural network (ANN) to predict the permanganate demand at the inlet of DWTP using raw water characteristics and operational parameters as input data.

Even sensitivity analysis and structural validation can contribute in understanding the inner mechanics of ANNs, one limitation of data-driven models is their lack of transparency (Olden & Jackson 2002; Humphrey *et al.* 2017). An EDSS should provide users with a justification for the proposed actions in order to build confidence among users and be a real aid for decision-making (Poch *et al.* 2004; Worm *et al.* 2010). It is also important the development of user-friendly and web-based systems for improving EDSS usability (Mannina *et al.* 2019).

The incorporation of artificial intelligence (AI) techniques into an EDSS has led to more accurate and reliable systems (Núñez *et al.* 2003). The present study investigates whether the lack of transparency of data-driven models can be overcome by reporting the propagation of model uncertainty, and also by comparing these uncertainties with a preliminary approach to a case-based reasoning (CBR) model. CBR has been used for modelling the experimental knowledge of wastewater treatment plants operation for more than two decades (Sànchez-Marrè *et al.* 1997), allowing the use of past experiences to solve new cases in a certain process. In the present study, a CBR model is approached for backing up the predictive model outputs with a distribution of solutions in the past given similar operating conditions. This way, the precision of the predictive model can be compared with the precision in past decisions for similar input conditions and, thus, the confidence in the use of the EDSS for operating a certain process can be strengthened.

The objective of this study was to 1) confirm the appropriateness of the predictive model by a systematic feature and model selection procedures and uncertainty analysis, and 2) integrate the predictive model with a CBR engine in an EDSS to strengthen confidence in the use of the tool for the daily operation of the process at the DWTP.

This paper is structured as follows. In the second section, a brief description of the case-study DWTP and the need for an advanced control system is presented. Then, the methodology for feature selection and model development for multiple-linear regression (MLR), MLP and an approach to a CBR model is described. In the first part of the third section, the input selection and the accuracy of the models is validated with an historical dataset. Then, in the second part, the integration of the different models in an EDSS and its potential is discussed. Finally, in the last section, the conclusions from this work are presented.

MATERIAL AND METHODS

Case study

Llobregat DWTP is located in Abrera (NE Spain) and provides water to the metropolitan area of Barcelona. It takes surface water from Llobregat river and has a maximum treatment capacity of 3.2 $m^3 \cdot s^{-1}$. The treatment train has several processes (pre-oxidation with permanganate, enhanced coagulation, oxidation with chlorine dioxide, sand and carbon filters, electrodialysis reversal and disinfection with sodium hypochlorite) in order to remove THM precursors and comply with Spanish regulation for drinking water (Valero & Arbós 2010). Llobregat river is a Mediterranean catchment that presents high variability in terms of quantity and quality throughout the year, which poses a challenge for treatment plant managers to produce constant effluent water quality. To help with that, Llobregat DWTP has an extensive analytical and on-line monitoring of the treatment process but no predictive tools are available. This fact motivated the development of data-driven models to predict the main operational set-points of the plant, like in potassium permanganate dosing (D_{KMnO4}) at the pre-oxidation step (Godo-Pla *et al.* 2019). This way, the utility's digital infrastructure is used to provide users with augmented decision-making capabilities on the operation of DWTPs.

Llobregat river flows from the Pyrenees to the Mediterranean Sea, with the presence of a system of reservoirs in the upper part of the basin that manage the environmental, domestic and industrial uses of water in the lower part of the basin. Changes in the river management have great effect on

the quality of Llobregat DWTP raw water. Therefore, a five-year period was suitably chosen as a time-space for representing the variations that Llobregat DWTP catchment may suffer due to the management plans of the upper part of the basin. A dataset with analytical values and operational data from daily samples collected at 7 a.m. was built for the period January 2013–December 2018.

The first step for developing a model is the selection of inputs and outputs. The selected output in this study was the potassium permanganate dose (D_{KMnO4}). For selecting the inputs, a pool of input candidates was considered based on the data availability (from commercially available sensors and probes) and on the existence of a known or suspected relationship with the output variable (Baxter *et al.* 2002). The pool of candidates includes all applicable variables for developing the data-driven model: Raw water temperature (T_{RW}), pH (pH_{RW}), total organic carbon (TOC_{RW}), Turbidity ($Turb_{RW}$), electrical conductivity (EC_{RW}), UV absorbance at 254 nm ($UV254_{RW}$), Color ($Color_{RW}$) and Inflow rate (Q_{RW}). Main characteristics of these parameters are summarised in Table 1.

Features selection

Among all the possible input subsets, a procedure has to be followed to systematically choose the one that provides sufficient prediction accuracy for subsequent model development. The *best subset selection* method was applied to a pool of predictors candidates ($p = 8$), including all the quality parameters listed in Table 1. This method consists of fitting models that consider every possible combination of input subsets, using p predictors, for $p = 1, \ldots, 8$ (James *et al.* 2013), being 2^p the total number of possible combinations. By doing this, a single best model can be chosen that minimises the cross-validation prediction error or maximises the adjusted R-squared. Adjusted R-squared was used because it adjusts the coefficient of regression to the number of terms in a model. This method is adequate when p is not large, since it is not computationally efficient. The software used in this study was MATLAB R2015b (MathWorks®).

Table 1 | Raw water characteristics of Llobregat DWTP

Parameter	Unit	Mean	St. dev	10th percentile	90th percentile
T_{RW}	°C	16.8	6	8.5	24.9
pH_{RW}	–	8.11	0.2	7.85	8.37
TOC_{RW}	$mg{\cdot}L^{-1}$	3.31	0.8	2.51	4.21
$Turb_{RW}$	NTU	39	35	5	76
EC_{RW}	$\mu S{\cdot}cm^{-1}$	1,347	262	1,033	1,659
$UV254_{RW}$	m^{-1}	6.94	1.9	5.30	9.10
$Color_{RW}$	mg Pt-Co $\cdot L^{-1}$	11.00	5.0	7.50	16.90
Q_{RW}	$m^3 \cdot s^{-1}$	1.85	0.7	0.90	2.80
D_{KMnO4}	$mg \cdot L^{-1}$	0.82	0.3	0.42	1.23

$N = 2{,}040$ samples from January 2013 to December 2018.

Predictive model selection

For modelling purposes, historical data was allocated into a calibration and test dataset (70 and 30%, respectively) to assess the model's ability to perform well on data that was not used to calibrate it (generalisation property). To ensure that data contained in these subsets contain similar statistical properties, allocation of the data was done using a self-organised map algorithm (May *et al.* 2010) with the *selforg()* function. Therefore, the calibration dataset was used for model fitting purposes and for assessing the replicative validation of the models whereas the test dataset (unseen during model calibration) was used for predictive validation. Two kinds of modelling techniques were compared: multiple linear regression (MLR) and multi-layer perceptron (MLP).

A MLR model can be represented in the form of Equation (1):

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \ldots + \beta_n.X_n \quad (1)$$

where Y is the response variable, β_0 is the intercept coefficient, $X_1, \ldots, X_n$ are input variables and $\beta_1, \ldots, \beta_n$ are coefficients estimated by a least squares technique. Input data was log-scaled before fitting the multiple linear regression model using the *fitlm*() function. Data points below 5% and above 95% of the empirical distribution cumulative function *ecdf()* were marked as outliers and removed from the original dataset after applying a robust regression method.

MLP are simple forms of artificial neural networks and consist of a minimum of three layers: the input layer, where each node corresponds to an input variable; hidden layers, where nonlinear activation functions connect input nodes with the following layer, and the output layer, where the final output is a linear combination of the hidden layer outputs. Regarding MLP model development, different number of nodes (K) ranging from 1 to 9 in the hidden layer were tested to find the model that best captures the underlying relationships in the experimental data. More details on

methodology used for MLP model development can be found in Godo-Pla *et al.* (2019).

Uncertainty analysis

A Monte Carlo scheme was used for quantifying the uncertainty of the models resulting from uncertainties in the parameter estimation step. This quantification in full-scale plants is important to increase the awareness of modelling robustness and to avoid bad modelling practices (Borzooei *et al.* 2019). To these means, 100 parameter sets were sampled from the joint distribution of parameter estimators ($\hat{\theta}$) using multivariate random sampling with *mvnrnd()* function. The probability density of Monte Carlo outputs for each observation can be computed using *ksdensity()* function.

Case-based reasoning model

CBR is an AI modelling technique that aims to provide solutions to new cases by looking at solutions of previous similar cases. In the present application, a CBR model was approached to provide the user with information about which permanganate doses were used in the past, given similar raw water and operational characteristics. A general CBR model is described by four processes: Retrieve, Reuse, Revise and Retain (Aamodt & Plaza 1994). In the present study, the preliminary approach to a CBR model only comprises the first two processes, which consist of (1) retrieving the most similar cases and (2) reusing the solutions (permanganate dosing rate) in these cases to support the predictive model outputs. Local and global similarity indices were used to find the most similar cases to the current one. Local similarity was assessed using domain expert knowledge in a binary basis, as expressed in Equation (2):

$$SIM_{local}(C_{i,k}, C_{j,k}) = \begin{cases} 1 & if \quad C_{j,k} \in [C_{i,k} - atr_k, C_{i,k} + atr_k] \\ 0 & otherwise \end{cases} \tag{2}$$

where SIM_{local} is the local similarity measure of attribute k between cases C_i and C_j, $C_{i,k}$ is the value of attribute k in case C_i and atr_k is the local similarity for attribute k. Domain knowledge was used to assign atr_k values. For $k =$ T_{RW}, $Turb_{RW}$, $UV254_{RW}$ and Q_{RW}, atr_k was set to 2.5 °C, 20 NTU, 1.5 m^{-1} and 0.5 $m^3 \cdot s^{-1}$, respectively.

Global similarity (SIM_{global}) between two cases (C_i, C_j) was also assigned on a true/false basis, being true only if all local similarities were true.

$$SIM_{global}\ (C_i, C_j) = \begin{cases} 1 & if \quad \forall\ k, \quad SIM_{local}(C_{i,k}, C_{j,k}) = 1 \\ 0 & otherwise \end{cases} \tag{3}$$

Given a case C_0, the probability density of solutions for all cases $C_1 \ldots C_N$ from the historical database where $SIM_{global} = 1$ can be computed as a means to illustrate operator's behaviour uncertainties in similar past situations. Note that in the present study it is not intended to find the most similar case and provide a unique solution rather than providing the user with a distribution of similar actions done in the past. Therefore, this preliminary approach to the CBR model gives a probability distribution of past actions that is comparable to the uncertainty analysis made for the predictive model.

A schematic of how the predictive and the CBR model outputs are integrated is shown in Figure 1.

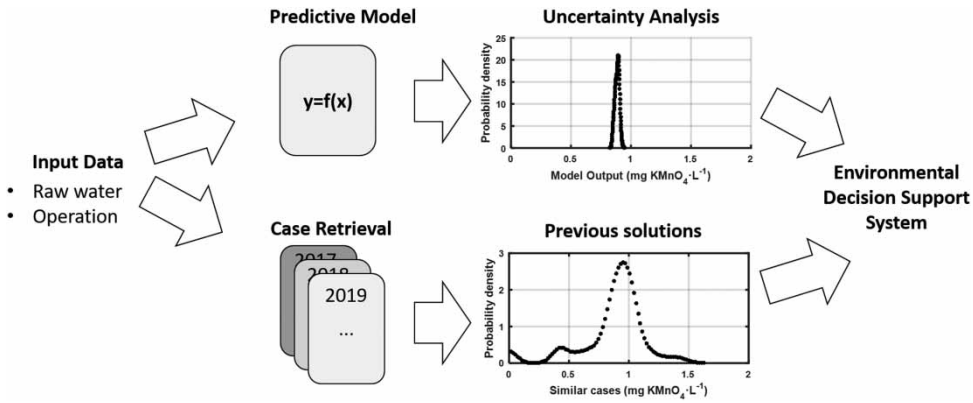

Figure 1 | Flow diagram of predictive and case-based reasoning model integration.

RESULTS AND DISCUSSION

Features selection

Best subset selection method was applied for feature selection of the data-driven models. Results after fitting MLR models with all possible input subsets containing from $p = 1 \ldots 8$ predictors are shown in Figure 2.

It can be seen that as the number of predictors increase, the model accuracy in terms of adjusted R-squared increases but at $p = 4$, the inclusion of an additional predictor in the MLR model does not correspond to a significant increase in the adjusted-R^2. Within all possible combinations including four variables, the subset that maximises model accuracy included the following state variables: T_{RW}, $Turb_{RW}$, $UV254_{RW}$ and Q_{RW}, with an adjusted R-squared of 0.54.

The selected subset was considered to have physical meaning in the pre-oxidation process. T_{RW} strongly affects the kinetics and solubility of permanganate in water, and seasonal variability is strongly related to this. Turbidity and UV254 are surrogate measures for suspended solids, organic matter and sediments, among others. These parameters are usually associated with organic loads resulting from river's runoff, being positively correlated with the permanganate dose. The inflow rate is inversely proportional to the contact time that water is in contact with permanganate in the pre-oxidation chamber and also in the clarifiers, thus affecting the oxidation process. Other parameters like pH_{RW} did not result in the best input subset. This might be because pH is adjusted at the inlet of the DWTP with a target range of 7.4-7.7 by a carbon dioxide dosing. Therefore, pH_{RW} does not play a key role, as expected.

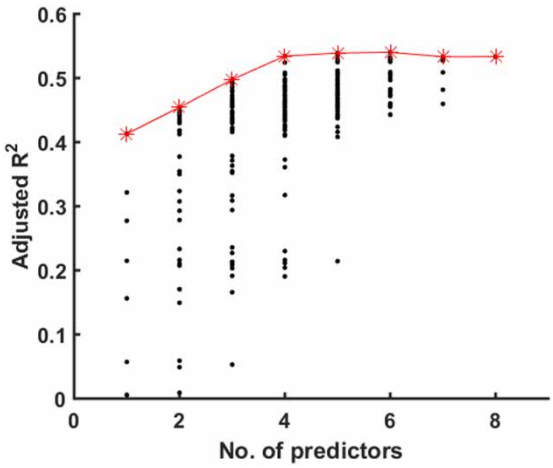

Figure 2 | Adjusted R-squared for all possible combinations of subsets containing from 1 to 8 predictors as input variables in an MLR model for predicting the permanganate dose.

Predictive model validation

The performance statistics of MLR and MLP models using the selected features were compared using both calibration (replicative validation) and test dataset (predictive validation). The results are shown in Figure 3.

MLR showed similar performance compared to MLP models. In terms of model predictive accuracy, it can be seen that MLP-1, MLP-4 and MLP-7 gave similar results with RSE of 0.132, 0.134 and 0.133 $mg \cdot L^{-1}$ respectively, while MLR had an RSE of 0.139 $mg \cdot L^{-1}$. Balancing the number of parameters involved (greater model parsimony) and performance, the MLP-1 model was selected for further analysis and integration into the EDSS. The selected model showed R^2 values of 0.76 and 0.74 for calibration and test dataset, respectively. Monte Carlo outputs of this model were computed for the test dataset (data unseen during model calibration), and it is shown in Figure 4.

It can be observed that Monte Carlo simulations resulted in narrow uncertainty bands on each sample. The developed model adjusted correctly the seasonal as well as smaller day-to-day variations on the permanganate demand. Seasonal variations were associated with changes in raw water temperature. Llobregat DWTP takes water from a river and the temperature has strong differences between the summer and winter period, ranging from 4 to 25 °C, leading to changes in the permanganate demand from 0.4 to 1.2 $mg \cdot L^{-1}$. Also, smaller fluctuations on the permanganate demand in the ± 0.2 $mg \cdot L^{-1}$ range were found because of day-to-day fluctuations of the natural organic matter or other components that modify the permanganate demand.

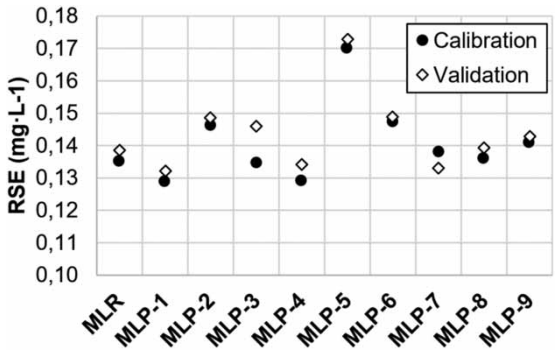

Figure 3 | Root squared error of model predictions for the MLR and the different MLP-K models, being $K = 1 \ldots 9$ nodes in the hidden layer.

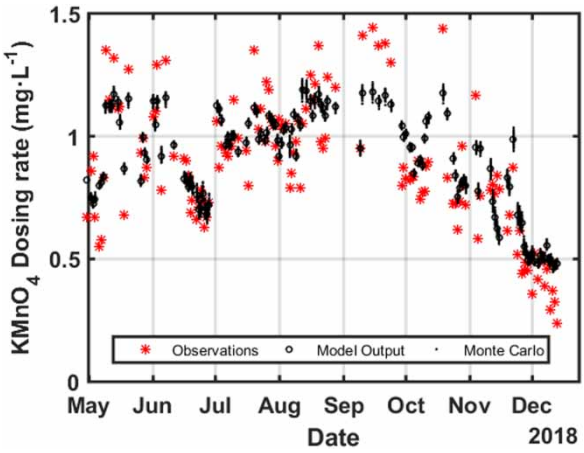

Figure 4 | Representation of uncertainty in MLP-1 predictions for $KMnO_4$ demand time-series, showing the experimental values vs the model output on the test dataset for the period May 2018 to December 2018.

Model integration in an EDSS

For implementation of the model in the real process and to allow the communication of the model outputs to the DWTP operators, models were integrated into an EDSS framework and a graphical user interface (GUI) was built. A screenshot of the developed GUI can be seen at Figure 5. The presented system is connected to the DWTP online data acquisition system and gathers real-time input data. The output of MLP-1 model and the probability distribution of Monte Carlo outputs and response of the CBR model are displayed.

Before running the control system in a closed loop and having the results of the real impact and limitations of this tool, building confidence among the users is needed. Therefore, as an initial step, the EDSS is running in parallel with the Supervisory Control and Data Acquisition (SCADA) system and is working as open-loop control system by recommending the operational set-points. The purposed permanganate dosing rate of the predictive model is backed up by the reporting of the uncertainty analysis and CBR model outputs. This way, the extent of uncertainty/precision of the predictive model but also of the historical behaviour of the operators in similar cases are shown. Generally, uncertainties regarding the predictive model output were in a lesser extent than variations of the permanganate dosing according to previous similar conditions given by the CBR model. The daily decision-making can be speeded-up and improved by offering consistent and robust results to the users, who can consult the model outputs at any time and according to real-time raw water characteristics and operation conditions of the plant.

Moreover, the EDSS architecture allows the addition of expert rules, which act as supervisory rules at the top of the control algorithm. It was considered necessary to add a rule for lowering the permanganate dose to 0.2 $mg \cdot L^{-1}$ in case of achieving manganese concentrations greater than 10 $\mu g \cdot L$-1

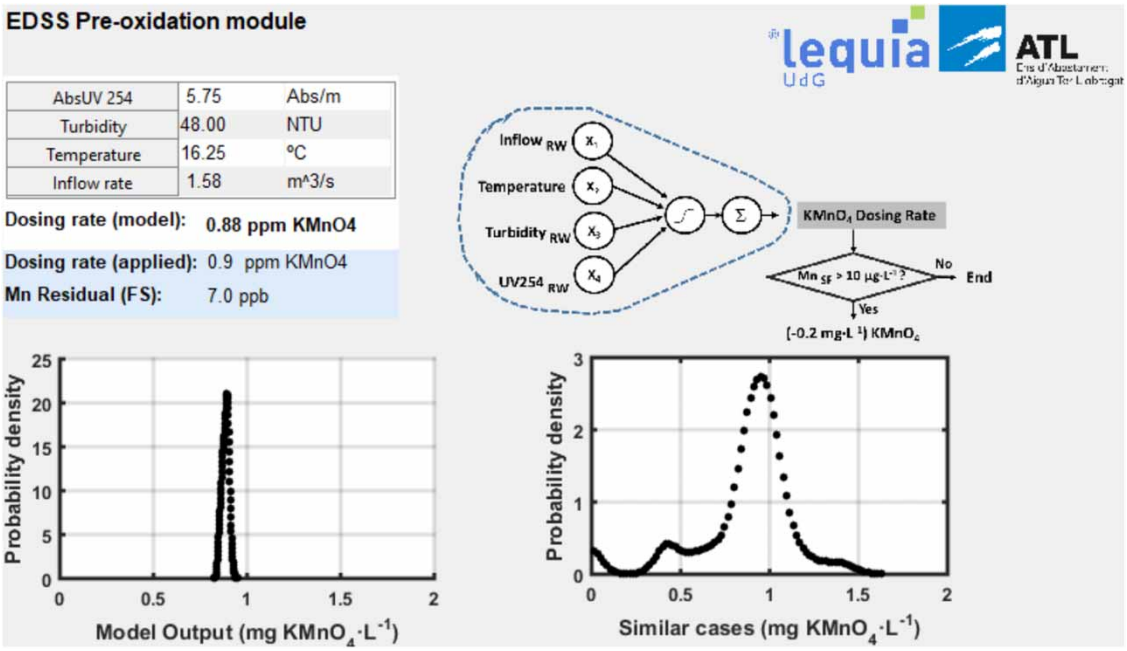

Figure 5 | Graphical user interface for the EDSS.

at the sand filters, to prevent potential overdosing of the chemical. The EDSS was implemented at Llobregat DWTP and was tested in the January-September 2019 period. Figure 6 shows the time-series of the daily average of permanganate dosing rate proposed by the predictive model versus the one applied at full scale.

During the studied period, it was up to DWTP users to apply the predicted dose or adjust it according to the actual concentration of residual manganese in water and previous experience. As a first approximation, the combination of the predictive model with the expert rule for lowering the dosing rate at high residual manganese levels was considered to be sufficiently good. It was shown that the purposed system did not lead to any overdosing of the system, especially in the March-April 2019 period, in which high concentrations of residual manganese (between 10 and 20 $\mu g \cdot L^{-1}$) led to a proposal for lower dosing rates. The laboratory measurements in these cases were all within quality specifications. Excess permanganate passing through the filters has to be avoided, since it may enter the distribution system and lead to an undesirable taste in water (Crittenden *et al.* 2012).

After the implementation phase, it was considered that benefits from EDSS include providing baseline operational set-points while maintaining the operator's added-value expertise in the process. Modifications on the baseline set-points made by the users were recorded for the follow-up of the implementation phase. It is also expected that DWTP users will gradually build confidence in predictive model outputs and implement them more consistently. It was noted that systems like the developed EDSS may contribute to train and support those technical personnel that have not accumulated sufficient experience to run the process, while it serves as a supporting tool for experienced users.

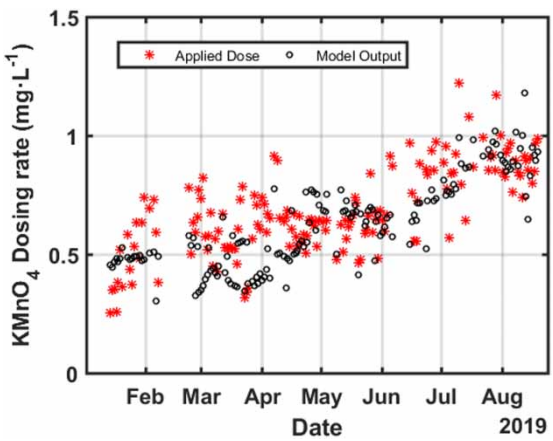

Figure 6 | Results of the implementation phase of the EDSS.

CONCLUSIONS

An environmental decision support system to help with the multi-parametric challenge of controlling the pre-oxidation step at a full-scale DWTP was implemented. To do this, first a systematic procedure for feature selection was done, showing that potassium permanganate dose can be best predicted using temperature, turbidity, absorbance at 254 nm and inflow rate as input variables. For model development purposes, multiple linear regression and multi-layer perceptron models were compared, and the best data-driven approach was shown to be a multi-layer perceptron with one node in the hidden layer, as shown in previous studies.

Uncertainties in the model output resulting from model development were quantified using a Monte Carlo scheme and validated against historical data. The root squared error and R^2 of the predictive model was 0.13 $mg \cdot L^{-1}$ and 0.76 respectively, which was considered sufficiently accurate for the utility needs. In lights of integrating the predictive model in an EDSS for aiding in day-by-day operation of a full-scale DWTP, a case-based reasoning model was developed in order to support model outputs and overcome the black-box nature of the predictive model.

The MLP and CBR models were integrated in an EDSS that gathers data from online sensors and analysers and provide real-time support for daily operation. By integrating these two kinds of model, the user is informed about uncertainty in model predictions, as well as uncertainties related to previous actions with similar operating conditions recorded in the historical database. We believe that this system can increase the robustness of model predictions and allows the user to become more confident in using the EDSS for aiding in decision-making rather than being guided only by previous experience. Also, the EDSS architecture demonstrates being adaptable to specific cases and situations out of the scope of the predictive model by the inclusion of expert rules at the supervisor level. The EDSS is currently implemented at Llobregat DWTP and has been operated as an open-loop control system for 9 months, providing the base-line permanganate dosing rate from which operators decided whether to apply it or adjust it according to their experience to fit more specific cases.

ACKNOWLEDGEMENTS

Lluís Godo-Pla, Hèctor Monclús and Manel Poch want to thank the company Ens d'Abastament d'Aigües Ter-Llobregat (ATL) for their collaboration in this work, especially to Llobregat DWTP treatment managers, Oriol Capdevila and Àngel Barceló. This work was partially supported by University of Girona and ATL with a PhD student grant (IFUdG2017-30) and by Retos de la Sociedad Project (CTM2017-83598-R). LEQUIA has been recognised as a consolidated research group by the Catalan Government (2017- SGR-1552).

REFERENCES

Aamodt, A. & Plaza, E. 1994 Case-based reasoning: foundational issues, methodological variations and system approaches. *AI Commun.* **7**, 39–59.

Baxter, C. W., Stanley, S. J., Zhang, Q. & Smith, D. W. 2002 Developing artificial neural network models of water treatment processes: a guide for utilities. *J. Environ. Eng. Sci.* **1**, 201–211. https://doi.org/10.1139/s02-014.

Borzooei, S., Amerlinck, Y., Abolfathi, S., Panepinto, D., Nopens, I., Lorenzi, E., Meucci, L. & Chiara, M. 2019 Data scarcity in modelling and simulation of a large-scale WWTP: stop sign or a challenge. *J. Water Process Eng.* **28**, 10–20. https://doi.org/10.1016/j.jwpe.2018.12.010.

Crittenden, J. C., Trussell, R. R., Hand, D. W., Howe, K. J., Tchobanoglous, G. & Borchardt, J. H. 2012 *MWH's Water Treatment Principles and Design*. John Wiley & Son, New York, NY.

Godo-Pla, L., Emiliano, P., Valero, F., Poch, M. & Sin, G. 2019 Predicting the oxidant demand in full-scale drinking water treatment using an artificial neural network: uncertainty and sensitivity analysis. *Process Saf. Environ. Prot.* **125**, 317–327. https://doi.org/10.1016/j.psep.2019.03.017.

Hu, J., Chu, W., Sui, M., Xu, B., Gao, N. & Ding, S. 2018 Comparison of drinking water treatment processes combinations for the minimization of subsequent disinfection by-products formation during chlorination and chloramination. *Chem. Eng. J.* **335**, 352–361. https://doi.org/10.1016/j.cej.2017.10.144.

Humphrey, G., Maier, H. R., Wu, W., Mount, N. J., Dandy, G. C., Abrahart, R. J. & Dawson, C. W. 2017 Improved validation framework and R-package for artificial neural network models. *Environ. Model. Softw.* **92**, 82–106. https://doi.org/10.1016/j.envsoft.2017.01.023.

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013 *An Introduction to Statistical Learning with Applications in R*. Springer, New York.

Mannina, G., Rebouças, T. F., Cosenza, A., Sànchez-marrè, M. & Gibert, K. 2019 Decision support systems (DSS) for wastewater treatment plants – A review of the state of the art. *Bioresour. Technol.* 121814. https://doi.org/10.1016/j.biortech.2019.121814.

May, R. J., Maier, H. R. & Dandy, G. C. 2010 Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Networks* **23**, 283–294. https://doi.org/10.1016/j.neunet.2009.11.009.

Núñez, H., Sànchez-Marrè, M., Cortés, U., Comas, J., Martínez, M., Rodríguez-Roda, I. & Poch, M. 2003 A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. *Environ. Model. Softw.* **19**, 809–819. https://doi.org/10.1016/j.envsoft.2003.03.003.

Olden, J. D. & Jackson, D. A. 2002 Illuminating the 'black box': a randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Modell.* **154**, 135–150. https://doi.org/10.1016/S0304-3800(02)00064-9.

Poch, M., Comas, J., Rodríguez-Roda, I., Sànchez-Marrè, M. & Cortés, U. 2004 Designing and building real environmental decision support systems. *Environ. Model. Softw.* **19**, 857–873. https://doi.org/10.1016/j.envsoft.2003.03.007.

Sànchez-Marrè, M., Cortés, U., R-Roda, I., Poch, M. & Lafuente, J. 1997 Learning and adaptation in wastewater treatment plants through case-based reasoning. *Comput. Civ. Infrastruct. Eng.* **12**, 251–266. https://doi.org/10.1111/0885-9507.00061.

Valero, F. & Arbós, R. 2010 Desalination of brackish river water using Electrodialysis Reversal (EDR). Control of the THMs formation in the Barcelona (NE Spain) area. *Desalination* **253**, 170–174. https://doi.org/10.1016/j.desal.2009.11.011.

World Health Organization 2004 *Water Treatment and Pathogen Control: Process Efficiency in Achieving Safe Drinking Water*. IWA Publishing, London, UK.

Worm, G. I. M., van der Helm, A. W. C., Lapikas, T., van Schagen, K. M. & Rietveld, L. C. 2010 Integration of models, data management, interfaces and training support in a drinking water treatment plant simulator. *Environ. Model. Softw.* **25**, 677–683. https://doi.org/10.1016/j.envsoft.2009.05.011.

First received 30 October 2019; accepted in revised form 20 March 2020. Available online 31 March 2020