Predicting the Oxidant Demand in Full-Scale Water Treatment using an Artificial Neural Network: Uncertainty and Sensitivity Analysis

Godo-Pla, Lluís; Emiliano, Pere; Valero, Fernando; Poch, Manel; Sin, Gürkan; Monclús, Hèctor

Published in:
Process Safety and Environmental Protection

Link to article, DOI:
10.1016/j.psep.2019.03.017

Publication date:
2019

Document Version
Peer reviewed version

Citation (APA):
Accepted Manuscript

Title: Predicting the Oxidant Demand in Full-Scale Water Treatment using an Artificial Neural Network: Uncertainty and Sensitivity Analysis

Authors: Lluís Godo-Pla, Pere Emiliano, Fernando Valero, Manel Poch, Gürkan Sin, Héctor Monclús

PII: S0957-5820(19)30261-7
DOI: https://doi.org/10.1016/j.psep.2019.03.017
Reference: PSEP 1700

To appear in: Process Safety and Environment Protection

Received date: 14 February 2019
Revised date: 12 March 2019
Accepted date: 13 March 2019

Please cite this article as: Godo-Pla L, Emiliano P, Valero F, Poch M, Sin G, Monclús H, Predicting the Oxidant Demand in Full-Scale Water Treatment using an Artificial Neural Network: Uncertainty and Sensitivity Analysis, Process Safety and Environmental Protection (2019), https://doi.org/10.1016/j.psep.2019.03.017

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Predicting the Oxidant Demand in Full-Scale Water Treatment using an Artificial Neural Network: Uncertainty and Sensitivity Analysis

Lluis Godo-Pla1,2, Pere Emiliano2, Fernando Valero2, Manel Poch1, Gürkan Sin3, Hèctor Monclús1

1LEQUIA, Institute of the Environment, University of Girona, Maria Aurèlia Capmany 69, E-17003, Girona, Catalonia, Spain
2Ens d’Abastament d’Aigua Ter Llobregat (ATL), Sant Martí de l’Erm, 30. 08970 Sant Joan Despí, Barcelona, Spain
3PROSYS research center, Department of Chemical and Biochemical Engineering, Technical University of Denmark (DTU), Building 229, DK-2800 Lyngby, Denmark

Graphical abstract

Highlights:
Potassium permanganate demand in raw water of a DWTP was modelled using ANN. Comprehensive methodology on parameter estimation and uncertainty analysis on ANN. The model input-output relationships were assessed by sensitivity analysis techniques. The model output gives support to the daily management of the treatment plant.

Abstract: Drinking Water Treatment Plants face changes in raw water quality and quantity and the treatment needs to be adjusted accordingly to produce the best water quality at the minimum environmental cost. The amount of data generated along drinking water treatment plants allows developing data-based models like artificial neural networks that are able to predict operational parameters and can be incorporated into environmental decision support systems. In the present study, an artificial neural network is developed for predicting the potassium permanganate demand at the inlet of a full-scale Drinking Water Treatment Plant. A systematic methodology is carried out for outlier detection and removal from the original dataset. Afterwards, model parameters estimation, uncertainty and sensitivity analysis is reported to assess prediction quality and uncertainty of the models. Bootstrap method was used for parameter estimation, and uncertainty of the inputs onto the model outputs was propagated using a Monte Carlo scheme. Several sensitivity analysis methods were evaluated to understand the contribution of the inputs on the output of the models, and this was in accordance with the knowledge of the process and other studies found in the literature. The selected architecture consisted of a feed-forward multi-layer perceptron with four inputs and one node in the hidden layer with a sigmoid activation function. The mean absolute error of the resulting model is 0.128 mg·L⁻¹, which was considered acceptable by the DWTP operators. The resulting model provided good results in terms of replicative, predictive and structural performance and is to be used for supporting decision-making in the daily operation of the plant.

Keywords: artificial neural network, uncertainty, sensitivity, water treatment, oxidation, permanganate

Abbreviations: ANN, artificial neural network; CI, confidence interval; CW, connection weight; DBP, disinfection by-product; DOC, dissolved organic carbon; DWTP, drinking water treatment plant; EC, electrical conductivity; ECDF, empirical cumulative distribution function; EDSS, environmental decision support system; MAE, mean absolute error; MLP, multi-layer perceptron; MC, Monte Carlo; MSE, mean squared error; NOM, natural organic matter; PaD, partial derivatives method; Q, inflow rate; RI, relative
importance; RMSE, root mean squared error; RW, raw water; SOM, self-organizing map; SUVA, specific ultraviolet absorbance; T, temperature; THM, trihalomethanes; TOC, total organic carbon; TSS, total suspended solids; Turb, turbidity; UV254, ultraviolet absorbance at 254 nm.

1 INTRODUCTION

Potassium permanganate (KMnO₄) has replaced chlorine at the inlet of Drinking Water Treatment Plants (DWTP) as an oxidant with different objectives: Oxidation of iron and manganese, oxidation of algal-derived taste and odour compounds, control of microorganisms and Disinfection By-Products (DBP) precursors (Hu et al., 2018; Kim et al., 2018). The main advantage of permanganate instead of chlorine at this point, is that it does not produce DBPs like trihalomethanes (THM) while oxidizing a wide range of organic and inorganic compounds. Contaminant reduction mechanisms include oxidation through addition to double bonds, abstraction of hydrogen or hydride and electron transfer (Waldemer and Tratnyek, 2006) and adsorption onto manganese dioxide formed during the reduction of permanganate (Colthurst and Singer, 1982). The dosing rate of KMnO₄ in the pre-oxidation process is site specific and typically varies between 1 and 3 mg·L⁻¹, with at least 1 hour of contact time (Crittenden et al., 2012). Operators of DWTP adjust the dosing rate daily according to the kinetic conditions (temperature, contact time), and the inlet quality (organic matter, presence of iron and manganese…). Maximising multiple-target oxidation process will have a positive effect in the subsequent treatment steps: Improves coagulation/flocculation, reduces the microbiological load (improving carbon filters performance) and reduces the potential of THM formation by partially oxidising the organic matter (Moyers and Wu, 1985). Hence, the oxidation capacity has to be maximised but at the same time, care must be taken to maintain a low residual of manganese (regulated to 100 ppb) after the coagulation/flocculation process. An overdose of this chemical results in an undesirable taste and can lead to the accumulation of deposits in the distribution system (World Health Organization, 2017). Therefore, the dosing control becomes important at an hourly or daily basis depending on the raw-water characteristics.

Specially in Mediterranean regions, where water quality at the inlet of DWTP can have important changes in short periods of time, adjusting and optimising the operational parameters becomes
challenging. In addition, the shift work of DWTP operators may also lead to problems in homogenising criteria to unify strategies and optimise operational parameters. To overcome this and in order to anticipate changes on the permanganate demand due to changes on water quality or operational conditions, several modelling approaches can be applied using knowledge and data from full-scale facilities and be integrated into Environmental Decision Support Systems (EDSS), aiding in the daily operational decision-making (Corominas et al., 2017; Hamouda et al., 2009; Poch et al., 2004). Integrating modern technologies like EDSS into industrial processes is in alignment with the “Industry 4.0” concept and can bring managerial benefits such as virtualization, real-time capability and service orientation among others (Kamble et al., 2018; Stock et al., 2018).

Some permanganate modelling studies have been done targeting the kinetics of microcystins oxidation (Jeong et al., 2017; Kim et al., 2018; Sharma et al., 2012) and identified pH, Dissolved Organic Matter (DOC), UV absorbance of the water at 254 nm (UV254) and specific ultraviolet absorbance (SUVA) as quality parameters that control the kinetics of this process. Some others have focused on the oxidation of organic contaminants and THM precursors removal (Colthurst and Singer, 1982; Hidayah and Yeh, 2018; Hu et al., 2018; Naceradska et al., 2017; Sharma et al., 2012; Zhang et al., 2009). Nonetheless, many DWTP practitioners do not focus on one specific target but want to predict the permanganate dose that can be applied without surpassing the residual (permanganate demand), therefore maximising the reduction of the contaminant load of different groups for the subsequent treatment steps. In this point, there is no universally accepted mathematical description of the overall process. In this context, artificial neural networks (ANN) have been reported as universal approximators and as a suitable tool for their predictive potential and the capacity of modelling multiple-variable nonlinear phenomena in water treatment (Baxter et al., 2002; Maier and Dandy, 2000).

ANN have been broadly investigated and applied in forecasting operational parameters in the drinking water treatment field, focusing mainly in the coagulation process (Baxter et al., 1999; Gagnon et al., 1997; Griffiths and Andrews, 2011; Juntunen et al., 2012; Maier et al., 2004) and on evaluating filter performance (Baxter et al., 2002; Hawari and Alnahhal, 2016). Other applications of ANN in similar environmental processes can be found in the literature regarding optimisation of metals and organic compounds sorption processes in the industrial water treatment field (Hamzaoui et al., 2017; Mazaheri
The great diversity of ANN that exists in the literature has led to the need of establishing frameworks and protocols in the last decade (Baxter et al., 2002; Humphrey et al., 2017; Maier et al., 2010; Wu et al., 2014) to homogenize criteria, avoid over-parametrization of ANN models and guarantee the reproducibility and scientific reporting of these models. However, uncertainty analysis of ANN is not usually reported and this could give additional confidence when using model predictions, especially at full-scale facilities. As outlined by Wu et al. (2014), uncertainty of ANN outputs is an area that needs more attention and can be addressed by implementing parameter estimation methods incorporating uncertainty and assessing the structural validity of ANN models.

The objective of this study is to develop an ANN model for predicting the potassium permanganate demand for a drinking water treatment plant that catches water from a Mediterranean river. In this case, the modelling effort does not result in a direct control of a specific water quality objective (e.g. Turbidity < 1 NTU), but on a finding the pattern in the historical data between raw water characteristics and the permanganate demand over a four-year study period. Given the data-driven nature of the proposed model and the aim of achieving a good modelling accuracy/prediction, a robust regression method is used for outlier detection and removal. Next, a comprehensive parameter estimation and uncertainty analysis is performed to assess the reliability of the predictions. Moreover, sensitivity analysis is performed for a structural validation of the evaluated models.

2 MATERIALS AND METHODS

2.1 Case Study

Llobregat DWTP is situated in Abrera, NE Spain, and managed by ATL to provide water to the metropolitan area of Barcelona. It takes the raw water directly from Llobregat river and has a maximum treatment capacity of 3.3 m$^3$·s$^{-1}$. The raw water presents high concentrations of salts due to the salt deposits located at the upper part of the basin and high organic matter and temperature. Because of this, the chlorination at the DWTP was located at the end of the process and several processes (pre-oxidation with permanganate, enhanced coagulation, carbon filters and electrolysis reversal) target the removal of THM precursors (Valero and Arbós, 2010). The dataset used in this study consisted in
1101 observations from analytical values and operational data collected daily at 7 am from January 2015 to September 2018, from ATL database. It is to note that there was a malfunctioning in the UV254 sensor between 1/1/2018 and 24/4/2018. Therefore, data points from this period are missing. Summary of this data is presented in Table 1. Total organic carbon of raw water (TOC) measurements were carried out by a Shimadzu TOC-VCSH analyser (Shimadzu Europa GmbH, Duisburg, Germany). Turbidity was measured by means of Hach 2100N Turbidimeter. Electrical conductivity (EC) was measured with an electrical conductivimeter (Multimeter MM41, Crison). Colour and UV254 were measured by means of an Agilent 8453 UV-Visible Spectrophotometer.

The first step of the conventional treatment consists in pre-oxidation with potassium permanganate. The range for potassium permanganate dosing rate is between 0.4 and 1.8 mg KMnO₄·L⁻¹. The permanganate demand depends mainly on the raw water load (measurable with UV254, TOC and Turbidity online sensors) and on the kinetics of the oxidation. The latter depends on the temperature and the hydraulic retention time (inversely proportional to the inflow rate).

2.2 Artificial Neural Network model

The software used in this study for building the models was MATLAB R2015b (MathWorks®).

For model development purposes, data has to be divided into calibration and validation dataset. Calibration data set is fed to error minimisation algorithms for parameter estimation and tuning. On the other hand, the validation dataset is unseen during model development phase, and is used for model evaluation. A discussion on data division methods for development for data-based algorithms can be seen in Bowden (2002) and Galelli et al. (2014). In this case a Self-Organizing Map (SOM) algorithm is used in this paper for data division into calibration and validation dataset. This method clusters the data in a way that have similar statistical properties. The selforgmap() function with the default parameters was used for implementing this algorithm. When data is clustered, 70% of the samples in each cluster were included in the calibration dataset and the remaining 30% of the samples were included in the validation dataset. This resulted in 68% of the samples in the Calibration dataset and 32% in Validation dataset.
Among different possible architectures, multi-layer perceptrons (MLP) are amongst the most used ANN in the prediction of water quality parameters (Maier et al., 2010; Wu et al., 2014). A structure for a single hidden layer MLP is presented in Figure 1. MLP have minimum three layers: Input layer, where each node represents one input variable; Hidden layer, where each input is weighted and connected with non-linear activated function (sigmoid) and the Output layer, where the output is calculated from the result of the previous layer nodes through a linear transfer function.

The architecture of the presented model is defined by the number of independent or state variables ($N_{var}$) and the number of nodes in the hidden layer ($K$). A scheme of the generic model is shown for $N=4$ and $K$ nodes in Figure 2. It consists of two different kind of inputs (State variables or inputs, and Parameters) connected with an output through a single-hidden-layer MLP. For the MLP architecture with a single hidden layer, there is a total amount of $K-(N_{var}+2)+1$ parameters to be estimated.

Input variables and outputs were standardized with Z-score transformation. In this operation, input variables are scaled and centred to zero according to the mean and standard deviation using Eq. (1).

$$x_{i,norm} = \frac{x_i - \bar{x}_{mean}}{x_{Std}}$$  \hspace{1cm} (1)

where $x_i$ is the $i^{th}$ sample for variable $x$, $\bar{x}_{mean}$ is the mean value of variable $x$, $x_{Std}$ is the standard deviation and $x_{i,norm}$ is the normalised value of $x_i$. Standardization of input variables is essential to ensure that same percentage change in the weighted sum of the inputs causes a similar percentage change in the unit output (Olden and Jackson, 2002).

Input variable selection is a relevant step in the development of data-driven models, and different algorithms can be used for this assessment to ensure that a subset of relevant and non-redundant input variables are chosen (Galelli et al., 2014). Given that the aim of present study is to develop a model for an EDSS to be tested at a full-scale facility, only variables which could be measured through available online sensors or analysers were considered. To increase the confidence with the ANN model among operators and treatment managers, a first set of input candidates were selected based on the domain knowledge and the availability of online measures. Different input candidate subsets were
tested to find the balance between model parsimony and goodness-of-fit metrics. After this trial-and-error process, the selected inputs for building the model were the raw-water temperature ($T_{RW}$), turbidity ($Turb_{RW}$), UV254 absorbance ($UV_{254RW}$), and the inflow rate ($Q_{RW}$). According to the domain knowledge, it is expected that $T_{RW}$, $Turb_{RW}$ and $UV_{254RW}$ have a positive effect on the permanganate demand. Operators have to lower the dose down to 0.4 mg/L in winter because of low reactivity (low solubility and slow kinetics) and increase up to 1.5 mg/L in summer season, where the oxidative potential of permanganate is the maximum. $UV_{254RW}$ measures the aromatic organic matter content in water, and is used as a surrogate for natural organic matter (NOM), which increases the demand of permanganate. $Turb_{RW}$ is also used as surrogate for many parameters including NOM, Total suspended solids (TSS), chemical precipitates such as iron and manganese, and protozoa (World Health Organization, 2017). On the other hand, $Q_{RW}$ is used as an inverse of the hydraulic retention time and is expected to negatively impact the permanganate demand. The pH of water is set to 7.6 at the inlet of the DWTP by a carbon dioxide dosing step, and therefore this parameter is not included as a potential permanganate demand input.

The Neural Network Toolbox™ was used to obtain parameter values for initialisation of the algorithms. ANN architectures corresponding to a single hidden layer MLP with number of nodes ranging from 1 to 9 in the hidden layer (namely ANN-$K$, being $K$ the number of hidden nodes) are assessed in this study for finding the model that best captures the underlying relationships in the experimental data.

### 2.3 Parameter Estimation

In a parameter estimation problem, a set of model parameter is calculated using an error minimisation algorithm following Eqs. (2) and (3):

\[ e_i = y_i - f(x_i, \theta) \]  
\[ \theta = \text{arg min} \sum_{i=1}^{N} e_i^2 \]
where \( x_i \) is the \( i \)th observation of independent (or input) variables, \( y_i \) is the \( i \)th observation of the dependent (or target) variable \( y \), \( \Theta \) is the set of unknown model parameters, \( e_i \) is the error between the observed and predicted value on the \( i \)th observation and \( N \) is the total number of observations. In this method, residuals \( e_i \) (difference between observations \( y_i \) and model outputs \( f(x_i, \theta) \)) are used to perform a minimisation algorithm using least squares method and find optimal parameter values \( \theta \).

Typical error minimisation algorithms used in parameter estimation, like least squared, assume a Gaussian distribution of errors and therefore, outliers can have a significant impact on the curve fit. Given the fact that data from a full-scale DWTP is used in this study, the presence of outliers resulting from abnormal conditions during start-up of the treatment or errors in data entry in the database is quite predictable. Lorentzian or Cauchy distributions can adjust the distribution of errors in a more appropriate manner than a Gaussian fit, because they have wider tails. Therefore, residuals between observations and model output can be accordingly weighted \( (\omega_{\text{Cauchy},i}) \) before applying a least-squares method to obtain parameter estimator set \( \theta^* \), as shown in Eq. (4) and (5).

\[
\omega_{\text{Cauchy},i} = \frac{1}{1+e_i^2} \tag{4}
\]

\[
\theta^* = \arg \min \sum_{i=1}^{N} \omega_{\text{Cauchy},i} \cdot e_i^2 \tag{5}
\]

A robust nonlinear regression method like the iterative reweighted least squares algorithm, assuming a Cauchy distribution of the errors will be used as a first parameter estimation method for advanced outlier treatment. This method was implemented with the nlinfit() function, specifying the “Cauchy robust weight function” option.

Following the methodology proposed in Frutiger et al. (2015), an empirical cumulative distribution function (ECDF) of the residuals between experimental data and predicted values from the robust nonlinear regression is used for outlier detection and removal. The basis of this method is that it does not assume a normal distribution of the errors a priori, instead the method estimates directly the underlying cumulative distribution function of the errors and use this to infer outliers. The ECDF is a step function that increases by \( 1/n \) in every data point \( (n \) is the number of observations) and was
calculated using the \texttt{ecdf()} function. According to this method, data points below 2.5% or above 97.5% the probability levels can be considered outliers and removed from the dataset.

Once outliers have been removed from the original dataset, bootstrap method was used as a statistical method for parameter estimation of the model. This method was chosen because it works with the actual distribution of the measurement errors and is considered to be valid to calculate parameter estimation when the underlying distribution of the errors is not known (Sin and Gernaey, 2016). In this method, a first parameter estimation is done using nonlinear least squares minimisation function \texttt{lsqnonlin()}, thus obtaining a first set of parameter estimators $\hat{\theta}_0$. The residuals from this process ($\hat{\epsilon}$) are collected and new residuals subsets $\hat{\epsilon}^*$ are created by random sampling and replacement of $\hat{\epsilon}$. Several synthetic datasets $\{y_1, y_2\ldots y_n\}$ are then generated by assigning $\hat{\epsilon}^*$ to the output dataset using a Monte Carlo scheme following Eq. (6).

$$y_i = y + \hat{\epsilon}^*_i \quad for \quad i = 1, 2 \ldots N_{BT} \quad (6)$$

where $N_{BT}$ is the number of bootstrap samples. This way, the measurement error is simulated in the new datasets. At this point, it is important to check that the distribution of the residuals of the model output is random. For each synthetic dataset, the parameters are re-estimated, thus obtaining a distribution of the parameter estimators $\hat{\theta} = \hat{\theta}_1, \hat{\theta}_2 \ldots \hat{\theta}_{N_{BT}} (N_{BT} \times p)$ ($N_{BT}$ is the number of synthetic datasets and $p$ the number of parameters estimated) which is supposedly centred to the real value of the parameters. MATLAB Scripts from (Sin and Gernaey, 2016) were adapted for parameter estimation using Bootstrap method.

After applying Bootstrap method for parameter estimation, several statistics can be computed to assess model performance in terms of replicative and predictive performance. While replicative performance refers to the ability of ANN to fit the calibration dataset, the predictive performance refers to the validation dataset. Mean Absolute Error (MAE), mean squared error (MSE), root mean squared error (RMSE) and Spearman’s rank correlation coefficient ($\rho^2$) can be computed for assessing replicative and predictive performance of the models.
The covariance and correlation matrix of parameter estimators can be obtained directly from $\hat{\theta} (N_B T \times p)$ using the functions `cov()` and `corr()`, respectively.

The number of parameters to be estimated and therefore, the size of the correlation matrix ($p \times p$) will depend on the architecture of the ANN. The correlation between two parameters will indicate if the parameter estimators are uniquely identifiable or not.

### 2.4 Uncertainty Analysis

A Monte Carlo (MC) scheme was used to quantify the uncertainty in the model output that result from the uncertainty of the parameter estimation problem. In this method, we randomly generate $N_{MC}$ possible parameter sets. The parameter sets are sampled using multivariate random sampling with the function `mvnrnd()` from the parameter estimator covariance matrix obtained in the previous step (bootstrap method), therefore having $N_{MC}$ sets of possible parameter estimator $\hat{\theta}_m$ sets. Other sampling techniques like latin hypercube sampling (LHS) could also be used. The model is computed for each set of parameters and some statistics like mean values, and the 10th and 90th percentile of the model predictions can be computed. MATLAB Scripts from Sin et al. (2009) and Sin and Gernaey (2016) were adapted for uncertainty analysis using MC method.

### 2.5 Sensitivity Analysis

Sensitivity analysis are used to study the effects of variation in the model inputs onto the model output. ANN sensitivity analysis methods differ from the typical ones used in statistical or mechanistic models in that their parameters (weights and biases values of the hidden and output layer) do not have any physical meaning by itself. Humphrey et al. (2017) highlights the importance of considering and reporting the structural validity of artificial neural network models (in addition to predictive and replicative validity) in order to make plausible models and to certify that the model has captured well the underlying relationship in the calibration data. In this regard, several sensitivity techniques to quantify the relative importance (RI) of ANN inputs are reported and discussed in the literature (Humphrey et al., 2017; Olden and Jackson, 2002; Sarle, 2000). Among them, Connection Weight (CW), Partial Derivatives (PaD) and Profile Method (PM) are chosen in this study as two different methods for quantifying RI of
inputs: The first one, based on the connection weights between input-hidden and hidden-output layers of ANN; and the second and the third one, based on the effects of variation of the inputs onto the output of the ANN model. Details of theory are given in the supplementary material, and for further details and discussions on these methods we direct the interested reader to Humphrey et al. (2017).

A scheme of the methodological approach covered in this study is summarised at Figure 3.

3 RESULTS AND DISCUSSION

Results for parameter estimation, uncertainty analysis and sensitivity analysis are shown in this section.

3.1 Outlier treatment

Data was fitted with robust nonlinear regression assuming Cauchy distribution of errors for every tested ANN architecture. Figure 4 shows on the one hand, the deviation of the distribution of the residuals of this fit comparing to a Gaussian one, and on the other one the empirical cumulative distribution function used to detect the outliers of the dataset. It can be observed how the residual distribution displays wide tails comparing to the Gaussian distribution, which confirms the necessity of using actual/empirical distribution rather than assume it is normal distribution. In this way, employing the ECDF can guide better outlier detection comparing to other methods that assumes normal distribution of errors (Frutiger et al., 2015). Points below 2.5% and above 97.5% of probabilities in ECDF were identified as outliers. Figure 5 shows the time-series plot of the robust model fit versus target data, including outlier identification in the dataset used for ANN-1.

To see the effects of the outlier treatment on the model performance statistics, a nonlinear least squares method was applied for parameter estimation on the original and on the outlier treated dataset. The results showing the effects of the outlier treatment are shown in Table 2.
It is shown that, by applying the outlier removal based on the robust regression and ECDF method, MAE improved from 0.15 to 0.13 mg·L\(^{-1}\) and \(\rho^2\) from 0.79 to 0.83.

### 3.2 Parameter estimation

Bootstrap method was applied on the outlier-treated dataset and selected ANN architectures. The residuals obtained using this method were checked and shown in Figure 6, which confirmed to follow a random pattern scattered around mean close to zero. This confirms the suitability of using bootstrap method to quantify the effect of measurement error on model performance (Efron, 1979; Frutiger et al., 2016). The mean of the parameter estimator matrix (\(\hat{\theta}\)) calculated after bootstrapping for every ANN architecture were used for model performance evaluation.

The performance statistics of ANN models on both calibration and validation dataset after exclusion of ECDF outliers and resulting from applying Bootstrap method for parameter estimation are presented in Table 3. The MAE and \(\rho^2\) for both calibration and validation dataset when increasing the number of hidden nodes in the ANN model from 1 to 9 can be seen in Figure 7. For evaluating model performance, the attention is focused on the validation dataset, since it is unseen data during the model development phase. In Figure 6, three minimum MAE areas can be observed for the validation dataset. The first one, at \(K=1\) with MAE= 0.128 mg·L\(^{-1}\); the second one at \(K=4\) with MAE=0.129 mg·L\(^{-1}\) and the third one at \(K=7,8\) and 9 with MAE= 0.127, 0.123 and 0.124 mg·L\(^{-1}\) respectively. ANN with 1, 4 and 7 nodes in the hidden layer showed better performance in terms of \(\rho^2\) (0.801, 0.804, and 0.810 respectively). As can be observed among ANN-1, ANN-4 and ANN-7, all of the performance statistics are in the same range for both calibration and validation dataset and there is no clear difference among them regarding these metrics. Llobregat DWTP treatment managers accepted model predictions that have discrepancies regarding their operational data in the 0-0.15 mg·L\(^{-1}\) range. Therefore, the values found for MAE, MSE, RMSE and \(\rho^2\) confirm the applicability of the models to predict the permanganate demand in terms of predictive performance. The selected models are candidates to further explore through uncertainty and sensitivity analysis.
The standard deviation of model parameters, the 95% confidence interval (CI) and the correlation matrix can be calculated from the covariance matrix obtained after applying bootstrap method. The correlation matrix of $\hat{\theta}$ can help in determining if parameter estimators are uniquely identifiable or not. The summary statistics and correlation values of $\hat{\theta}$ for the different ANN models are provided in the supplementary material.

Without taking into account the correlation of a parameter with itself, ANN-1 showed 47.62% of the correlation coefficients (in absolute terms) above 0.5. These means that a big part of parameter estimators are strongly correlated (have similar sensitivity to the model output) and thus, not uniquely identifiable. On the other hand, ANN-4 and ANN-7 showed only 3.78 and 1.00% of the correlation values above this number.

Overall these identifiability issues (the fact unique parameter estimation is not possible) means that when ANN model is used for prediction, the simulations with the ANN model should be performed considering not only the mean values but also the covariance matrix of the parameters of the model, i.e. by performing MC simulations. The outcome of MC simulations will provide mean and variance of the ANN predictions. Identifiability of ANN parameters is not usually reported. Parameters with identifiability issues should not be attributed physical meaning since their values are not unique (Frutiger et al., 2016).

### 3.3 Uncertainty Analysis

MC method was used for sampling different parameter subsets from the parameter estimators covariance matrix. In total, 100 different parameter estimator subsets $\Theta_i$ were obtained using multivariate random sampling. The model output for ANN-1, ANN-4 and ANN-7 including 10th and 90th percentile and mean value of MC simulations are shown in Figure 8. The spread of MC output is also represented by showing the probability density of the outputs at one specific data point. The higher extent of uncertainty in parameter estimators, the larger spread of MC outputs for each data point. Time-series plots of model output versus observed data including 10th and 90th percentile of MC simulations for the whole studied period are included in the supplementary material.
In general, the extent of uncertainty in the three tested models was shown to be low. Also, there were no significantly variations on the extent of uncertainty along the studied time period (e.g. variations between summer and winter periods). The smallest uncertainty was found for ANN-1, whereas it slightly increased at ANN-4 and ANN-7. Low uncertainty bounds can be due to negatively and highly correlated model parameters. The increase in the number of parameters and the better identifiability of ANN-4 and ANN-7 model parameters enlarged the uncertainty bounds and thus, providing a slightly better fit with the observed data. Due to the narrow uncertainty bands, only 8.99%, 18.43% and 12.89% of the experimental values for permanganate demand were shown to lie within the 10th and 90th percentile of MC outputs of ANN-1, ANN-4 and ANN-7 respectively.

### 3.4 Sensitivity Analysis

The RI values of the four inputs selected for three models were calculated using Connection Weight, Partial Derivatives and Profile Method, and results are shown in Figure 9. The CW and PaD method illustrated the relative importance of inputs in the same order ($\text{Turb}_{RW} > \text{TRW} > \text{UV254}_{RW} > \text{Q}_{RW}$), whereas $\text{RI}_{Profile}$ showed different order of importance especially in ANN-1. Turbidity is surrogate measure of sediments in water, metals, and NOM (World Health Organization, 2017) and thus is expected be positively correlated with the permanganate demand. UV254 measure accounts for the aromatic content of dissolved organic matter. In a study conducted by (Kim et al., 2018) UV254 and specific UV254 (SUVA254) were identified as the main DOM characteristics contributing to the permanganate consumption in natural waters, showing high correlation ($R^2=0.95$ and 0.96). This is in alignment with the RI of UV254 in the different models, especially in ANN-1. Temperature has positive RI for all the studied models, since it has an effect on the oxidation rate constants of permanganate in natural waters (Crittenden et al., 2012; Kim et al., 2018; Rodríguez et al., 2007). Finally, $\text{Q}_{RW}$, which is inversely proportional to the hydraulic retention time, showed negative values of RI and was placed to the least important input variable. A sufficient contact time prevents permanganate to pass through the downstream filters and enter the distribution system (Crittenden et al., 2012), and this is why operators have to adjust the dose according to the corresponding inflow rate.
In addition to computed \( R_{CW} \), \( R_{PaD} \) and \( R_{Profile} \), the profile method is also reported for assessing the plausibility of the models input-output relationships. This method is suitable for visually identifying the patterns fixed by the different ANN models. Very similar response curves were found for ANN-1 and ANN-4 regarding all input variables but less plausible relationships were found for ANN-7. The profile method plots for ANN-1, ANN-4 and ANN-7 can be find in the supplementary material section.

Even ANN-1 had the simplest structure, it showed the most plausible input-output relationships according to a-priori knowledge of the system.

Similar findings can be seen in study conducted by Humphrey et al. (2017), where the predictive and structural performance of three different ANN models with 1, 12 and 14 nodes in the hidden layer were compared for predicting surface water turbidity. It was shown that even a MLP with one neuron had the simplest architecture (low number of parameters and architecture) compared with one node in the hidden layer (the simplest tested architecture) captured the underlying relationships between inputs and outputs better than the larger architectures, and this was demonstrated through sensitivity analysis techniques.

4 CONCLUSIONS

An artificial neural network for predicting the potassium permanganate demand in the raw water of a Drinking Water Treatment plant has been developed using historical data from a full-scale facility. In this novel approach, the model objective was to find the pattern in the data between raw water characteristics and the permanganate dosing rate over a four-year period. The developed model used the temperature and the inflow rate as kinetic parameters, and UV254 and Turbidity as surrogate measures for the natural organic matter, chemical precipitates and microorganisms.

A comprehensive and systematic procedure for parameter estimation, uncertainty analysis and sensitivity analysis were applied to different ANN architectures and the results were discussed in order to choose the model which showed the better fit to the data and adjustment to the underlying
input/output relationships. To this end, several architectures were systematically evaluated through replicative, predictive and structural validity procedures. The most plausible monotonic relationships between input variables and output of the model, especially regarding variations on the raw-water inflow rate, were found for a feed-forward single-layer MLP with sigmoid activation function and one hidden node.

Moreover, uncertainty of the model output resulting from the propagation of uncertainty in the data, modelling assumptions and model parameters was quantified. This information is important because it reports the accuracy of the model prediction and thus, contributes in that operators and plant managers build confidence in data-driven models. Even a low percentage of the observed data lied within the 10th and 90th percentile model predictions due to narrow uncertainty bands, the good results in terms of replicative, predictive and structural validity confirmed the adequacy of the developed model. Given the number of nonlinearities and operational limitations of permanganate in the pre-oxidation step, the proposed model is an approximate for the general trend of permanganate demand for the presented study case. Targeting specific pollutants would allow optimisation of the process, which needs to be addressed according to each DWTP specific needs. Nevertheless, the resulting model is to be integrated in an environmental decision support system that can be fed by on-line data to improve and accelerate the decision-making process of plant managers and operators.

ACKNOWLEDGEMENTS

Lluis Godo-Pla, Hèctor Monclús and Manel Poch want to thank the company ATLL CGC SA for their collaboration in this work, especially to DWTP Abrera treatment managers, Santiago González, 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 Ministry of Science, Innovation and Universities of Spain with postdoctoral grant (IJCI-2015-23159). LEQUIA has been recognized as consolidated research group by the Catalan Government (2017- SGR-1552).
REFERENCES


LIST OF FIGURES

Figure 1. Single-layer MLP structure with N input variables, K nodes in the hidden layer and one output.

Figure 2. Generic model structure for a multi-layer perceptron with a single hidden layer and sigmoid activation function for the prediction of potassium permanganate. $w_{n,k}$ is the weight from input “n” to node “k”; $w_k$ is the weight from node “k” to the output layer; $b_k$ is the bias value for node “k”; $b_{K+1}$ is the bias value for the output layer node.
Figure 3. Flow diagram for methodological approach of the work.
**Figure 4.** A) Normal probability plot of measured residuals vs. Gaussian distribution. B) Empirical cumulative distribution function of residuals of ANN-1 model.

**Figure 5.** Outlier detection with robust regression and ECDF method presented for ANN-1.

**Figure 6.** Residuals of ANN-1 after applying bootstrap method for parameter estimation.
Figure 7. MAE of model predictions for the different ANN architectures tested, ranging from 1 to 7 hidden neurons.
Figure 8. Representation of uncertainty predictions for KMnO₄ demand time-series using Monte Carlo method for ANN1, ANN4 and ANN7, showing the Monte-Carlo simulation outputs, with mean value, 10th and 90th percentile, and the observed KMnO₄ demand for the period 13/5/2015-16/5/2015. Probability density of Monte-Carlo outputs on one date (15th May 2015) is also shown.

Figure 9. RI computed for ANN-1, ANN-4 and ANN7 models using Connection Weight and Partial Derivatives Method.
LIST OF TABLES

Table 1. Characterisation of Llobregat DWTP influent. Samples from period January 2015 -June 2018. (N=1101)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unit</th>
<th>Mean</th>
<th>St. Dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>ºC</td>
<td>16.2</td>
<td>6.0</td>
<td>1.2</td>
<td>27.7</td>
</tr>
<tr>
<td>TOC</td>
<td>mg·L⁻¹</td>
<td>3.35</td>
<td>0.81</td>
<td>1.32</td>
<td>7.96</td>
</tr>
<tr>
<td>Turbidity</td>
<td>NTU</td>
<td>41.9</td>
<td>35.8</td>
<td>1.3</td>
<td>420</td>
</tr>
<tr>
<td>UV254</td>
<td>Abs·m⁻¹</td>
<td>6.92</td>
<td>1.87</td>
<td>0.07</td>
<td>23.2</td>
</tr>
<tr>
<td>Color</td>
<td>Pt·Co⁻¹</td>
<td>12.91</td>
<td>8.01</td>
<td>0.81</td>
<td>97.61</td>
</tr>
<tr>
<td>Electrical</td>
<td>µS·cm⁻¹</td>
<td>1371</td>
<td>236</td>
<td>542</td>
<td>2920</td>
</tr>
<tr>
<td>Conductivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflow</td>
<td>m³·s⁻¹</td>
<td>1.92</td>
<td>0.71</td>
<td>0.45</td>
<td>3.30</td>
</tr>
</tbody>
</table>
Table 2. Effects of outlier treatment on different model performance statistics on ANN-1, using nonlinear least squares method for parameter estimation.

<table>
<thead>
<tr>
<th></th>
<th>Original dataset</th>
<th>Outlier treated dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAE</strong></td>
<td>0,15</td>
<td>0,13</td>
</tr>
<tr>
<td><strong>MSE</strong></td>
<td>0,04</td>
<td>0,03</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td>0,20</td>
<td>0,16</td>
</tr>
<tr>
<td><strong>ρ² (Spearman)</strong></td>
<td>0,73</td>
<td>0,78</td>
</tr>
</tbody>
</table>

Table 3. Statistics used for evaluating the model performance for ANN-1, ANN-4 and ANN-7.

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of Parameters</th>
<th>Dataset</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>ρ² (Spearman)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN-1</td>
<td>7</td>
<td>Calibration</td>
<td>0,13</td>
<td>0,03</td>
<td>0,16</td>
<td>0,77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Validation</td>
<td>0,13</td>
<td>0,03</td>
<td>0,16</td>
<td>0,80</td>
</tr>
<tr>
<td>ANN-4</td>
<td>25</td>
<td>Calibration</td>
<td>0,13</td>
<td>0,03</td>
<td>0,16</td>
<td>0,80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Validation</td>
<td>0,13</td>
<td>0,03</td>
<td>0,16</td>
<td>0,80</td>
</tr>
<tr>
<td>ANN-7</td>
<td>43</td>
<td>Calibration</td>
<td>0,12</td>
<td>0,03</td>
<td>0,16</td>
<td>0,80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Validation</td>
<td>0,13</td>
<td>0,03</td>
<td>0,16</td>
<td>0,81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KMnO₄ Dosing rate</th>
<th>mg·L⁻¹</th>
<th>0.87</th>
<th>0.29</th>
<th>0.40</th>
<th>2.00</th>
</tr>
</thead>
</table>