Journal of Environmental Chemical Engineering 3 (2015) 2829–2838

Contents lists available at ScienceDirect

Journal of Environmental Chemical Engineering

journal homepage: www.elsevier.com/locate/jece

Coagulation modeling using artificial neural networks to predict both turbidity and DOM-PARAFAC component removal

Marla J. Kennedy[a,*], Amir H. Gandomi[a,b], Christopher M. Miller[a]

[a] Department of Civil Engineering, The University of Akron, Akron, Ohio 44325, United States
[b] BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, Michigan 48824, United States

ARTICLE INFO

Article history:
Received 30 July 2015
Received in revised form 6 October 2015
Accepted 7 October 2015
Available online 20 October 2015

Keywords:
Artificial neural networks
Fluorescence
PARAFAC
Coagulation Modeling
Turbidity

ABSTRACT

In this study, four different neural network models were evaluated for predicting both turbidity and dissolved organic matter (DOM) removal during the coagulation process at the Akron Water Treatment Plant (Akron, Ohio, USA). DOM was monitored and characterized using fluorescence spectroscopy and parallel factor (PARAFAC) analysis, building upon previous research which identified three unique fluorescence components (C1, C2, and C3). Neural network models were built using operational data to predict each of the fluorescence components and turbidity after coagulation based on variable raw water quality and chemical doses. Correlation coefficients between measured and model predicted values for the final turbidity, C1, C2, and C3 models on an unseen test data set were 0.91, 0.95, 0.97, and 0.51, respectively. The predictive capability of the top performing model for each parameter was evaluated using parametric analysis, external validation criteria, and the absolute relative error distribution. Results suggest that the models for settled turbidity and the three settled component scores are valid and can be used to predict the removal of individual fractions of DOM (as measured by PARAFAC components) as a function of chemical dose and raw water quality, providing the water plant the ability to simultaneously manage two key water quality treatment objectives.

1. Introduction

In recent years, artificial neural networks (ANN) have emerged as a viable method for modeling complex water treatment processes [1–6]. ANN offers several advantages over traditional modeling approaches (e.g. linear regression) (i) The associations between inputs and outputs are "learned" from historical data without having to specify the form of the model, (ii) ANN are robust to noisy or discontinuous data, (iii) It is not necessary to have a detailed understanding of the processes, only an understanding of the factors that influence the processes, and (iv) Increases in computer processing speeds have reduced the time needed to train and evaluate these models [2,7,8].

Several different ANNs have been reported in the literature including multi-layer perceptron (MLP) [9,10], radial basis function (RBF) [11,12], and generalized regression neural network (GRNN) [13]. MLP are feedforward networks consisting of one input layer, one output layer, and one or more hidden layers with a variable number of nodes whose activation functions are often sigmoid or log sigmoid. These networks are typically trained using back-propagation algorithms. The training process for MLP networks is more computationally and time intensive than the training process for GRNN and RBF networks, which occur in one pass. MLP models also have a tendency to become stuck in local minima during training. To avoid this, the weights of the MLP can be initialized using global search methods including simulated annealing (SA) and genetic algorithms (GA). This ensures that the weights converge to the global optimum [14,15].

RBF and GRNN models include an input layer, a hidden layer, and a linear output layer. Gaussian distributions are commonly used as the activation functions in the hidden radial basis function nodes. In both models, each input is shown to every node in the hidden layer and the output signal from each node is calculated based on the Euclidean distance between the input vector and the center of the node. This signal decreases exponentially with distance from the node with the rate of decrease depending on the width of the RBF. In the linear layer of the RBF, the output of the radial basis function nodes are multiplied by the weights and a bias term is added, while in the linear layer of the GRNN model, the dot product of the weights and the output from the hidden nodes is computed. When designing RBF and GRNN models, the centers and widths of the radial basis functions must first be selected, then the

* Corresponding author at: Auburn Science and Engineering Center (ASEC), 210, The University of Akron, Akron, OH 44325, United States. Fax: +1 330 972 6020.
E-mail address: mjk81@zips.uakron.edu (M.J. Kennedy).

http://dx.doi.org/10.1016/j.jece.2015.10.010
2213-3437/

weights are optimized [16]. The centers can be selected as individual training vectors, but to improve the generalization of the model, centers can be selected using clustering algorithms including sub-sampling, *K*-means, or Kohonen self-organizing maps [16].

The primary goal of the coagulation process is to promote the aggregation of small particles for subsequent removal during sedimentation and/or filtration processes [17]. Traditionally, optimization of the coagulation process has focused on turbidity removal. Turbidity is used a surrogate measure of health related contaminants in water, including pathogens [18]. High turbidity water can also impede the disinfection process, effectively shielding organisms from exposure to disinfectant [17]. Turbidity removal is dependent on coagulant type and dose as well as raw water quality parameters such as turbidity, temperature, pH, and alkalinity [17]. In practice, coagulant doses are routinely determined using jar tests; however, several researchers have developed plant specific models for predicting turbidity removal and optimum coagulant dosing based on raw water quality parameters [1,2,19,20].

As a requirement of the Stage 1 Disinfectant and Disinfection by Products Rule, many water treatment plants have begun to practice enhanced coagulation. During enhanced coagulation, treatment focuses on the removal of total organic carbon (TOC), rather than turbidity, as a means to reduce disinfection byproduct formation potential. TOC is used as a bulk measure of the amount of dissolved organic matter (DOM) in the water. DOM is a complex mixture of organic compounds that can vary temporally and geographically, and can encompass a range of physical and chemical properties [21–23]. Coagulation performance is highly dependent on the nature of DOM, which cannot be determined using TOC measurements, as well as coagulant type and dose, coagulation pH, turbidity, and alkalinity [24–26].

Given the need for a measure more specific than DOC/TOC, fluorescence spectroscopy has emerged as a rapid, cost effective approach for characterizing DOM and monitoring DOM removal [27,28]. During fluorescence analysis, the fluorescence intensity of a sample is measured over a range of excitation and emission wavelengths and recorded as a three dimensional excitation emission matrix (EEM). This EEM can contain hundreds of data points, and must be processed to derive meaningful information. Parallel Factor Analysis (PARAFAC) of the EEM can be used to identify unique chemical components in the sample, assuming that the fluorescence signal at any excitation/emission pair is the result of the combination of these components [29,30]. Using PARAFAC analysis, humic-terrestrial groups, protein-like groups, and microbial-humic groups have been identified and correlated with dissolved and/or total organic carbon concentrations (DOC) and used to assess the coagulation process based on DOC removal [27,31–33]. However, the effectiveness of the coagulation process depends on the presence and concentration of the individual organic groups, as they are not all efficiently removed during coagulation [24,25,34].

The purpose of this research was to develop robust models that can be used evaluate turbidity and DOM (via PARAFAC components) removal during coagulation. Four different ANN models were trained for each of four predicted parameters (turbidity and three PARAFAC components) using operational data provided by Akron Water Treatment Plant (Akron, Ohio, USA). The models were evaluated using a combination of statistical, parametric, and sensitivity analyses. This research is the first attempt to model both turbidity removal and the change in DOM-PARAFAC components during the coagulation process in a full-scale scenario based on chemical doses and variable raw water quality.

2. Material and methods

2.1. Data collection and preprocessing

The data utilized for model development was provided by Akron WTP in Akron, Ohio. As shown in Fig. 1, Akron WTP uses a combination of potassium permanganate, powdered activated carbon (PAC), chlorine dioxide, and aluminum sulfate (alum) prior to flocculation/sedimentation. Model data, provided by plant personnel, included daily measurements of raw and settled water quality as well as chemical dose information. The raw water quality measures reflect conditions prior to chemical addition. The data utilized for the turbidity model encompassed a four year period from November 2009 through October 2013 and included 1338 individual daily measurements of raw and coagulated turbidity, coagulant dose, PAC dose, chlorine dioxide dose, raw water hardness, raw water alkalinity, raw water temperature, and raw water pH.

Data utilized for the fluorescence component models was collected from March 2014 through January 2015. During this time, a total of 428 samples of raw and coagulated water were collected for fluorescence analysis. Samples were filtered using a 0.45 μm filter and pH adjusted (pH ~3) [28]. EEM and absorbance measurements were collected at Akron WTP using a Horiba-Jobin Yvon (Irvine, California) Aqualog® benchtop fluorometer. Excitation wavelengths were set at 220–800 nm in 2 nm increments and emission spectra were recorded from 249 to 827 nm in 4.66 nm increments. The integration time was 10 s. Samples were corrected for first and second order Rayleigh and Raman scatter and inner filter effects using the software included with the Aqualog® and normalized to Raman units [35].

The PARAFAC models were built in Matlab v. 8.3 using the N-way v. 3.00 Toolbox [36,37]. Details of the development of the PARAFAC model for the City of Akron WTP were presented in detail elsewhere [34]. Briefly, prior to model development, EEMs were preprocessed to exclude signals at excitation wavelengths less than 224 nm, and signals in the regions of Rayleigh and Raman scatter

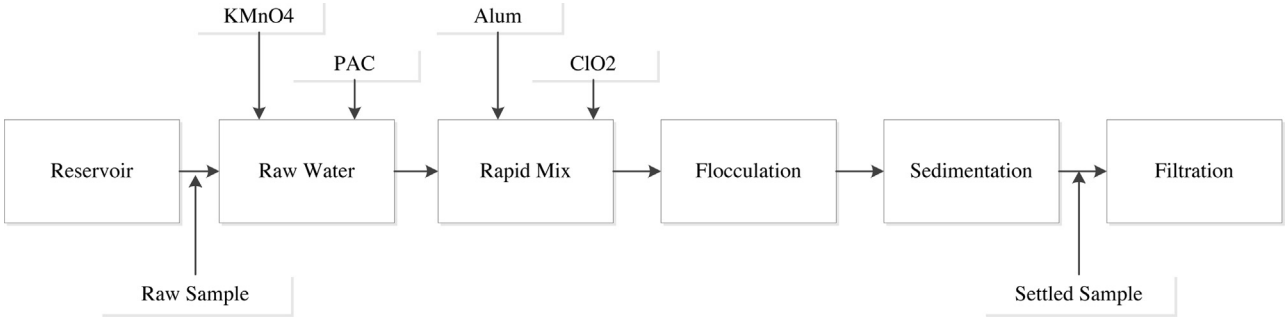

Fig. 1. Akron WTP treatment train with chemical dose application points and sample points.