Journal of Geochemical Exploration xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Applicability of a processes-based model and artificial neural networks to estimate the concentration of major ions in rivers

Clemêncio Nhantumbo[a,b,*], Frede Carvalho[c], Cintia Uvo[a], Rolf Larsson[a], Magnus Larson[a]

[a] Lund University, Water Resources Engineering, Box 118, SE-221 00 Lund, Sweden
[b] Eduardo Mondlane University, Faculty of Engineering, Av. de Moçambique km 1.5, C. Postal 257 Maputo, Mozambique
[c] Federal University of Alagoas, Academic Center of Technology, Av. Lourival Melo Mota, S/N Tabuleiro do Martins, ZIP: 57072-900 Maceó, AL, Brazil

ARTICLE INFO

Keywords:
River monitoring
Water quality
Major ions
Model comparison and neural networks

ABSTRACT

Modelling is an alternative solution to reduce the cost of water quality monitoring. Commonly, concentration of pollutants is estimated based on limited sampling information. Concentration of ions in rivers can be estimated using modelling strategies that involve statistics and artificial intelligence as well as the understanding of physical processes. Therefore, the performance of feedforward neural networks that employs the Levenberg-Marquardt optimization method was compared to the PPBM recently proposed. Both ANN and PPBM were used to estimate the concentration of major ions (Na^+, K^+, Mg^{2+}, Ca^{2+}, HCO_3^-, SO_4^{2-}, Cl^-, and NO_3^-) in river water based on pH, alkalinity, and temperature. Root-mean-square error and Pearson correlation coefficient (R) together with its p-value were used to evaluate the quality of results of both models. The ANN model provides better estimates compared to the PPBM in most cases. However, the PPBM has the possibility to evaluate its predictions by using the difference between the estimated and measured electrical conductivity. If the predictions are not good the PPBM can be recalibrated, whereas the ANN model is limited in this respect. Another disadvantage of ANN models is that they are developed based on historical data and if limited data are available, such models cannot be used. This latter disadvantage makes the PPBM superior in developing countries, where often little or no consistent historical data exist.

1. Introduction

Water quality monitoring is a costly activity that cannot be afforded by most developing countries, at least not at a sufficient level of detail. However, there are several measures that can be adopted to reduce the cost of water quality monitoring, where still a reasonable understanding of the water quality changes in time and space is acquired (Nhantumbo et al., 2015, 2016). Such measures include the usage of monitoring strategies and predictive models to improve and optimize the data collection and interpretation; thus, extensive sampling and laboratory analysis may be reduced (Nhantumbo et al., 2015).

Models used to optimize water quality monitoring can be developed based on physical processes, traditional statistics, and artificial intelligence methods. In some cases, these modelling approaches can be combined. Statistical and artificial intelligence models require comprehensive historical data to be used, whereas physical processes based models (PPBMs) do not, or require less historical data (Razi and Athappilly, 2005).

Despite the fact that some studies have shown that the artificial neural networks (ANN) tend to overfit and question its applicability outside the range of training data, most studies argue that artificial intelligence models such as ANN perform better than traditional statistical regression analysis (Abebe et al., 2000; Zahedi et al., 2009; Paliwal and Kumar, 2009a, 2009b). ANN models are particularly useful when the physical relationships between the variables are not known (Paliwal and Kumar, 2009a, 2009b). PPBMs do not rely on historical data for their development, besides at the validation stage; thus, once developed, they can be used at any location. However, PPBMs require a deep understanding of phenomena and processes governing the water quality and their applicability is often a function of the simplifications introduced, liming them to certain conditions (Ciou et al., 2007).

A PPBM was recently developed that estimates the concentration of major ions in rivers using as input data on the pH, alkalinity, and temperature (Nhantumbo et al., 2016). The model has two options to estimate the concentration of major ions, denoted as the generalized and customized methods. The customized method uses specific river data to calibrate the model and it provides typically more accurate predictions (Nhantumbo et al., 2016).

* Corresponding author at: Lund University, Water Resources Engineering, Box 118, SE-221 00 Lund, Sweden.
E-mail address: clemencio.nhantumbo@tvrl.lth.se (C. Nhantumbo).

https://doi.org/10.1016/j.gexplo.2018.07.003
Received 24 April 2017; Received in revised form 17 May 2018; Accepted 3 July 2018

The aim of this paper is to evaluate the performance of ANN in comparison with the PPBM developed in for estimating the concentration of major ions in rivers using data from four stations in Swedish rivers. In other applications, ANN models have been demonstrated to perform better simulations of natural systems compared to traditional statistical models (Paliwal and Kumar, 2009a, 2009b). The model comparison will also be used as a basis for recommendations on the situations in which it might be preferable to apply ANN or PPBM.

2. Background

Modelling, in a wide context, has previously been used as an aid in both surface and ground water quality monitoring (Honti et al., 2017; Amanollahi and Kaboodvandpour, 2016). Most of the surface water models have been developed with focus on monitoring pollutants from agriculture and municipal wastewater. However, modelling has also been carried out for special pollutants generated in connection with mining activity (Valente et al., 2013). An example of such modelling employing statistical techniques is estimating the concentration of metals from acid mine drainage based on pH, electrical conductivity, and sulfate using a fuzzy interface system (Valente et al., 2013). Another example involves the usage of PPBM to estimate the concentration of major ions in rivers (Nhantumbo et al., 2016).

Models to predict the concentration of major ions are less developed compared to the models related to oxygen concentration, organic compounds, and nutrient leaching and transport (Scavia et al., 2017). In such applications, both ANN and PPBMs can potentially be used to increase the sustainability of water quality monitoring programs by reducing extensive sampling and analysis, which make these models particularly suitable for countries with limited resources.

2.1. Physical processes based model (PPBM)

A PPBM to predict the concentrations of major ions in a river was developed combining the following theories and relationships: i) carbonate equilibrium; ii) total alkalinity; iii) conductivity of salts in water; and iv) calibration using baseline-data (Nhantumbo et al., 2016). The model has two options to predict the concentrations of major ions: (1) a generalized method that uses continental averages of relative concentrations of major ions available in the database of the model, and (2) a customized method that requires river-specific baseline-data for calibration (Nhantumbo et al., 2016).

The baseline data include the concentration of the specific major ions, converted to eq/l, in percentage of the total concentration of positive ions or total concentration of negative ions depending on whether this ion is positive or negative, respectively. These concentrations are denoted as relative concentrations of major ions. In the development of the model it was assumed that the concentrations do not change over time. It was shown that although there are variations, these variations are typically not significant for periods smaller than 10 years (Nhantumbo et al., 2016).

Combining the theory of alkalinity and carbonate equilibrium made it possible to determine the concentration of carbonate species, hydrogen, and hydroxide ions. The relative concentrations of major ions and the concentration of bicarbonate, from carbonate species, yielded estimates of the concentration of major ions. Finally, the theory of conductivity was used to check the accuracy of the model introducing a parameter called (DiffEC), which is the difference between the electrical conductivity calculated using the concentrations of ions estimated by the model and the measured electrical conductivity (Nhantumbo et al., 2016). The model also estimates the equilibrium concentration of some minor ions, Fig. 1.

For the customized method, the value on DiffEC should be higher than zero during a base period used for calibration and it should not attain values higher than ± 20%. For the general method, DiffEC should be close to zero and reasonable estimates are provided by the model, if its value is between −15% and +40. Keeping the value of DiffEC within these limits, there is 80% probability of having RMSE (%) below 15% (Nhantumbo et al., 2016).

2.2. Artificial neural networks (ANN)

ANN have been applied for modelling in many fields ranging from signal processing in telecommunications to pattern recognition in medicine, business, and engineering (Razi and Athappilly, 2005). ANN are applied in medical disciplines for different purposes, for example in cardiovascular studies (Itchhaporia et al., 1996) and for decision support in cancer studies (Lisboa and Taktak, 2006). In business and financing ANN are applied in forecasting stock market returns (Enke and Thawornwong, 2005; Kim, 2006). Neural networks are also applied in engineering, for example, in water quality monitoring (Kim et al., 2006).

Models of ANN are divided in two categories, the feedforward neural networks and the recurrent neural networks (Rankovic et al., 2010). The most popular and widely used is the feedforward neural networks model, which propagate data linearly from input to output, Fig. 2.

The processing elements in neural networks are the neurons (Rankovic et al., 2010). Each neuron receives one or more inputs and processes them to generate a single output (Razi and Athappilly, 2005). The main elements that take part in processing information in ANN are: inputs (x), weights (ω), summation function (Σ), transformation or activation function (f), and outputs.

The neural network finds its parameters by minimizing the difference between the modelled values y_{calc} and the desired output y_{meas}, the so-called performance function (Rankovic et al., 2010). A simple example of a performance function is given by Eq. (1). The minimum of the function is calculated using optimization methods.

$$P = -\frac{1}{2}(y_{meas} - y_{calc})^2 \quad (1)$$

Normally the artificial neural networks are composed by layers of neurons, the hidden and output layers. The number of neurons is not necessarily equal to the number of input variables. The rule of thumb is that the number of neurons in the hidden layers is equal to the number of input parameters plus one and the output layer is composed by a single neuron (Beale et al., 2015). Lippmann also proposes that the number of hidden neurons should be $s(i + 1)$, where s is the number of output parameters and i is the number of input parameters (Tatibana and Kaetsu, 2017).

3. Methods

Data from the four monitoring stations in Swedish rivers were used to compare the estimates obtained by the two models. The river and the station names, together with the geographical coordinates of the selected stations are given as follow, Skellefte älv, Slagnas (lat: 65.675; long: 18.146); Vindelälven, Maltbrännan (lat: 64.576; long: 19.291); V. Dalälven, Mockfjärd (lat: 60.483; long: 14.897); and Klarälven, Edsforsen (lat: 60.066; Long: 13.526) (Nhantumbo et al., 2016). The stations were selected because they have comprehensive and high-quality data to evaluate model performance.

Using data from the selected stations, the correlation between the input parameters (pH, alkalinity, and temperature) with the output parameters (major ions) was evaluated to check which input parameters are strongly correlated with the output to support the discussion of the results.

In order to do a fair comparison between the models the same input data were used. The input parameters for both models were pH, alkalinity, and temperature. This is because the theoretical model was developed to use pH, alkalinity, and temperature as input parameters and it is not possible to change this. Fig. 3 shows the schematic