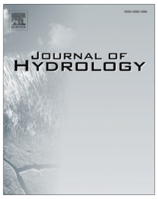
Journal of Hydrology 519 (2014) 1895–1907

Contents lists available at ScienceDirect

Journal of Hydrology

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



Modelling of dissolved oxygen in the Danube River using artificial neural networks and Monte Carlo Simulation uncertainty analysis

Davor Antanasijević [a,*], Viktor Pocajt [b], Aleksandra Perić-Grujić [b], Mirjana Ristić [b]

[a] University of Belgrade, Innovation Center of the Faculty of Technology and Metallurgy, Karnegijeva 4, 11120 Belgrade, Serbia
[b] University of Belgrade, Faculty of Technology and Metallurgy, Karnegijeva 4, 11120 Belgrade, Serbia

ARTICLE INFO

Article history:
Received 14 June 2014
Received in revised form 27 August 2014
Accepted 1 October 2014
Available online 13 October 2014
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Fi-John Chang, Associate Editor

Keywords:
DO
GRNN
MCS
VIF
Genetic algorithm
Correlation analysis

SUMMARY

This paper describes the training, validation, testing and uncertainty analysis of general regression neural network (GRNN) models for the forecasting of dissolved oxygen (DO) in the Danube River. The main objectives of this work were to determine the optimum data normalization and input selection techniques, the determination of the relative importance of uncertainty in different input variables, as well as the uncertainty analysis of model results using the Monte Carlo Simulation (MCS) technique. Min–max, median, *z*-score, sigmoid and tanh were validated as normalization techniques, whilst the variance inflation factor, correlation analysis and genetic algorithm were tested as input selection techniques. As inputs, the GRNN models used 19 water quality variables, measured in the river water each month at 17 different sites over a period of 9 years. The best results were obtained using min–max normalized data and the input selection based on the correlation between DO and dependent variables, which provided the most accurate GRNN model, and in combination the smallest number of inputs: Temperature, pH, HCO_3^-, SO_4^{2-}, NO_3-N, Hardness, Na, Cl^-, Conductivity and Alkalinity. The results show that the correlation coefficient between measured and predicted DO values is 0.85. The inputs with the greatest effect on the GRNN model (arranged in descending order) were *T*, pH, HCO_3^-, SO_4^{2-} and NO_3-N. Of all inputs, variability of temperature had the greatest influence on the variability of DO content in river body, with the DO decreasing at a rate similar to the theoretical DO decreasing rate relating to temperature. The uncertainty analysis of the model results demonstrate that the GRNN can effectively forecast the DO content, since the distribution of model results are very similar to the corresponding distribution of real data.

1. Introduction

Programs that monitor water quality help to understand various processes that have an impact on the overall quality of water and provide necessary information for the management of water resources in general. The quality of a water body is usually described by sets of physical, chemical and biological variables that are mutually interrelated (Khalil et al., 2010). The river waters have been contaminated as a result of the discharges from wastewater containing degradable organics, nutrients, domestic effluent, and agricultural waste (Dimitrovska et al., 2012). All of the aforementioned contaminants directly or indirectly negatively affect key river quality parameters such as dissolved oxygen (DO) content, temperature, pH, conductivity, transparency, viscosity and total dissolved solids. Among them, the DO is the most severely affected, since the diffusion of oxygen into the river body (re-aeration) is an inherently slow process. This in turn puts additional strain on the other very important contributor to DO, namely the generation of oxygen from photosynthetic aquatic plants (Araoye, 2009). Furthermore, the above-mentioned water contamination, among other parameters (e.g. the amount of light, species and abundance of plants), also influence the factors which control the rate of photosynthesis, which makes the quantification of DO content in rivers one of the primary concerns for water resource managers (Wen et al., 2013).

Water quality modeling as a basis for water pollution control are commonly used to predict trends in water quality based on current water conditions, including pollutant concentrations (Najah et al., 2011). The major issue in the application of water quality models, such as IWA River Quality Model No. 1 (Reichert et al., 2001), QUAL2K (Chapra and Pelletier, 2003), WASP6 (Wool et al., 2006), is the requirement for more information regarding the river system than is often available (Mannina and Viviani, 2010). A constant need for less complex models for the DO forecasting led to the application of artificial neural networks

[*] Corresponding author. Tel.: +381 11 3303 642; fax: +381 11 3370 387.
E-mail address: dantanasijevic@tmf.bg.ac.rs (D. Antanasijević).

http://dx.doi.org/10.1016/j.jhydrol.2014.10.009
0022-1694/© 2014 Elsevier B.V. All rights reserved.

(ANN) in this field (Chang et al., 2013; Chen and Chang, 2009). The advantage of ANNs over deterministic models is that they require less data and they are well suited for forecasting (Kisi et al., 2012). In addition, the ANN approach does not require a complex and explicit description of the underlying process in a mathematical form (Nayak et al., 2005). The design of ANNs originated from a desire to emulate human learning, which led to the application of massive parallel, distributed processing and computing techniques inspired by biological neuron processing. ANNs are proved to be highly effective for modeling non-linear problems, with application to diverse large-scale problems (Banerjee et al., 2011).

Successful application of ANN models for the forecasting of DO is associated with several challenges, the key issues being proper data normalization and the selection of the model inputs that have the most significant impact on model performance. Employing a large number of inputs to an ANN model usually increases the network size, resulting in a decrease in processing speed, a reduction in the efficiency of the network (Arhami et al., 2013), and also may ultimately result in a model that is not suitable as a practical forecasting tool. One of the important subjects in ANN modeling studies is the analysis of uncertainty and the influence of input data uncertainty on the model results. The term uncertainty refers to lack of knowledge or information on the models, parameters, constants, input data, and beliefs/concepts. Information on the total model uncertainty, for models which support decision-making, is essential and it is as important as the modeling results themselves (Borrego et al., 2008). The Monte Carlo Simulation (MCS) technique is a widely used method for the analysis of uncertainty in hydrological modeling and it allows the quantification of the model output uncertainty resulting from uncertain model parameters, input data or model structure (Shrestha et al., 2009).

In recent years, considerable progress has been made in the development of ANN models for the forecasting of DO. Some examples of the application of ANNs for the modeling of DO at a single location, include models developed for the Melen River, Turkey (Samandar, 2010), Bow River, Canada (He et al., 2011), Foundation Creek in Colorado, USA (Ay and Kisi, 2012), the Danube River in Bezdan, North Serbia (Antanasijević et al., 2013a) and the Upper Klamath River in Oregon, USA (Heddam, 2014). In those papers, the authors tested a variety of ANN architectures (feed-forward, recurrent, radial basic and general regression neural network), applied for various periods of time, as well as using different data representations (for details please see Appendix Table A1). In contrast, the application of ANNs for the modeling of DO across multiple sites was limited only to the use of multilayer perceptron (for examples and details see Appendix Table A2).

In this paper, we propose an integrated ANN model, based on the general regression neural network (GRNN) architecture, for the forecasting of DO across multiple sites; the model is in this instance applied to all monitoring stations located on the Danube River, covering its 588-km course through the territory of Serbia. Different methods for data normalization and input selection were in order to enhance the performance of the model and to reduce the number of inputs needed for DO forecasting. The performance of the created ANN models were analyzed using multiple statistical metrics. Finally, the impact of input data uncertainty on the model output and the analysis of uncertainty of the results were performed using the Monte Carlo Simulation (MCS) technique.

2. Materials and methods

2.1. Study area and water quality data

The Danube is the longest river on the Balkan Peninsula and the second longest river in Europe, after the Volga. It is an international waterway that connects Germany, as well as other Central European and Balkan countries with the Black Sea. The Danube flows for 2857 km and passes through or touches the borders of ten countries: Germany, Austria, Slovakia, Hungary, Croatia, Serbia, Bulgaria, Romania, Ukraine, and Moldova (ICPDR, 2014). Around 10% of its basin is located in Serbia and on its 588-km course the quality of river water is monitored at 17 separate monitoring stations (Fig. 1).

The dataset used in this study has been generated through continuous monitoring of the water quality of the Danube River in the territory of the Republic of Serbia. The water quality was monitored regularly each month (monthly or semi-monthly) at 17 different sites over a period of 9 years (2002–2010) and the data was obtained from the Serbian Agency for Environmental Protection (SEPA, 2013). The availability of data, number of data patterns (input vectors) per year and number of data patterns per site for the studied period are presented in Table 1. There were between 131 and 252 available patterns per year, while the number of available patterns per site was between 53 and 128.

All water samples collected during the study period were analyzed for a large number of different water quality parameters, from which, 19 were selected as inputs for the model (Table 2). In total, the dataset contained 1512 data patterns with 20 water quality parameters, which provided more than 30,000 individual data points. The basic statistics of the selected input/output parameters are presented in Table 2.

2.2. ANN architecture

An artificial neural network, which employs the model structure of a biological neural network, is a very powerful computational technique for modeling complex non-linear relationships particularly in situations where the explicit form of the relationship between the variables involved is unknown (Singh et al., 2009). The basic and the most commonly used ANN architecture consists of an input layer, a series of hidden layers and an output layer. Each of these layers consists of a number of interconnected neurons (processing units). In this study, the ANN architecture known as general regression neural network (GRNN) was used, since it proves to be an effective alternative to the basic Feed-forward ANNs (Heddam, 2014). The GRNN is based on the non-linear regression theory and is a universal approximator for smooth function. It consists of four layers, which are presented in Fig. 2.

GRNN is a one-pass supervised learning network, which means that weights (W_{ij} and W_{S1}) between neurons in different layers are determined by the values of variables, there W_{ij} are weights defined by the ith input variable and the jth training pattern, while W_{S1} is equal to the values of output variable. In this architecture, the number of neurons in the input and output layer corresponds to the number of input and output variables, while the number of pattern neurons is equal to the number of data patterns. The number of neurons in the summation layer can be expressed as $N_o + 1$, where N_o is the number of output neurons. In this case, the pattern neurons are connected to two neurons in the summation layer, since the model has only one output.

Being a supervised network, the GRNN basically measures the distance (D_j) of the training patterns in N-dimensional space, where N is the number of inputs, and estimates the output accordingly (Hanna et al., 2007). The calculated D_j, e.g. Euclidean distance (1), is then processed using an exponential activation function (2).

$$D_j = \sqrt{\sum_{i=1}^{n}(w_{ij} - x_i)^2} \qquad (1)$$

$$f(D_j) = \exp\left(\frac{-D_j}{2\sigma^2}\right) \qquad (2)$$