Journal of Environmental Management 143 (2014) 8–16

Contents lists available at ScienceDirect

Journal of Environmental Management

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

A hybrid evolutionary data driven model for river water quality early warning

Alejandra Burchard-Levine [a,*], Shuming Liu [b], Francois Vince [c,1], Mingming Li [d], Avi Ostfeld [e,2]

[a] Tsinghua University – Veolia Environnement Joint Research Center for Advanced Environmental Technology, School of Environment, Tsinghua University, Beijing, China
[b] School of Environment, Tsinghua University, Beijing, China
[c] Veolia Environnement, 8 North Dongsanhuan Road, Beijing 100004, China
[d] School of Environment, Tsinghua University, Beijing, China
[e] Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

A R T I C L E I N F O

Article history:
Received 1 July 2013
Received in revised form
20 January 2014
Accepted 21 April 2014
Available online 13 May 2014

Keywords:
Genetic algorithm
Artificial Neural Networks
Water quality
Early warning system
China

A B S T R A C T

China's fast pace industrialization and growing population has led to several accidental surface water pollution events in the last decades. The government of China, after the 2005 Songhua River incident, has pushed for the development of early warning systems (EWS) for drinking water source protection. However, there are still many weaknesses in EWS in China such as the lack of pollution monitoring and advanced water quality prediction models. The application of Data Driven Models (DDM) such as Artificial Neural Networks (ANN) has acquired recent attention as an alternative to physical models. For a case study in a south industrial city in China, a DDM based on genetic algorithm (GA) and ANN was tested to increase the response time of the city's EWS. The GA-ANN model was used to predict NH_3–N, COD_{mn} and TOC variables at station B 2 h ahead of time while showing the most sensitive input variables available at station A, 12 km upstream. For NH_3–N, the most sensitive input variables were TOC, COD_{mn}, TP, NH_3–N and Turbidity with model performance giving a mean square error (MSE) of 0.0033, mean percent error (MPE) of 6% and regression (R) of 92%. For COD, the most sensitive input variables were Turbidity and COD_{mn} with model performance giving a MSE of 0.201, MPE of 5% and R of 0.87. For TOC, the most sensitive input variables were Turbidity and COD_{mn} with model performance giving a MSE of 0.101, MPE of 2% and R of 0.94. In addition, the GA-ANN model performed better for 8 h ahead of time. For future studies, the use of a GA-ANN modelling technique can be very useful for water quality prediction in Chinese monitoring stations which already measure and have immediately available water quality data.

1. Introduction

China's fast pace industrialization, rapid population growth and increase in urbanization are some of the main reasons for continuous occurrences of river water pollution contamination events. Cities situated downstream of vulnerable rivers and dependent on these for drinking water supply need to have appropriate early warning systems (EWS) in order to protect citizens from receiving contaminated drinking water. Although the Chinese government has emphasized on the development and installation of early warning systems, accidents have continued to occur at high cost and affecting the health of large populations. Some of the recent incidents are the blast of Jihua petroleum factory upstream of the Songhua River in Jilin Province, in November 2005 (Hu, 2009), and the cadmium spill which contaminated over 100 km of the Longjiang River in southwestern China's Guangxi region, in January 2012 (Cadmium spill threatens water supply in Liuzhou, China, 2012).

The major issue is that EWS for drinking water protection are still incomplete in many areas of China. A complete EWS namely

* Corresponding author.
E-mail addresses: burchard-levin10@mails.tsinghua.edu.cn, alejandra.burchard@gmail.com (A. Burchard-Levine), shumingliu@tsinghua.edu.cn (S. Liu), francois.vince@veolia.com (F. Vince), limingming.china@gmail.com (M. Li), ostfeld@tx.technion.ac.il (A. Ostfeld).
[1] Tel.: +86 10 5953 2700.
[2] Tel.: +972 4 8292782; fax: +972 4 8228898.

http://dx.doi.org/10.1016/j.jenvman.2014.04.017
0301-4797/

contains detection, confirmation and characterization components. From the aforementioned, a challenging component is the characterization one which usually consists of advanced mathematical models which can predict the fate of pollution events. This component would allow water utilities to better understand the characteristics of a possible pollution event, eventually acquire more time to deal with it and use appropriate measures.

In the context of drinking water source protection and EWS, the specific type of modelling of interest involves the modelling of transient water quality conditions associated with contamination events and their potential impact on drinking water intakes (Grayman, 2001). Throughout the literature, many hydrological models whether a public tool or licensed tool can be used to model an accidental event and characterize its behaviour. There are many ways of classifying these hydrological models. For simplicity, these were divided into two categories (Burchard-Levine et al., 2012):

- Physically-based “white-box” models that calculate and integrate all the physico-chemical mechanisms occurring in the river to predict pollutant fate (Hao, 2006).
- Empirically-based or data-driven “black-box” models that aim at predicting pollutant fate using a purely numeric approach and without any detailed understanding of physico-chemical mechanisms.

Physical models for water contamination prediction and forecasting are very popular and useful when all the information needed is available. Meanwhile, data-driven models (DDM) for surface water quality predictions have gained a lot of attention in order to overcome challenges such as non-linearity which are not taken into account in physical models. In addition, in the case of China where data availability is a constraint, DDM, such as Artificial Neural Networks (ANN), can overcome this issue with its flexibility which allows it to predict and forecast using available data (Burchard-Levine et al., 2012). Burchard-Levine et al. (2012) provide a more detailed review on the state of EWS and water quality modelling in China.

In a review by Maier et al. (2010), it is declared that ANN were used to predict water quantity variables in more than 90% of the papers, of which flow was by far the most popular. On the other hand, water quality variables were predicted in fewer than 10% of the papers. Hence, the application of ANN for water quality prediction is relatively recent. This fact is reflected in China where many applications of ANN based models have been studied for forecasting and prediction of river flows and watershed rainfall-runoff; fewer can be found for water quality studies. Zhao et al. (2007) studied a water quality, COD and DO, forecast model for the Yuqiao reservoir in Tianjin.

In most water resources ANN applications, very little attention has been given to the task of selecting appropriate model inputs (Maier and Dandy, 2000; Bowden et al., 2005). Bowden et al. (2005) presented a review of the current state of input determination methods in water resources applications of ANN modelling. To address the lack of a robust procedure for selecting important inputs, Bowden et al. (2005) presented and recommended two input determination methodologies out of which one of them was a Genetic Algorithm coupled with ANN for the selection of the final subset of important model inputs.

Genetic Algorithms are well suited for the task of selecting an appropriate combination of inputs to a model, as they have the ability to search through large numbers of combinations where interdependencies between variables may exist (Bowden et al., 2005). Kuo et al. (2006) used a hybrid neural GA for water quality management of the Feitsui Reservoir in Taiwan. Abrahart et al. (1999) used a genetic algorithm to optimise the inputs to an ANN model used to forecast runoff from a small catchment.

The objective of this research was to develop a water quality prediction model which would enhance the characterization component of China's EWS while taking into account data availability constraints of the country. The use of a DDM based on GA and ANN, referred to as the GA-ANN model, instead of a classical physical model, was explored and tested in order to increase the capacity for water utilities to characterize and increase the response time of a possible contamination event. The GA-ANN model was used to predict downstream physico-chemical water quality using available upstream physico-chemical water quality variables ahead of time. This could help to characterize whether the downstream water quality is good enough as a drinking water source. Bearing in mind, this method can be used for accidents only occurring upstream. A case study is presented for a large industrial city X in the south of China which depends on the River Z as the drinking water source. *The GA-ANN model was first tested for the prediction of* NH_3*–N,* COD_{mn} *and TOC at Station B using past measured water quality variables at Station A, located 12 km upstream, and Station B at minimum 2 h ahead of time.* Finally, the model was tested in order to see its performance for larger time delays, meaning its capacity to predict ahead of time.

2. Materials and methods

ANN are powerful adaptive tools which imitate the central nervous system by recognizing patterns in data and allow mimicking those patterns. ANN are able to forecast data without the need to formalize how these patterns work and what inner mechanisms/formula are involved (Adeli, 2001; Maier et al., 2010). One of the most important steps in the ANN development process is the determination of significant input variables. Usually, not all of the potential input variables will be equally informative since some may be correlated, noisy or have no significant relationship with the output variable being modelled.

In the application of ANN, it is always a challenge to select the appropriate input variables. Over-parameterization might lead to bad outputs. In this paper, GA is connected to ANN to optimize the water quality input variables and prevent over-parameterization in model training. Hence, this technique is referred to as the application of GA for optimizing the process of selection of input variables for an ANN based water quality prediction model. The following section describes the methodology to configure and test an ANN model which uses GA for input variable optimization: Raw data pre-processing, ANN model, GA input variable selection and final ANN model.

2.1. Raw data organization and pre-processing

Raw water quality data was received from the local environmental monitoring center of the southern Chinese City X in question. At first, the raw dataset received needs to be sorted and dates of measurements need to be matched. If one dataset measures a parameter every half hour while another measures it every 2 h, the average of four data points will be taken in order to match it with the corresponding 2 h measurements. In addition, it is quite usual that raw water quality data contain a lot of false measurements and noise. There are many ways and tools which exist to assess whether a measurement should be discarded, i.e. pre-process data. As a preliminary pre-processing step, all measurements above and below two standard deviations can be eliminated. In order to smooth the data, all missing and eliminated points can be replaced by the average of adjacent measurements.