Journal of Hydrology 548 (2017) 170–183

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Conjunction of wavelet transform and SOM-mutual information data pre-processing approach for AI-based Multi-Station nitrate modeling of watersheds

Vahid Nourani [a,c,*], Gholamreza Andalib [a], Dominika Dąbrowska [b]

[a] Department of Water Resources Engineering, Faculty of Civil Engineering, University of Tabriz, Tabriz, Iran
[b] Department of Hydrogeology and Engineering Geology, Faculty of Earth Sciences, University of Silesia, Poland
[c] Department of Civil Engineering, Near East University, P.O. Box: 99138, Nicosia, North Cyprus, Mersin 10, Turkey

ARTICLE INFO

Article history:
Received 18 November 2016
Received in revised form 21 February 2017
Accepted 1 March 2017
Available online 2 March 2017
This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Gokmen Tayfur, Associate Editor

Keywords:
Nitrate modeling
Wavelet
Self-Organizing Map
Mutual Information
Artificial intelligence
Little River watershed

ABSTRACT

Accurate nitrate load predictions can elevate decision management of water quality of watersheds which affects to environment and drinking water. In this paper, two scenarios were considered for Multi-Station (MS) nitrate load modeling of the Little River watershed. In the first scenario, Markovian characteristics of streamflow-nitrate time series were proposed for the MS modeling. For this purpose, feature extraction criterion of Mutual Information (MI) was employed for input selection of artificial intelligence models (Feed Forward Neural Network, FFNN and least square support vector machine). In the second scenario for considering seasonality-based characteristics of the time series, wavelet transform was used to extract multi-scale features of streamflow-nitrate time series of the watershed's sub-basins to model MS nitrate loads. Self-Organizing Map (SOM) clustering technique which finds homogeneous sub-series clusters was also linked to MI for proper cluster agent choice to be imposed into the models for predicting the nitrate loads of the watershed's sub-basins. The proposed MS method not only considers the prediction of the outlet nitrate but also covers predictions of interior sub-basins nitrate load values. The results indicated that the proposed FFNN model coupled with the SOM-MI improved the performance of MS nitrate predictions compared to the Markovian-based models up to 39%. Overall, accurate selection of dominant inputs which consider seasonality-based characteristics of streamflow-nitrate process could enhance the efficiency of nitrate load predictions.

1. Introduction

The diffusion of nitrate pollution in watersheds is due to complex biochemical and hydrological procedures linked to the cycle of nitrogen and water. Nitrate load comes from different sources such as wastewater treatment plants, runoff of fertilized lawns and cropland, failing on-site septic systems, runoff of animal manure storage areas, and industrial discharges that contain corrosion inhibitors. Loss of nitrate to surface and groundwater can reduce farm productivity, harm the environment, and affect drinking water quality. Large uncertainties involved in the water quality processes (such as nitrate) limit physical interpretations and consequently make black box modeling of such processes more attractive over physical-based (white box) investigations. Therefore, several black box approaches have been developed and used to simulate water quality processes (e.g. see, Najah et al., 2012; Ma et al., 2014; Ay and Kisi, 2014; Sarkar and Pandey, 2015; Ravansalar et al., 2015; Olyaie et al., 2016). In this regard, Artificial Neural Network (ANN) and Support Vector Machine (SVM) as black box approaches have revealed their capabilities in modeling and forecasting non-linear water quality processes in general and nitrate time series modeling in particular (e.g., Markus et al., 2003; Anctil et al., 2009; Dixon, 2009; Hosseini and Mahjouri, 2014). Noori et al. (2015) introduced an appropriate methodology for determination of SVM's uncertainty for biochemical oxygen demand prediction. Arabgol et al. (2016) developed SVM to predict nitrate concentration of Arak plain, Iran; overall, the results showed that SVM model could be used as a fast, reliable, and cost-effective method for assessment and predicting groundwater quality. In spite of ability of artificial intelligence (AI) models sometimes there is a shortage when signal fluctuations are highly non-stationary and physical hydro-environmental process

* Corresponding author at: University of Tabriz, 29 Bahman Ave., Tabriz 5166616471, Iran.
E-mail address: vnourani@yahoo.com (V. Nourani).

http://dx.doi.org/10.1016/j.jhydrol.2017.03.002
0022-1694/

operates under a large range of scales varying from one day to several decades. In such a situation, AI models may not be able to cope with non-stationary data if pre-processing of the input data is not performed. The Wavelet Transform (WT) is an appropriate temporal pre-processing method that can be utilized to extract a variety of features from the data, such as short-term and long-term fluctuations, by decomposing the time series into different sub-series. The wavelet decomposition of a non-stationary time series into various scales provides an interpretation of the time series structure and extracts significant information about its history. As a result of these features, the WT can clarify spectral and temporal information of time series by combination with ANN and SVM as hybrid Wavelet-ANN (WANN) and Wavelet-SVM (WSVM) models for prediction of hydro-environmental processes (Nourani et al., 2014; Nourani and Andalib, 2015). In this paper, WANN and Wavelet Least Square Support Vector Machine (WLSSVM) models are employed for Multi-Station (MS) modeling of nitrate load. However in any data driven modeling, some of the inputs may have no significant relationship with output variables. Therefore, determination of dominant input variables, which are independent and informative is one of the important challenges of time series forecasting (e.g., Bowden et al., 2005; May et al., 2008; Tran et al., 2015) particularly in the WANN and WSVM development procedure. Therefore, to extract main features and inputs of the WANN and WLSSVM methods, two kinds of data pre-processing methods of Self-Organizing Map (SOM) based clustering and Mutual Information (MI) concepts are employed in this study. SOM as a robust clustering tool is a sort of unsupervised ANN system, and recently has widespread in several hydro-environmental fields (e.g. see, Kalteh et al., 2008; Ismail et al., 2011; Nourani and Parhizkar, 2013; Kar et al., 2015; Li et al., 2016; Zhou et al., 2016; Chang et al., 2016), also in this study clustering results of SOM was compared with linear K-means tool (Hsu and Li, 2010). On the other hand, MI as a nonlinear measure of data substance and information content can be a valuable tool in obviating the choice of viable inputs among tremendous numbers of wavelet-based sub-series (Nourani et al., 2015). Altogether, selection of dominant input data for the model has a crucial role to obtain accurate spatial variations of the nitrate loads. Therefore, spatio-temporal investigation, identification and using all sub-basins records as a MS analysis can improve prediction of nitrate diffusion in watershed. Although MS models are employed in some fields of hydrology (e.g. see, Turan and Yurdusev, 2009; Nourani and Komasi, 2013), to the best of the authors' knowledge, although there is a lack of literature focusing on AI-based prediction of streamflow nitrate time series; previous studies presented Markovian-based modeling of some other quality parameters of streamflow such as chemical oxygen demand, total organic carbon, electrical conductivity and ammoniacal nitrogen (e.g., Burchard-Levine et al., 2014; Deng et al., 2015). However, owing to the seasonality (periodic characteristics) of the hydrologic cycle and its impact on the nitrate transformation through the watershed, it is necessary to consider multi-scale seasonality of the process in the modeling framework. Furthermore, there is no research employing a MS framework in order to pattern the nitrate load (water quality) of a watershed and other researches just confined to consider outlet of watersheds and didn't pay attention to spatial variation of nitrate load. According to this, the effort of this paper is not only MS modeling of outlet nitrate of the Little River watershed (LRW) which nitrate process is one of the dynamical environmental stresses of this watershed but also notifying the nitrate yield from the inside of the watershed for getting informed from variation of nitrate loads of interior sub-basins. Hence, MS nitrate modeling is considered whereby nitrate loads of the inside of the LRW could be predicted. As a more explanation, the nitrate of upper sub-basins are employed for predictions of the interior sub-basins nitrate loads, and then, central sub-basins are participated in outlet nitrate prediction of the LRW. So, MS model can prepare a reliable platform to get information about the amount of nitrate in crucial places of the LRW. For this purpose, two scenarios with distinct views are used for MS nitrate modeling to identify the suitable strategy for future hydro-environmental researches. In the first scenario, Markovian characteristics of the streamflow-nitrate process are proposed as the base of the MS model, where antecedent of streamflow and nitrate time series of sub-basins are shared in nitrate modeling. On the other hand, non-linear feature extraction criterion of MI that is more suitable measure comparing to the linear measure of Correlation Coefficient (CC) is employed for the selection of appropriate inputs of the LSSVM and Feed Forward Neural Network (FFNN) models to avoid from the time consuming trial-error process of input selection, also performance of FFNN and LSSVM models was compared to the conventional linear prediction method of ARIMAX (Auto Regressive Integrated Moving Average with exogenous input) in the first scenario. In the second scenario, seasonality-based characteristics of the streamflow-nitrate process are focused. Where, streamflow and nitrate time series of the sub-basins are decomposed by the WTs at a suitable level for clarifying spectral and temporal information of the time series. Then, as a new feature extraction method, both SOM and MI are respectively employed for clustering homogeneous sub-series and selecting clusters' proper agents, to be fed into LSSVM and FFNN models for MS nitrate load modeling of the LRW.

2. Materials and methods

2.1. Study area and data set

The southeast United States is an important region from economical, agricultural, and social points of view, but its rapidly growing population adds water demand and environmental stress to its degraded ecosystem (Bosch et al., 2006). One of the dynamical environmental stresses is nitrate which its limitation is established by the US environmental protection agency as Maximum Contaminant Level (MCL) of 1 mg L^{-1} for nitrite (measured as nitrogen) as well as 10 mg L^{-1} MCL for total nitrate plus nitrite (measured as nitrogen). The production of nitrate (NO_3^-) can be described via two reactions, the first reaction is conducted by the Nitrosomonas genus as it oxidizes ammonia (NH_4^-) to nitrite (NO_2^-, Eq. (1)) and the second reaction is conducted by the Nitrobacter genus as it oxidizes nitrite into nitrate (NO_3^-, Eq. (2)).

$$NH_4 + 1.5O_2 \rightarrow 2H^+ + H_2O + NO_2^- \quad (1)$$

$$NO_2^- + 0.5O_2 \rightarrow NO_3^- \quad (2)$$

The nitrification process can be performed in the soil by nitrifying bacteria using ammonium came from ammonification of plants, fertilizer, waste and animal dung. Nitrate is more water-soluble and can be diffused in the watershed by streamflow. Therefore, the land cover/use can have direct relation with nitrate load production of sub-basin. The LRW in Tifton, southeast Georgia, southeast US as a kind of stereotyped coastal plain watershed was surveyed for the present study goal. Indwelling a wide zone in the southeast US, coastal plain watersheds have typified via platy, flat sedimentary flood plains with sandy soils, and slow moving flows. The LRW covers 334 km^2, of which approximately 40% of the southern half and 30% of the northern half are agricultural croplands (Bosch et al., 2004) and it is comprised of eight sub-basins of size varying from 2.62 to 334 km^2 (Fig. 1a). Via a filed investigation, the land cover of the LRW was classified into five categories of tilled fields, open water, pasture/lawn, total forest and urban (Bosch et al., 2006, Fig. 2). The observed nitrate data show