A Study of Real-Time Forecasting for the Urban Lake-Groundwater Coupled System Using Surrogate Models

Chuankun Liu^{1,2,*}, Yue Hu^{2,3}, Ting Yu¹, Qiang Xu^{2,3}, Chaoqing Liu¹, Quan An¹ and Chao Shen¹

¹POWERCHINA Chengdu Engineering Corporation Limited, 610072 Chengdu, China

²College of Environment and Civil Engineering, Chengdu University of Technology, 610059 Chengdu, China

³State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, 610059 Chengdu, China

Abstract. The real-time forecasting of flooding event and pollution emergency has a significant impact on the robust of urban lake and groundwater coupled system. However, the traditional statistical based prediction method is too rough while numerical based method is very time-consuming. In this study, a framework integrating surface water-groundwater coupled numerical model and surrogate model for real-time forecasting was proposed. The Artificial Neural Network (ANN) algorithm was used to train the surrogate model. The performance of the surrogate model was assessed with the number of training samples and hidden neurons as variates. More training samples would help improve the performance of the surrogate model indicated with R square value getting close to 1. The complex ANN with more hidden neurons performed better than the simple networks in the condition of enough training samples, and complex network without enough supporting training samples would be inferior to the simple network.

1 Introduction

The lake in urban areas, including the artificial lake, plays an important role in the civic environment maintenance[1]. The urban lake is benefit for providing recreation area, storm water management and polluted surface water purification[2]. Meanwhile, the urban lake is a fragile ecosystem since the dense contamination threatens and poor eco-plasticity compared with lake in the natural area. The urban lake is tightly connected with groundwater[3]. Once the urban lake is heavily polluted through sudden events, the irreversible damage would happen to the groundwater, and vice versa[4]. Hence, managers needs to figure out the response of urban lake under various uncertain threatens in time with more frequent extreme weather happening, and the real-time forecasting for water level, flow speed and water quality in the urban lake is very necessary[5]. Traditional statistical based forecasting method cannot match high accuracy requirement during the management of urban environment, and the numerical based forecasting method has unacceptable time consumption.

Building surrogate model can provide accurate prediction with high computational efficiency[6]. The training method of surrogate modelling is a key challenge of using the surrogate model in real time forecasting for the urban lake-groundwater coupled system. The Artificial Neural Network (ANN)[7] is one of the most popular and powerful machine learning algorithms applied in hydrological[8, 9] and environmental issues[10, 11]. The neural network in

ANN method is developed by imitating the communication systems of human neurons, which can learn behavior logic from the training data by adjusting the connection weights between different layers of neurons in the network. The ANN algorithm would be used to train the surrogate model of coupled surface water (lake) and groundwater numerical models, providing a tool of real time forecasting for the urban lake-groundwater coupled system in this study.

2 Methods

2.1 Numerical Model

2.1.1 MT3D-USGS

The MT3D-USGS is a three-dimensional multi-species solute transport model from United States Geological Survey, which can be used for solving advection, dispersion, and chemical reactions of contaminants in surface water, unsaturated and saturated groundwater systems[11]. In MT3DM-USGS, the solute transport simulation module interfaces directly with the USGS finite-difference flow model MODFLOW for the flow solution.

MT3D-USGS was developed from MT3DMS, by adding functions like unsaturated-zone transport simulation and lake-groundwater interaction simulation, and MT3DMS is not a USGS software product (the support information of MT3DMS is available at

* Corresponding author: liuchuankun0615@126.com

<http://hydro.geo.ua.edu/mt3d/>). The first generation of this series software is named as MT3D developed by Zheng in 1992[12], MT3D can only be used for single-species mass transport modeling. On the basis of MT3D, MT3DMS was developed for multi-species solute transport modeling and has been widely used in research projects and practical engineering applications[13].

2.1.2 Lake Package of MT3D-USGS

In MT3D-USGS, the simulation of solute transfer between lake features and the underlying aquifer was accomplished by Lake Transport (LKT) package. LKT assumes that contaminant (solute) entering the lake would be instantaneously fully mixed throughout the entire volume in the lake water, and the constituent concentration in lake seepage into the underlying aquifer is equal to the concentration in the lake. The instantaneous mixing assumption means flow dynamics within lakes and the effects of spatial stratification are not considered in LKT of MT3D-USGS.

In this study, a simulation case was chosen as the original numerical model for generating training datasets for the surrogate model, which is used to verify the accuracy of the LKT package. In the case, the transient simulation period is set to be 5000 days which is divided into 100 time steps. The spatial grid discretization of this model is shown in Fig. 1, where the lake located in the middle part of the simulated domain. The initial constituent concentration in the lake is set to be 100 mg/L, and 0 mg/L in the groundwater. Precipitation is set to be a constant value of 0.0115 ft/d and evaporation rate is 0.0103 ft/d. The water head in the west boundary is set to be a constant value of 160 ft, and the east boundary is 140 ft. The north and south boundary are set to be a linear decrease from 160 ft to 140 ft in west to east direction. The simulation results of this numerical case are compared with an analytical solution, ensuring the accuracy of LKT package.

Fig. 1. (a) Plan view of the simulated domain; (b) Profile view of the simulated domain.

2.1.3 Generation of the Training Samples

The surrogate model needs samples generated from the original numerical model for the training process. In this study, a MATLAB (Mathworks, Natick, MA, USA) code is used to generate the input data set of MT3D-USGS, then invoke MT3D-USGS program and run it automatically.

Five input variables are chosen as the training input parameters, including porosity, hydraulic conductivity along rows, lakebed hydraulic conductivity, the vertical hydraulic conductivity and storage coefficient. Latin hypercube sampling method was used to cover the entire sampling space. The solute concentration at any spatial point in the simulated domain can be determined as the output information, and the lake solute concentration was chosen as the training output. The number of the training samples needed by the surrogate model maybe large, therefore the input and output information for surrogate model training is stored dynamically for saving storage space requirement.

2.2 Introduction to ANN

The ANN algorithm is used to train the surrogate model in this study[7]. An ANN contains an input layer, a hidden layer and an output layer as shown in Fig. 2. In the input layer, neuron elements, which has the same dimension as the input data, collect input information from the input data. The input information is transferred to the hidden layer in accordance with a predetermined non-linear function (activation function). The neuron number in the hidden layer can be adjusted by the developer, more neurons means more complex network, which is more powerful but needs more training samples. Specifically, the Back Propagation (BP) Neural Network is used in the surrogate model training. The learning process adjusts the contact mechanism between neurons in different layers based on prepared input and output training samples. The number of neurons in the hidden layer and training algorithm are the key parameters which should be carefully optimized or chosen.

Fig. 2. The sketch map of the structure of ANN.

2.3 Training Surrogate Model

The training process of the surrogate model is conducted in MATLAB environment. In the training process, the training samples would be divided into training group, validation group and testing group. The number of samples for different use can be adjusted artificially. Different number of hidden neurons would be tested to analysis the space complexities' influence on the performance of the network. Levenberg-Marquardt training algorithm (trainlm) has high time efficiency and is able to stop automatically when generalization stops improving, thus trainlm is used to train the surrogate model in this case.

The training process of ANN algorithm could be influenced by randomness, the same training strategy would run 100 times and the R square value of each strategy, which is the indicator of the training performance, would be calculated from the 100 results. Support Vector Machine (SVM) is also used to train the surrogate model in this study for the purpose of evaluating the performance of the ANN algorithm.

3 Results and Discussion

A series of numerical experiments were set up to test the performance of ANN with a different number of training samples and hidden neurons. The number of hidden neurons varied from 7 to 16, and the network (surrogate model) was trained by 50 to 200 samples. 1000 samples were used to test the fitness between the predicted values by the trained surrogate model and the real values. Each scenario was run 100 times to eliminate the noise influence from the randomness of the ANN training process. The performance of the ANN algorithm was shown in Fig. 3. The result shows that for this kind of low dimension surrogate models, the ANN algorithm can capture the mapping relationship between input data and output data well and truly. With the number of training samples increases from 50 to 200, the performances of networks with different hidden neurons improve notably, characterized by the R square value getting more close to 1. R square is an evaluation indicator of fitness between predicted and real values, and R square value increases from 0 to 1 indicating the improvement of fitness.

Fig. 3. The surrogate model performance of ANN with the number of training samples and hidden neurons varying.

The neuron number in the hidden layer has more complex influences on the performance of the ANN algorithm compared with the number of training samples.

Fig. 4 is a plan view map showing the R square value with different numbers of training samples and hidden neurons. The algorithm performs better with a larger number of training samples, while the result shows that the optimal neuron number in the hidden layer varies with the number of training samples changing. According to this knowledge, the structure of the surrogate model should be in harmony with the available training samples in the application of real-time forecasting. A complex network needs more supporting training samples, and has worse performance compared with a simple network when the number of training samples is less than 100 in this study.

Fig. 4. The plan view showing the R square value with a different number of training samples and hidden neurons.

The potential of a complex network with more hidden neurons can be activated when the training samples are enough to support the training process. Fig. 5 shows that the performance of the trained network improves notably with enough training samples, i.e. from 400 to 600 training samples in this study. A complex network has a higher risk of overfitting, setting penalty coefficient and enough cross validation samples can reduce the risk of overfitting.

Additionally, the ANN algorithm has a linear time and space complexity ($O(n)$), where n represents the number of training samples. The linear time and space complexity of the ANN algorithm should be taken into account in the actual application process.

Fig. 5. The surrogate model performance of ANN improves with an increasing number of hidden neurons when training samples is enough.

The SVM algorithm is another machine learning training method which is also widely used. The result shows that the SVM algorithm needs much more training samples to improve its performance compared with the ANN algorithm in this specific case.

4 Conclusion

A new surrogate based approach was developed for the real-time forecasting of urban lake and groundwater coupled system. Surface water and groundwater coupled numerical model MT3D-USGS was used to simulate the urban lake and groundwater interaction process and generate training samples for the surrogate model. A batch program was developed to generate the input file of MT3D-USGS, invoke the core program and save the output information dynamically. ANN was used to train the surrogate model, and the performance of the surrogate model was analyzed. The surrogate model can greatly decrease computational costs compared with the original numerical model, meeting the strict computational efficiency of real-time forecasting.

The proposed method offers a reliable guideline for the real-time forecasting of urban lakes and groundwater coupled system. More training samples helps to improve the performance of the surrogate model, but the number of hidden neurons has a more complex influence pattern. With limited training samples, simple network should be preferred since the complex network may not be well trained. Without doubt, more hidden neurons do help improve the performance with enough training samples. Traditional approaches to forecast the urban lake and groundwater coupled system are mainly based on the statistical method. The statistical method is rough while numerical based approaches have a high computational cost, which are both unacceptable in real-time forecasting. Compared with traditional measures, this mathematically based approach proposed in this study performs more scientifically and matches the tight schedule of real-time forecasting at the same time.

The approach proposed in this study faces challenge when dealing with high dimensional surrogate modeling problems. In the real-time forecasting problem, better representativeness needs more temporal and spatial

control points. On this background, the high dimensional surrogate models should be trained and used. Deep learning is a kind of ANN with more hidden layers, which is a potential solution for the complicated surrogate modeling problems.

Foundation items

- (1). The National Natural Science Foundation of China (NSFC, Grant No.: 41807407);
- (2). P38018 Research Project of POWERCHINA Chengdu Engineering Corporation Limited;
- (3). P32916 Research Project of POWERCHINA Chengdu Engineering Corporation Limited;
- (4). Major Project for Specialized Science and Technology Fund of Sichuan Province (Grant No.: 19ZDZX0033).

References

1. R. B. E. Shutes, Artificial wetlands and water quality improvement, *Environ. Int.*, **26**, 5, 441-447 (2001)
2. C. Makropoulos and D. Butler, A multi-objective evolutionary programming approach to the 'object location'spatial analysis and optimisation problem within the urban water management domain, *Civ. Eng. Environ. Syst.*, **22**, 2, 85-101 (2005)
3. C. H. Mortimer, Chemical exchanges between sediments and water in the great lakes-speculations on probable regulatory mechanisms, *Limnol. Oceanogr.*, **16**, 2, 387-404 (1971)
4. H.-B. Moon, M. Choi, J. Yu, R.-H. Jung, and H.-G. Choi, Contamination and potential sources of polybrominated diphenyl ethers (PBDEs) in water and sediment from the artificial Lake Shihwa, Korea, *Chemosphere*, **88**, 7, 837-843 (2012)
5. J. Su, X. Wang, S. Zhao, B. Chen, C. Li, and Z. Yang, A structurally simplified hybrid model of genetic algorithm and support vector machine for prediction of chlorophyll a in reservoirs, *Water*, **7**, 4, 1610-1627 (2015)
6. S. Razavi, B. A. Tolson, and D. H. Burn, Review of surrogate modeling in water resources, *Water. Resour. Res.*, **48**, 7 (2012)
7. R. S. Govindaraju and A. R. Rao, *Artificial neural networks in hydrology*. (Springer Science & Business Media, 2013)
8. S. Mohanty, M. K. Jha, S. Raul, R. Panda, and K. Sudheer, Using artificial neural network approach for simultaneous forecasting of weekly groundwater levels at multiple sites, *Water. Resour. Manag.*, **29**, 15, 5521-5532 (2015)
9. M. Behzad, K. Asghari, M. Eazi, and M. Palhang, Generalization performance of support vector machines and neural networks in runoff modeling, *Expert. Syst. Appl.*, **36**, 4, 7624-7629 (2009)
10. K. Ostad-Ali-Askari, M. Shayannejad, and H. Ghorbanizadeh-Kharazi, Artificial neural network

for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran, Ksce. J. Civ. Eng., **21**, 1, 134-140 (2017)

11. V. Bedekar, E. D. Morway, C. D. Langevin, and M. J. Tonkin, *MT3D-USGS version 1: A US Geological Survey release of MT3DMS updated with new and expanded transport capabilities for use with MODFLOW*, (US Geological Survey, 2016)
12. C. Zheng, *MT3D: A modular three-dimensional transport model for simulation of advection, dispersion and chemical reactions of contaminants in groundwater systems*, (SS Papadopoulos & Associates, 1992)
13. C. Zheng and P. P. Wang, *MT3DMS: a modular three-dimensional multispecies transport model for simulation of advection, dispersion, and chemical reactions of contaminants in groundwater systems; documentation and user's guide*, (Alabama Univ University, 1999)