Survey of Methods for Data-Scarce Processing Based on Mechanism Model

Guobo Wang, Minglu Ma and Liansheng Xu

China Waterborne Transport Research Institute, No.8 xitucheng road, Haidain District, Beijing 100088, PR China.
Email: wangguobo@yeah.net

Abstract. Due to the imbalance in global economic development, the lack of response data is a prominent shortcoming in information concerning many river basins. The uncertainties incurred by data scarcity reduce the reliability of the simulation results of river basins and pose difficulties in the simulation of non-point source pollution and maritime search and rescue, which in turn affects the environment of the basin. Based on a simulation conducted using the mechanism model in the context of a lack of response data from rivers, this paper analyses a variety of methods of data processing in case of missing data. In the era of big data, this paper provides a variety means of using data-scarce simulation methods as an effective reference for simulations of water quality.

1. Introduction
China is facing a severe water shortage among many countries in the world. Its water resources per capita are less than that of the worldwide average. Rapid economic development has driven the development and utilization of water resources. At the same time, water shortages and contamination caused by a lack of protective policies have gradually become serious. China at present faces the dual pressure of water shortage and water pollution. With regard to the latter, the first concern has been point source pollution. As this has been controlled gradually, non-point source pollution has emerged as a concern. Non-point source pollution not only brings environmental problems, but also brings great difficulties to the search and rescue on the water. In the process of maritime search and rescue, the alluvial deposits of runoff and pollutants have a great impact on the water environment. However, the effective simulation of runoff and pollutant emissions can help search and rescue personnel to grasp the underwater dynamics in time, which is of great help to improve the efficiency of search and rescue and ensure the safety of search and rescue. Currently, model simulations are the major means to assess non-point source pollution. Insufficient monitoring and natural conditions lead to a lack of model-driven data and response data, which significantly affects the results of simulations.

Therefore, it is of great practical significance to carry out simulation on non-point source pollution in data-scarce watersheds, for the prevention and control of non-point source pollution in data-scarce watersheds, and to improve the ability of maritime search and rescue.

2. Status of Scarcity of Response Data
The load estimation for non-point source pollution mainly relies on two methods: the empirical statistical model and the mechanism model. Based on the characteristics of basins, the non-point source pollution model is established by building a statistical relationship among rainfall, hydrology, and water quality. However, the statistical model requires a large amount of data. Currently available statistical data cannot accurately reflect the complexity of the hydrological and environmental
processes in the basin. In particular, a statistical model to simulate non-point source pollution for a given event of rainfall cannot be constructed owing to the short monitoring time series and inconsistent frequencies of monitoring. The mechanism model can adequately describe the continuous process of non-point source pollution. By constructing numerical relationships among elements in the basin, such indicators as river flow and pollutant concentration can be obtained. Therefore, most current non-point source simulations are based on the mechanism model. The data for this model is mainly composed of driving data and response data. Models based on driving data are mostly based on meteorological data, such as rainfall, and underlying surface data, such as elevation, land use, and soil. The response data are mainly used for the calibration and verification of the model, and generally include hydrological and water quality-related data. However, most such data are obtained through manual monitoring, and the measurement are affected by human error. Manual monitoring is also costly and thus rare. The volume of data for many river basins is too small to meet the requirements of the general process of verifying the calibration of the model.

Influenced by the imbalance in regional development and topographical factors, some basins have only a single or few hydrological and water quality monitoring stations, which leads to significant differences in the distributions of the data collected. For river basins with only a single such site, the cost of monitoring different types of response data varies widely. The cost of monitoring water quality, such as total nitrogen and phosphorus content, is high while the cost of monitoring hydrological data, like flow, is relatively low. Therefore, most water quality-related data are monitored monthly while hydrological data are mostly monitored daily. Due to the influence of monitoring frequency and data processing, the hydrological and water quality-related monitoring data from stations in basins do not match, and are partially or completely missing in some cases, which renders the calibration and verification of the model challenging. These factors have made the scarcity of response data a common problem for simulations of non-point source pollution in river basins.

3. Processing Missing Response Data
Simulations of hydrology and water quality based on time series using response data from a single site have employed a variety of methods, including data assimilation, data fusion, and statistical models. Because data fusion technologies mostly belong to the category of data assimilation methods, this article presents it only from the perspective of the application of data assimilation and statistical models.

Data assimilation is a data processing technology originally derived from numerical weather forecasting to provide an initial field for it. This method has now been widely used in assessing the atmosphere and water quality. Because data assimilation can be applied to multiple fields of Earth system research, experts in different fields have different expressions for the intention and extension of data assimilation. Data assimilation can be said to be composed of four basic elements: dynamic models that simulate natural processes, direct or indirect observation data of state quantities, a data assimilation algorithm that continuously incorporates new data into process model calculations, corrects the model parameters, and improves simulation accuracy, and basic parameter data driving the model. The main task of data assimilation is to merge observation data from various sources, different error-related information, and different spatiotemporal resolutions into a dynamic numerical model. According to mathematical theory, an optimal solution between the model solution and observations can thus be found. This solution can continuously provide the initial field for the dynamic mode, which circulates and drives the results of the mode constantly approximating to the observed values. Common methods of data assimilation include optimal interpolation, variation assimilation, and Kalman filtering. Starting from the demands for surface water source monitoring and evaluation, and combining it with the characteristics of satellite remote sensing monitoring, Yu et al. (2014) studied and established an assimilation model for ground water level monitoring data and water area monitoring data by satellite remote sensing. Based on the assimilated information, the curve of characteristics of the reservoir was modified, and a reservoir storage monitoring model based on the assimilation of spatial and ground-related data was established to improve the accuracy of reservoir storage monitoring. Li et al. (2014) developed a multi-model collaborative retrieval algorithm to retrieve the concentration of chlorophyll A in Taihu Lake based on methods of data assimilation.
Seven modes of concentration retrieval of chlorophyll A were constructed using water reflectance data from hyperspectral remote sensing spectra in Taihu Lake from 2006 to 2009. It was shown that the accuracy of the multi-model collaborative retrieval algorithm was higher than that of the single-model retrieval method. The accuracy and errors of the products of retrieval were effective evaluated by calculating the confidence interval of the multi-model collaborative retrieval product. Model datasets based on data assimilation are also available in China, such as the CMADS dataset (China Meteorological Assimilation Driving Datasets for the SWAT model), and can be used for free. It was established by using the China Land Data Assimilation System (CLDAS) technology and using multiple techniques, such as data loop nesting, resampling, model estimation, and bilinear interpolation. The multi-grid method for three-dimensional variational assimilation (STMAS) can be used for a basic elemental field analysis based on the NCEP/GFS background field. In areas outside China, only NCEP/GFS background data are processed through terrain adjustment, variable diagnosis, and interpolation to the analyzed grid points. Using the STMAS algorithm, the pre-processed NCEP/GFS background data are fused with automatic station observations within China and stitched with data from outside China. Meng et al. (2016) introduced the Chinese Land Data Assimilation System (CLDAS) and established a CMADS dataset. In a case study on Heihe River Basin, the dataset using CMADS V1.0 was utilized to drive the SWAT model. The results of SWAT models driven by CFSR and traditional weather stations were compared. The analysis showed that the CMADS dataset was more advantageous than traditional weather stations in driving large-and meso-scale hydrological models. Although assimilation processing is currently performed only on SWAT model-driven data, work on response data is also underway.

The empirical statistical model is widely applied to hydrological simulations and those of water quality in areas lacking such data because its empirical parameters are geographically representative, and are not affected by the time scale of the simulation. Commonly used methods include the pollution load division method, output coefficient method, correlation relationship method, mean concentration method and the LOADEST model. There are also large differences in the applicable conditions of these methods. The pollution load division method, output coefficient method, and correlation relationship methods require a large amount of data for pre-analysis, and mostly use annual data. The LOADEST model uses continuous daily flow data and limited, discrete water quality-related data to establish load regression equations for water quality, and estimates the transport flux of rivers at different time scales. It is widely used in to predict water quality in areas worldwide, especially in basins lacking remote sensing data. Zhang et al. (2014) selected a typical basin in Zhejiang province as sample to estimate the annual NO-3-N flux and net anthropogenic nitrogen inputs (NANI) of rivers based on water quality and nitrogen source data from 1980 to 2010, and the LOADEST model. The results showed that the interannual change in riverine NO-3-N flux was not only significantly correlated to NAIN and the fertilizer nitrogen input, but was also strongly related to the river's annual flow or rainfall. A regression model developed by incorporating both NANI and water flow as independent variables can adequately simulate the NO-3-N flux change in rivers. Chen et al. (2015) used the LOADEST model to estimate the daily total nitrogen load in 18 basins in the southeastern United States with scarce data for them. The results showed that the LOADEST model can be used to simulate the total nitrogen load in most basins, and its performance in dry and normal seasons is relatively good.

4. Conclusion
The scarcity of monitoring data, especially the significant scarcity in the response data compared with the volume demanded by simulations of the mechanism model, poses challenges to simulations of non-point source pollution. Prevalent studies have carried out simulations of non-point source mechanism models in case of scarce data from various perspectives. Considerable progress has been made that can provide guidance for simulations of this kind. However, at present, regional environmental protections for water resources are not adequate to ensure that sufficient data resources are available. Environmental protection departments at all administrative levels are often constrained when carrying out measures to protect the regional water environment. Therefore, in future work, it is
important to combine multiple data assimilation methods and technologies to provide technical support for achieving the water environment simulation evaluation under data scarcity.

5. Acknowledgments
This project was supported by the National Science and technology support program(2015BAG20B01),the National Key R&D Program of China(2017YFC1404705).

6. References
[1] Xiang C, Wang Y, and Liu H 2017 A scientometrics review on nonpoint source pollution research Ecological engineering 99 400-408.
[2] Jabbar F K and Grote K 2019 Statistical assessment of nonpoint source pollution in agricultural watersheds in the Lower Grand River watershed, MO, USA Environmental Science and Pollution Research, 26(2) 1487-1506.
[3] Zhang L, Lu W, Hou G, Gao H, Liu H and Zheng Y 2019 Coupled analysis on land use, landscape pattern and nonpoint source pollution loads in Shitoukoumen Reservoir watershed, China Sustainable Cities and Society 51 101788.
[4] Ouyang W, Yang W, Tysklind M, Xu Y, Lin C, Gao X and Hao Z 2018 Using river sediments to analyze the driving force difference for non-point source pollution dynamics between two scales of watersheds Water Research 139 311-320.
[5] Read E K, Carr L, De Ciecco L, Dugan H A, Hanson P C, Hart J A and Winslow L A 2017 Water quality data for national-scale aquatic research: The Water Quality Portal Water Resources Research 53(2) 1735-1745.
[6] Emili L A and Greene R P 2013 Modeling agricultural nonpoint source pollution using a geographic information system approach Environmental Management 51(1) 70-95.
[7] Cho J and Mostaghimi S 2009 Dynamic agricultural non-point source assessment tool (dansat): model development Biosystems Engineering 102(4) 486-499.
[8] Chahor Y, Casali J, Giménez R, Bingner R L, Campo M A and Goñi M 2014. Evaluation of the annagpns model for predicting runoff and sediment yield in a small mediterranean agricultural watershed in navarre (spain) Agricultural Water Management 134, 24-37.
[9] Diaz-Ramirez J N, McAnally W H and Martin J L 2011 Analysis of hydrological processes applying the HSPF model in selected watersheds in Alabama, Mississippi, and Puerto Rico Applied Engineering in Agriculture 27(6) 937-954.
[10] Chen L, Shen Z Y, Yang X H, Liao Q and Yu S L 2014 An interval-deviation approach for hydrology and water quality model evaluation within an uncertainty framework Journal of Hydrology 509 207-214.
[11] Lu Z X, Cai X H, Zhou S B, Long A H and Xu B R 2012 Application of SWAT model in the upstream of Ili River Basin with Scarce Data Arid Land Geography 35(3) 399-407.
[12] Li L, Dong X H, Yu D, Liu J and Zhou Q P 2013 Study on runoff simulations on Qingjiang River Basin by SWAT model Yangtze River 44(22) 25-29.
[13] Yu X 2014 Study of water body monitoring of surface water source region by remote-sensing technology Pearl River Water Transport (16) 72-73.
[14] Li Y, Li Y M, Lv H, Zhu L, Wu C Q, Du C G and Wang S 2014 Muti-model collaborative retrieval of chlorophyll a in Taihu Lake based on data assimilation Environmental Science 9 022.
[15] Meng X Y, Shi C X, Liu S Y, Wang H, Lei X H and Liu Z J 2016 CMADS datasets and its application in watershed hydrological simulation: A case study of the Heihe River basin Pearl River 37(07) 1-19.
[16] Zhang B F and Chen D J 2014 Dynamic response of riverine nitrate flux to net anthropogenic nitrogen inputs in a typical river in Zhejiang province over the 1980-2010 period Environmental Science 35(8) 2911-2919.