A new forecasting model for groundwater quality based on short time series monitoring data

To cite this article: Ling Yao and Yunqiang Zhu 2019 IOP Conf. Ser.: Earth Environ. Sci. 227 062014

View the article online for updates and enhancements.
A new forecasting model for groundwater quality based on short time series monitoring data

Ling Yao\textsuperscript{1,2,3} and Yunqiang Zhu\textsuperscript{1,2,4}

\textsuperscript{1}Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; \textsuperscript{2}Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing Normal University, Nanjing 210023, China; \textsuperscript{3}Guangzhou Institute of Geography, Guangdong Academy of Sciences, Guangzhou 510070, China.

\textsuperscript{4}Email: zhuyq@lreis.ac.cn.

Abstract. Groundwater is an important part of regional water resource, rapid urban development often witness deterioration of regional groundwater quality. This paper proposed a missing-aware-weighted hidden markov model (MWMO-HMM) combining expectation maximization algorithm (EM) with a weighted multi-order HMM to build groundwater quality prediction model with incomplete short-term observations. The proposed model was used to predict hydrogen ion concentration (PH) and chemical oxygen demand (COD) of groundwater in five representative cities. The Nash–Sutcliffe model efficiency coefficients of MWMO-HMM prediction results are respectively 61.51\% and 98.06\%. Compared with prediction results achieved by auto-regressive and moving average model (ARMA) and gray model (GM), the results show that MWMO-HMM is superior to ARMA and GM, ARMA and GM demonstrate an unstable performance of forecasting. In addition, missing value has a greater effect on ARMA than GM. Furthermore, the integral observations filled with EM algorithm indicates that COD concentration of karst groundwater in Guizhou is affected to some extent by the surface precipitation. The proposed model can predict groundwater quality effectively and meet the management requirements in groundwater prediction based on disintegrated small sample datasets. It would assist decision makers to enhance the decision making for future sustainable development.

1. Introduction

Groundwater is an important part of human living environment, providing freshwater resources [1]. Nowadays, groundwater pollution extends from point to surface extension because of human activities, and the contaminated groundwater will seriously affect the living environment of mankind in turn. Consequently, critical attention has recently focused on ground water resources, especially during dry periods [2].

Guizhou province suffers severe shortage of water resources due to its particular Karst landform, with a grievous desertification greatly that curbs the economic and social development. However, a large number of meteorological and geological observations indicate that groundwater resources are plentiful in Guizhou, the key problem is how to utilize the huge reservoir of groundwater. In recent years, urbanization and the development of industry and mining in Guizhou province has caused
varying degrees of contamination to karst groundwater. Since groundwater contamination is difficult to repair, accurate prediction of groundwater quality is of great realistic significance.

Classical process-based modeling approaches can provide satisfied estimations of water quality variables, but usually require input data with long sequence and model parameters as well as approximations of various processes [3]. Process-based models can provide satisfying simulation or prediction of each variable involved in the process, on condition that the initial conditions, boundary conditions, and the input-output relationships in groundwater movement are clearly identified. Unfortunately, groundwater is a very complex system with a lot of interior and exterior unknown factors, which reduces the reliability of process-based models.

There are also numerous data-driven techniques providing effective alternatives to process-based modeling. Regression analysis, gray theory, time series, fuzzy reasoning, and machine learning (such as neural network, deep learning) [4-9] are commonly used in groundwater quality forecasting and have yielded some positive results. These data-driven prediction techniques are computationally faster and require fewer input parameters than process-based models, the shortcoming of these data-driven models lies in the strong requirements of information of influential factors and a slew of training samples. Models built with machine learning methods always not only have high complexity and low commonality but also require data from multiple dimensions. Auto-Regressive and Moving Average Model (ARMA) and Gray Model (GM) are commonly used in water quality prediction [10]. ARMA establishes appropriate mathematical model to fit the historical trend and predicts the trend of the future time series according to the established model. GM uses a generated series data to build a grey differential equation model and then restore the original appearance for prediction.

Based on these previous applications, this paper demonstrated a missing-aware-weighted multi-order hidden markov model (MWMO-HMM) to improve the prediction accuracy of short-time series with missing values. The model was applied to predict the groundwater quality of five representative cities in China’s Guizhou Province. The results show that MWMO-HMM has better performance than traditional ARMA and GM methods, and could be used as a prediction tool, which complements the process-based model and ongoing field monitoring program in the region.

The rest of this paper is organized as follows: in Section 2, we give a brief description of the study area, water quality data sources, proposed method, and performance evaluation. Then in Section 3.1 we discuss and compare forecasting accuracy of the proposed method with traditional techniques. Section 3.2 discusses the temporal variations of PH and COD in the study area. Finally, we summarize our work and draw some conclusions.

2. Materials and methods

2.1. Study area

Although the average annual groundwater resources in Guizhou is about 47.939 billion cubic meter per year, and 80.5% of them are karst groundwater, Guizhou suffers severe rocky desertification, with light rocky desertification area has peaked at 35,900 km², accounting for about 20.39% of the study area [11]. Characterized by the water bearing features, water enrichment and utilization conditions of carbonate rocks, karst groundwater can be divided into three main subsets: 1) The Middle-Lower Triassic dolomite and limestone karst water, mostly lies in the middle and southwest of Guizhou province, and Guiyang, Zunyi, Anshun are located in this area; 2) the Middle-Upper Cambrian dolomite karst fissure water, mostly lies in the north, northeast, and southeast of Guizhou province, including Kaili (belongs to Qiandongnan Miao and Dong) and parts of Zunyi; 3) The upper Carboniferous and Lower Permian limestone karst water, clustered in the southern and northwest of Guizhou province, in which Liupanshui is located. The study area shown in Figure 1 is five representative cities of Guizhou province in the southwest of China, including Guiyang, Anshun, Kaili, Liupanshui, Zunyi.
Figure 1. Study area and water quality monitoring stations in Guiyang province of China.

2.2. Water quality datasets

PH and chemical oxygen demand (COD) data were collected from 2001 to 2014 at 10 monitoring sites in the above mentioned five cities (Figure 1), including the well, driven well, spring, specialized sampling hole, and surface water sampling point. Figure 2 demonstrates that both PH and AOD datasets exist a certain level of data missing. Anshun monitoring sites have the loss of PH data between 2001 and 2005, and the whole COD sampling information of the whole research period. Liupanshui also has a deletion of PH and COD observation during 2001-2007, COD concentration also has missing data of 2009 at Kaili monitoring stations.

Figure 2. Sampling data of groundwater quality at monitoring stations ranges from 2001 to 2014, existing many missing values even in such a short-term observation series.
2.3. Proposed prediction model (MWMO-HMM)

A missing-aware-weighted multi-order hidden Markov model (MWMO-HMM) is proposed to deal with short-term groundwater quality prediction considering missing values. Considering that the missing values in the short-term groundwater quality data series would affect the prediction of water quality, Expectation Maximization Algorithm (EM) is introduced to fill the missing values. Then, the weighted autoregressive Hidden Markov Model (HMM) is used to predict the time series by setting and continuously adjusting initial parameters. For conventional HMM, low-order HMM does not take full account of the historical information of the time series, while high-order HMM prediction faces the problem of low coverage. In this paper, prediction results of different orders HMM are combined to improve the prediction accuracy. The performance of our proposed model is also compared with other two traditional methods, ARMA and GM.

EM is typically used to compute maximum likelihood estimations given incomplete samples [12]. The EM algorithm estimates parameters of a iterative model. Starting from initial values, each iteration consists of an E step (Expectation step) and an M step (Maximization step). EM algorithm can be used to fill the missing data in samples, discovering the value of latent variables, estimating the parameters of HMMs, estimating parameters of finite mixtures, unsupervised learning of clusters. The strategy assumes that all data series follow Gaussian distribution. For time series with missing data (v'), we establish a complete time series at first, then use K-means to give the original parameters of Gaussian distribution and get first estimation of missing data v. This process is repeated for n' times until all complete series are taken into consideration. Finally, the missing data can be estimated as v' = (v_1 + v_2 + ⋯ + v_{n'})/ n'.

Weighted autoregressive HMM was used to forecast numerically the time series. The model includes six elements, such as order parameter (maximizing to K order), two state sets (hidden state S and measurable state O) and three probability matrices (initial state matrix π, hidden state transition matrix A and measurable transition matrix B). S can be described as S = [s_i, i = 1, 2, ..., N] where N represents the number of hidden states. The state at time t is s_t, which belongs to S. The states satisfy the Markov property and cannot be measured. The measurable state O equals to [o_i, i = 1, 2, ..., T] with T the length of observations. O is connected with the hidden state and can be observed, while M is the number of measurable states, which is not necessarily the same as N. The initial state matrix π is the probability matrix of hidden states at t=1. The matrix A gives the transition probabilities of all states of the HMM model. a_{ij} = P(q_{t+1} = s_j | q_t = s_i), i, j = 1, 2, ..., N is the probability of being s_j at time t+1 given s_i at time t. B is the probability of the measurable state to be o_i given that the hidden state is s_j at time t, in which b_{ij} = P(o_i | s_j), 1 ≤ i ≤ M, 1 ≤ j ≤ N.

Sliding window with one-step size was used to divide time series by frame. Time series O = {o_i, i = 1, 2, ..., T} can be extracted from auto regression (AR) coefficients of each frame φ_r (r = 1, 2, ..., R). The hidden state with the largest probability is the most possible AR model for representing the state, which can be defined as:

\[ x_t = \alpha + \sum_{i=t-1}^{t-p} \beta_if_i(x_i) + \varepsilon \]  

where, \( \alpha \), \( \beta_i (i=1,2,..,m) \) are the corresponding regression coefficients of the hidden state of HMM model. p is the order of AR model and \( \varepsilon \) is zero mean random disturbance term with variance of \( \sigma^2 \).

For given initial parameters \( \lambda_0 = (\pi, A, B) \), Baum-Welch algorithm [9] was used to re-estimate the parameters until the model reaches the convergence error or the maximum iteration step.
\[
\begin{align*}
\pi_t & = \gamma_t(i), \\
a_{ij} & = \sum_{t=1}^{T-1} \xi_t(i,j) / \sum_{t=1}^{T-1} \gamma_t(i), \\
b_j(k) & = \sum_{\xi(i) = \gamma_k} \gamma_t(i) / \sum_{\gamma(i) = \gamma_k} \gamma_t(i).
\end{align*}
\]  
(2)

where, \( \xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j|O, \lambda), \gamma_t(i) = P(q_t = S_i|O, \lambda) = \sum_{i=1}^{N} \xi_t(i,j). \)

The complete Markov process which includes a hidden layer and measurable layer can describe a stochastic process very well [13]. It is possible to forecast states of the hidden Markov chain based on weighted forecast method. In other hand, the state of system can be forecasted by using multi-steps related transition probabilities to get the weighted probability distribution. Autocorrelation coefficients of the steps reflect the relationships between time series with different steps, analyzing the correlation between transition probabilities and time series can make full use of information. The multi-dimension correlation coefficients \( r_k \) can be defined as:

\[
\begin{align*}
r_k & = \frac{\sum_{t=1}^{T-k}(o_t - \bar{o})(o_{t+k} - \bar{o})}{\sqrt{\sum_{t=1}^{T-k}(o_t - \bar{o})^2 \cdot \sum_{t=1}^{T-k}(o_{t+k} - \bar{o})^2}}
\end{align*}
\]  
(3)

State sequence can be reconstructed in the one-order model with Viterbi decoding algorithm, then is possible to obtain the hidden states corresponding to the measurable K values before the time \( t' \), which are taken as the initial states. By using the multi-step transition probability matrix, we can get the probability distribution of states at the forecast time:

\[
P = \begin{bmatrix}
p_{11}^{(1)}(t') & p_{12}^{(1)}(t') & \cdots & p_{1N}^{(1)}(t') \\
p_{21}^{(2)}(t') & p_{22}^{(2)}(t') & \cdots & p_{2N}^{(2)}(t') \\
\vdots & \vdots & \ddots & \vdots \\
p_{11}^{(K)}(t') & p_{21}^{(K)}(t') & \cdots & p_{N1}^{(K)}(t')
\end{bmatrix}
\]  
(4)

The weighted probability of the same state acts as the forecast probability at time \( t' \).

\[
P_i(t') = \sum_{k=1}^{K} w_k p_i^{(k)}(t'), \quad 1 \leq i \leq N
\]  
(5)

where, \( w_k \) is the standardized weight of autocorrelations: \( w_k = |\eta_k| / \sum_{k=1}^{K} |\eta_k|, 1 \leq k \leq K. \)

The framework of our proposed model can be described as Figure 3.

3. Results and discussion

3.1. Missing data estimation

Figure 4 shows the time series dataset filled with missing values using EM algorithm. The accumulated COD concentration reached a maximum of 4.7 mg/L at 2008, much higher than other years. It is worth mentioning that the average annual precipitation of Guizhou was 1842mm at 2008, while the annual precipitation in Guizhou was between 1100mm and 1300mm. This may emulate the extreme value of COD appeared in 2008 to some degree.

PH value of groundwater generally appears Weak Alkaline (around 7.5) through the filtration of silicate layer. However, Figure 4 demonstrates that groundwater acidification occurs at some stations in Guizhou, especially at Zunyi. PH value of Zunyi presents thin acid, which is much lower than that of other stations until 2007, as low as 6.17 at 2006. This may be affected by the formation lithology and urbanization process.
Figure 3. Flowchart summarizing the framework used to predict short time series groundwater quality with missing value.

Figure 4. Temporal variation of the measured concentrations of PH and COD at monitoring stations in Guizhou province of China. Since Anshun is excluded from this study, because there are no observation samples of COD throughout the whole period (2001-2014).

3.2. Model performance evaluation
Time series data from 2001-2013 was used to predict PH and COD in 2014 with three different methods, among which MWMO-HMM, ARMA, and GM. It is worth highlighting that the missing data reparation process is only contained in MWMO-HMM. Three performance indicators, such as root mean square error (RMSE), the mean absolute error (MAE), and the Nash–Sutcliffe model efficiency coefficient ($R^2$) [14] for MWMO-HMM, ARMA, and GM for PH and COD predictions.
were derived (Table 1). In particular, ARMA displays the greatest errors for both PH and COD prediction, RMSE equals respectively 0.3829, 0.3645, while $R^2$ of ARMA also shows the worst correlation. MWMO-HMM has the optimal performance on both PH and COD predictions, RMSE are 0.1339 and 0.0387, $R^2$ rises to 0.9806.

| Water quality | Model type     | RMSE  | MAE  | $R^2$  |
|---------------|----------------|-------|------|--------|
| PH            | ARMA           | 0.3829| 0.3211| 0.0509 |
|               | GM             | 0.2487| 0.2134| -0.5267|
|               | MWMO-HMM       | 0.1339| 0.1166| 0.6151 |
| COD           | ARMA           | 0.3645| 0.2666| 0.1577 |
|               | GM             | 0.2573| 0.2226| 0.4022 |
|               | MWMO-HMM       | 0.0387| 0.0294| 0.9806 |

Figure 5 shows that predicted result of MWMO-HMM is closer to 1:1 line, and ARMA prediction results of both PH and COD deviate from 1:1 line. As can be seen from numerical distribution, PH prediction results of GM is relatively lower than the measured values, while GM demonstrates the opposite in COD prediction. The results state that GM method is less stable for time series prediction.

Figure 6 demonstrates the absolute errors (AEs) of results predicted with different methods, the closer to zero, the higher the prediction accuracy is. The AEs of COD predicted by MWMO-HMM are smaller than 0.1 at all monitoring stations and the AEs of PH prediction results are lower than 0.22. The forecasting precision of ARMA is quite unstable, since the AEs at several stations are much higher than GM and MWMO-HMM, this could signal that ARMA method shows more dependence to the integrity of original data set. By contrast, MWMO-HMM projects the least prediction error because MWMO-HMM not only takes the influence of missing value into account, but also considers weighting low-order and high-order HMM, which can effectively avoid the impact of data uncertainty.
Figure 6. Absolute error of the predicted concentrations of PH and COD at monitoring stations with ARMA, GM and MWMO-HMM methods.

4. Conclusions
This study provides a significant contribution to groundwater quality prediction of short time series observations with missing values. The developed MWMO-HMM model integrates EM algorithm with a weighted multi-order HMM model. The following primary conclusions were reached:

1) EM algorithm is a good way of filling the missing values in the observation dataset. Furthermore, the inter-annual variations of COD in Guizhou seem to be affected by surface precipitation to a certain extent (Figure 4).

2) The performance of MWMO-HMM is better than the other two commonly used methods (ARMA and GM) on both PH and COD forecasting. The Nash–Sutcliffe model efficiency coefficients of PH and COD results predicted by MWMO-HMM are respectively 61.51% and 98.06%. As a contrast, GM and ARMA both demonstrate an unstable predictive state when dealing with prediction on short-term sample dataset including missing value (Table 1).

These results come as a new perspective of this study particularly, considering that no prior models had been conducted to tackle the realistic problem that how to predict groundwater quality based on small sample data sets with missing values. In the future research, the applicability of MWMO-HMM can be tested by predicting other environmental pollution factors or other regions with more available observations.

Acknowledgment
This work is jointly supported by grants from GuiZhou Welfare and Basic Geological Research Program of China (Grant number QianGuoTuZiYuanHan[2014]No.23,[2016]No.269), the National Natural Science Foundation of China (Grants number 41771380), the Guangdong Innovative and Entrepreneurial Research Team Program (2016ZT06D336), and the Guangdong Academy of Sciences’ Project of Science and Technology Development (2017GDASCX-0801, 2017GDASCX-0101).

References
[1] Froukh L J 2003 Water Resour. Manag. 17 175
[2] Izady A, Abdalla O, Joodavi A, Chen M 2017 Water 9(3) 161
[3] Palani S, Liong S Y, Tkalich, P 2008 Marine Pollution Bulletin 56(9) 1586
[4] Wang X Y, Zhao X P, Liu Z W, Dong S Q 2011 Computer Simulation 28(1) 17
[5] Faruk D Ö 2010 Engineering Applications of Artificial Intelligence 23(4) 586
[6] Mahapatra S S, Nanda S K, Panigrahy B K 2011 Advances in Engineering Software 42(10) 787
[7] Eynard J, Grieu S, Polit M 2011 Engineering Applications of Artificial Intelligence 24(3) 501
[8] Jalalkamali A 2015 Earth Science Informatics 8(4) 885
[9] Xu L, Liu S 2013 Mathematical & Computer Modelling 58(s 3–4) 807
[10] Zhu C, Wu L 2009 Computational Intelligence and Industrial Applications 46
[11] Xiao S Z, Xiong K L, Lan J C, et al. 2015 Environmental Science 5 1590
[12] Dempster A P, Laird N M, Rubin D B 1997 J. Roy. Stat. Soc. B. 39 1
[13] Shen Y L, Wu L X, Di L P, et al. 2013 Remote Sensing 5(4) 1734
[14] Heij C, de Boer P, Franses P H, et al. 2004 Econometric Methods with Applications in Business and Economics; Oxford University Press Inc.: New York, NY, USA