Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2017.20186

64 (2017) 48–53
February

Grey prediction model of water quality based on clustering and fusion

Yuhong Du[a,b,*], Kunpeng Wei[a,b,*], Enhua Liu[a], Liancheng Wang[c], Qiyin Feng[a,b], Guangyu Dong[a,b]

[a]*Tianjin Polytechnic University, School of Mechanical Engineering, No. 399 West Binshui Road, Xi Qing District, Tianjin, China 300387, Tel. +86-156-2073-3265; email: 386731096 @qq.com (Y. Du), Tel. +86-137-5266-8841; email: 13752668841@163.com (K. Wei), Tel. +86-137-5215-3077; email: 308729945@qq.com (E. Liu), Tel. +86-156-2051-4875; email: 1286873614@qq.com (Q. Feng), Tel. +86-150-4981-3256; email: 609295126@qq.com (G. Dong)*
[b]*Advanced Mechatronics Equipment Technology, Tianjin, China 300387*
[c]*Tianjin Geothermal Exploration and Development-Designing Institute, Tianjin, China 300250, Tel. +86-186-3039-6615; email: 691815779@qq.com*

Received 20 April 2016; Accepted 21 August 2016

ABSTRACT

To effectively predict water quality parameters of water treatment control systems, this paper proposes a grey prediction model of water quality based on clustering fusion. The model uses the clustering fusion method to process the data collected by the sensors, and the processed data are used as the original input data of the grey forecasting control. At the same time, the output data of the grey forecasting control are compared with the sensor data after fusion to determine the forecast value. Finally, accurate forecast values of water quality in the system are obtained. Many data were obtained after running the system; grey forecasting control model based on clustering and fusion provides system parameters. The mean absolute percentage error of water quality characteristics of conductivity, dissolved oxygen and turbidity was 0.38%, 9.91% and 9.16%, respectively. The results highlight that the proposed method is better than the single grey forecasting method; thus, it can guarantee the different water quality parameters remain stable and meet the water quality requirements.

Keywords: Grey forecasting; Clustering fusion; Water treatment

1. Introduction

The current water treatment control system with large hysteretic nature and nonlinear, multi-parameter properties seriously affects product water quality [1]. The effective prediction of product water quality parameters change trend has become an important issue in improving water treatment control system and effectively stabilizes the water quality. Based on existing data to predict the future data of the system, grey predictive control theory is the most widely used forecasting method of reducing overshoot and response time to improve the stability of water treatment system [2,3]. Shi et al. [4] combined the grey metabolism (GM) (1,1) model and the BP neural network model in a model whose accuracy was better than that of the above-mentioned models taken individually. Seasonal artificial neural network (ANN) models were designed by Ying [5] to improve purification ability of wastewater. Xue et al. [6] used Markov chain correction residual error method to improve grey neural network, make the correction value closer to the actual value and improve the prediction accuracy. Li et al. [7] designed an adaptive grey predictive controller based on grey predictive control theory, improving control precision and adaptability. Xu [8] used the grey system and the ANN to predict and investigate the error of both predicted and actual values, and verified the model prediction accuracy. Xu [9] adopted the dynamic matrix predictive control algorithm to get satisfactory control,

[*] Corresponding author.

and Yan et al. [10] put forward a nonlinear grey Bernoulli model, which can significantly improve prediction accuracy. The prediction accuracy of Wang's unequal time distance weighted grey prediction model (UWGM (1,1 ω)) and the extended grey forecasting model (EGM (1,1)) were analyzed, concluding that EGM (1,1) is much more accurate in predicting the change trend of fouling thermal resistance [11].

According to specific differences, advantages, and disadvantages of the above-mentioned prediction analysis, based on the actual situation of water treatment, this paper proposes a grey prediction model based on cluster fusion. According to the model, the water quality information collected by multiple sensors is fused and processed, and processed data serve as the original data of grey prediction control. In the predictive control process, the relative error between the output data and the fusion data is monitored, and appropriate data through conditional judgment are selected. Providing a large number of practical data analyses, the model is more suitable for multi-sensor parameters of water treatment control system than a single grey prediction model.

2. Materials and methods

2.1. Grey prediction model of water quality based on cluster fusion

Grey forecasting model is a prior control, which predicts the future trend of development law of the system by its own behavior characteristics and determines the corresponding control decision. Grey prediction is required for the original data; however, the deviation of the predicted original data can lead to large differences between the predicted result and the true value. The existing measured data sets related to clustering and fusion are the original input data set of grey prediction by the method of cluster fusion. At the same time, the grey prediction value is compared with real-time fusion value, monitoring the error trend. When the error exceeds the set range, grey prediction data is modified according to fusion data. Fig. 1 shows the control strategy of the grey prediction model based on cluster fusion.

2.1.1. Grey prediction of water quality parameters

The steps of grey prediction modeling for the water treatment of the original data sequence with n variables $x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n)\}$ are as follows:

(1) Testing data: calculate first the class ratio of series:

$$\lambda(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, \quad k = 2,3,\cdots,n.$$

If all the class ratios fall within the range of the volume coverage $X = (e^{\frac{-2}{n+1}}, e^{\frac{2}{n+1}})$, the data column $x^{(0)}$ can be used to establish the grey model and thus for grey prediction. Otherwise, the data must be properly handled.

(2) Generated sequence processing: by taking the first-order accumulated generating operation (1-AGO) on $x^{(0)}$, the new sequence is:

$$x^{(1)} = \{x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n)\} \tag{1}$$

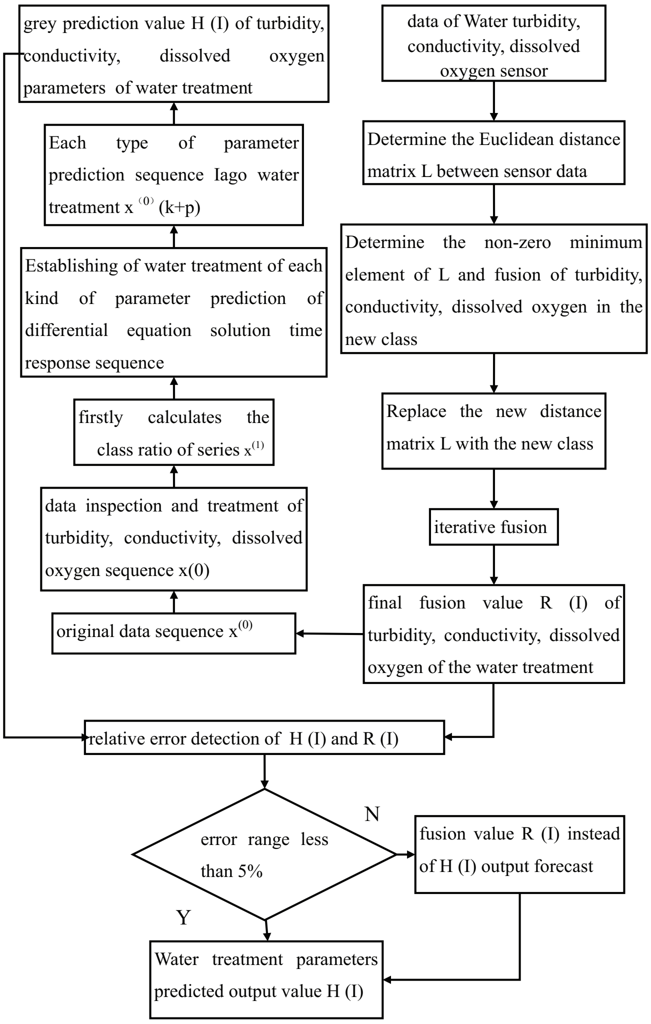
grey prediction value H (I) of turbidity, conductivity, dissolved oxygen parameters of water treatment
data of Water turbidity, conductivity, dissolved oxygen sensor
Each type of parameter prediction sequence Iago water treatment x (0) (k+p)
Determine the Euclidean distance matrix L between sensor data
Establishing of water treatment of each kind of parameter prediction of differential equation solution time response sequence
Determine the non-zero minimum element of L and fusion of turbidity, conductivity, dissolved oxygen in the new class
firstly calculates the class ratio of series x(1)
Replace the new distance matrix L with the new class
data inspection and treatment of turbidity, conductivity, dissolved oxygen sequence x(0)
iterative fusion
original data sequence x(0)
final fusion value R (I) of turbidity, conductivity, dissolved oxygen of the water treatment
relative error detection of H (I) and R (I)
N
error range less than 5%
fusion value R (I) instead of H (I) output forecast
Y
Water treatment parameters predicted output value H (I)

Fig. 1. Grey prediction model of water quality based on clustering and fusion.

where $x^{(1)}(k) = \sum_{m=1}^{k} x^{(0)}(m), \quad k = 1,2,\ldots n.$

(3) Model building: the differential equation of the $x^{(1)}$ sequence is:

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = b \tag{2}$$

The least square method is used to identify the model parameters a and b:

$[a\ b]^T = (B^T B)^{-1} B^T X_n$

where

$$B = \begin{bmatrix} -0.5[x^{(1)}(1) + x^{(1)}(2)] & 1 \\ -0.5[x^{(1)}(2) + x^{(1)}(3)] & 1 \\ \vdots & \vdots \\ -0.5[x^{(1)}(\mathrm{n}-1) + x^{(1)}(\mathrm{n})] & 1 \end{bmatrix}$$

$$x_n = [x^{(0)}(2), x^{(0)}(3), \ldots, x^{(0)}(n)]^T$$

After determining a and B, the solution of Eq. (2) is:

$$x^{(1)}(k+1) = \left[x^{(0)}(1) - \frac{b}{a}\right]e^{-ak} + \frac{b}{a} \tag{3}$$

where a is the development coefficient; b is the grey action quantity. Use the inverse accumulated generating operation (1-IAGO) to get predictive sequence, that is:

$$x^{(0)}(k+1) = x^{(1)}(k+1) – x^{(1)}(k)$$

The p step forward to the predictive value of the original data column is [12]:

$$x^{(0)}(k+p) = (\mathrm{x}^{(0)}(1) - \frac{b}{a})(1-e^{a})e^{-a(k+p-1)} \tag{4}$$

(4) Testing residual: calculated relative residuals:

$$\varepsilon(k) = \frac{x^{(0)}(k) \text{ - } \hat{x}^{(0)}(k)}{x^{(0)}(k)}, \quad k = 1, 2, \ldots, n$$

If the absolute range of absolute value of the residual error is satisfied, i.e., $|\varepsilon(k)| < 0.1$, it is assumed that the prediction accuracy is high. If the absolute range of absolute value of the residuals is satisfied when $|\varepsilon(k)| < 0.2$, the prediction accuracy can comply with general requirements.

2.1.2. Clustering fusion of water quality parameters

There are k group sensor measured data of n water treatment characteristic parameters. The n vectors, $X_i = (X_{i1}, X_{i2} \cdots, X_{in})^T$, represent the i-th group data. The distance between two measured data X_b and X_a is defined according to the Euclidean distance formula [13]:

$$L_{ab} = [(X_a - X_b)^T(X_a - X_b)]^{1/2} \tag{5}$$

The smaller the value of L_{ab} the closer X_b and X_a otherwise, the greater the deviation. The distance matrix L for all sensors is:

$$L = \begin{bmatrix} l_{11} & \cdots & l_{1\mathrm{k}} \\ \vdots & & \vdots \\ l_{\mathrm{k}1} & \cdots & l_{\mathrm{kk}} \end{bmatrix}$$

After getting the distance matrix, the fusion is carried out by the following steps:

(1) Consider the measured data of the i-th sensor X_i as a class, expressed with Φ_i (being $i = 1, \cdots, k$). Assume the selected element is l_{ij}, then Φ_i and Φ_j are merged into a new class, denoted by $\Phi_f = \{\Phi_i \Phi_j\}$.

(2) After eliminating the i-th line and the j-th column of the distance matrix L, add new rows and columns that are consistent with a new class of Φ_f with other unincorporated categories of the distance, and get new distance matrix $L_{(1)}$.

(3) Based on $L_{(1)}$, repeat steps (1) and (2) to obtain $L_{(2)}$. Iterate the procedure until the k group measured data together as a class.

(4) The new class $\Phi_f = \{\Phi_i \Phi_j\}$ uses Odeberg's algorithm for fusion [14]:

$$f(x_{ia}, x_{ja}) = \frac{c(x_{ia} + x_{ja}) + (c-1)^2 x_{ia} \cdot x_{ja}}{1 + c^2 - (c-1)^2(x_{ia} + x_{ja} - 2x_{ia} \cdot x_{ja})} \tag{6}$$

where $f(x_{ia}, x_{ja})$ is a numerical value corresponding to the a-th component fusion of sensor i, j measured data X_i, X_j.

According to the actual situation, the data of three parameters and six sets of measured data are used. Table 1 summarizes the six groups of measured values for conductivity, dissolved oxygen and turbidity.

From Table 1, $n = 3$ and $k = 6$; for each individual considered as a class, the distance matrix L is obtained by formula (5):

$$L = \begin{bmatrix} 0 & 364.5018 & 60.5010 & 242.0005 & 684.5018 & 162.0010 \\ & 0 & 128.0036 & 12.5040 & 50.0037 & 40.5037 \\ & & 0 & 60.5005 & 338.002 & 24.5000 \\ & & & 0 & 112.5015 & 8.0005 \\ & & & & 0 & 180.5002 \\ & & & & & 0 \end{bmatrix}$$

The smallest element of L $l_{46} = 8.0005$. Therefore, merge Φ_4 and Φ_6 into a new class Φ_7, to get the distance between Φ_7 and the rest of the class:

$$l_{71} = \min\{l_{41}, l_{61}\} = 162.0010$$

$$l_{72} = \min\{l_{42}, l_{62}\} = 12.5040$$

$$l_{73} = \min\{l_{43}, l_{63}\} = 24.5000$$

$$l_{75} = \min\{l_{45}, l_{65}\} = 112.5015$$

Table 1
The six groups of measured values for three water quality parameters

	X_{i1} Conductivity (μs/cm)	X_{i2} Dissolved oxygen (mg/l)	X_{i3} Turbidity (NTU)
1	402	0.09	0.13
2	429	0.15	0.12
3	413	0.07	0.09
4	424	0.06	0.12
5	439	0.08	0.07
6	420	0.07	0.09

Elimination of Φ_4, Φ_6 corresponding rows and columns in L plus the distance of the corresponding rows and columns of Φ_7 to Φ_1, Φ_2, Φ_3, Φ_5 is $L_{(1)}$. Repeat all the steps above steps until the rest of the class are merged into a large class. Use formula (6), and take $c = 1.1$, at first fuse the measured data X_4, X_6 gives:

$$f(x_{41},x_{61}) = \frac{c(x_{41}+x_{61})+(c-1)^2 x_{41}\cdot x_{61}}{1+c^2-(c-1)^2(x_{41}+x_{61}-2x_{41}\cdot x_{61})} = 422$$

$$f(x_{42},x_{62}) = 0.0648$$

$$f(x_{43},x_{63}) = 0.1047$$

Therefore, X_7 = $(422\ 0.0648\ 0.1047)^T$. Similarly, X_7, X_5 fusion gets $X_8 = (430\ 0.0721\ 0.0871)^T$; X_8, X_3 fusion gets X_9 = $(421\ 0.0708\ 0.0882)^T$; and X_9, X_1 fusion gets X_{10} = (411 0.0801 0.1088)^T. Finally fuse X_{10}, X_2. The final fusion value is X_{11} = $f(X_1 \ldots X_6) = (420\ 0.1147\ 0.1140)^T$.

2.2. *Water treatment process*

Nanofiltration technology is the core of water treatment technology project. After adding the fungicide and coagulant in the coagulation pool and mixing, water flows through the inclined tube sedimentation tank for slurry separation and then into the sand filter tank. Then water passes through deoxygenation tank for removing most dissolved oxygen in the water; precision filter removes the impurities. Then, water enters the nanofiltration membrane system. The addition of a small amount of alkali and oxygen scavenger in the membrane treated water allows meeting the requirements of water treatment. Electrical conductivity, turbidity, dissolved oxygen content and pH values are the main parameters that determine water quality. Turbidity is a key test of water clarity. Conductivity can indirectly reflect the total concentration of inorganic salts and charged colloids in water, and it is an important indicator to detect the effect of desalination. Dissolved oxygen is the main parameter to measure water pollution. Therefore, this paper chooses the aforementioned three water quality parameters to investigate the grey prediction model of water quality based on cluster fusion.

The system adopts feedforward and feedback control method. Fig. 2 illustrates the detection control system flow chart.

3. Results and discussion

150 groups of water quality data in the same given time period were obtained by field measurements. The first 100 groups were regarded as learning data, while the latter 50 were used as sample data of the test. In cluster fusion process, a total of 6 values, corresponding to different water quality parameters, are a set of data. Then, data are clustered into a fusion value. At the same time, the fusion value and the first five fusion values are taken as a set of data; then the set of data is used as the original input data of grey prediction. Grey prediction step size take $P = 4$. The error monitoring range of grey prediction value and fusion value is 5%.

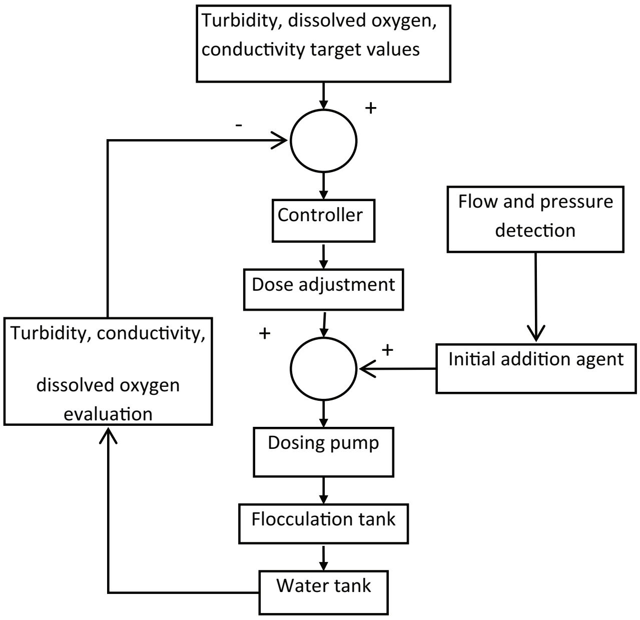
Turbidity, dissolved oxygen, conductivity target values
+
-
Controller
Flow and pressure detection
Dose adjustment
+
+
Turbidity, conductivity, dissolved oxygen evaluation
Initial addition agent
Dosing pump
Flocculation tank
Water tank

Fig. 2. Detection and control flow chart.

The comparisons of the prediction accuracy of the single cluster fusion method, the single grey prediction method, and the grey prediction model based on cluster fusion to conductivity, dissolved oxygen and turbidity are introduced in the following sections.

3.1. *Comparative analysis of prediction models for electrical conductivity*

Fig. 3(a) shows that the change tendency of the predicted output of the single grey forecast, multiple model forecast and the change tendency of the measured value are basically the same. The mean absolute difference of single grey forecast, single clustering fusion forecast and multiple model forecast are 1.8845, 2.8154 and 1.5583 μs/cm, respectively. Moreover, the mean absolute percentage error of single grey forecast, single clustering fusion forecast and multiple model forecast are 0.45%, 0.68% and 0.38%, respectively. The electric conductivity predicted accuracy of the grey prediction model based on cluster fusion is better than that of the other prediction models.

3.2. *Comparative analysis of prediction models for dissolved oxygen*

In Fig. 4(a). the fluctuation range of single grey prediction curve is beyond the range of measured value. The change tendency of the predicted output of multiple model forecast and single cluster fusion are basically the same. The mean absolute difference of single grey forecast, single clustering fusion forecast and multiple model forecast are 0.0901, 0.0730 and 0.0715 mg/l, respectively. The mean absolute percentage error of single grey forecast, single clustering fusion forecast and multiple model forecast are 12.97%, 10.15% and 9.91%,

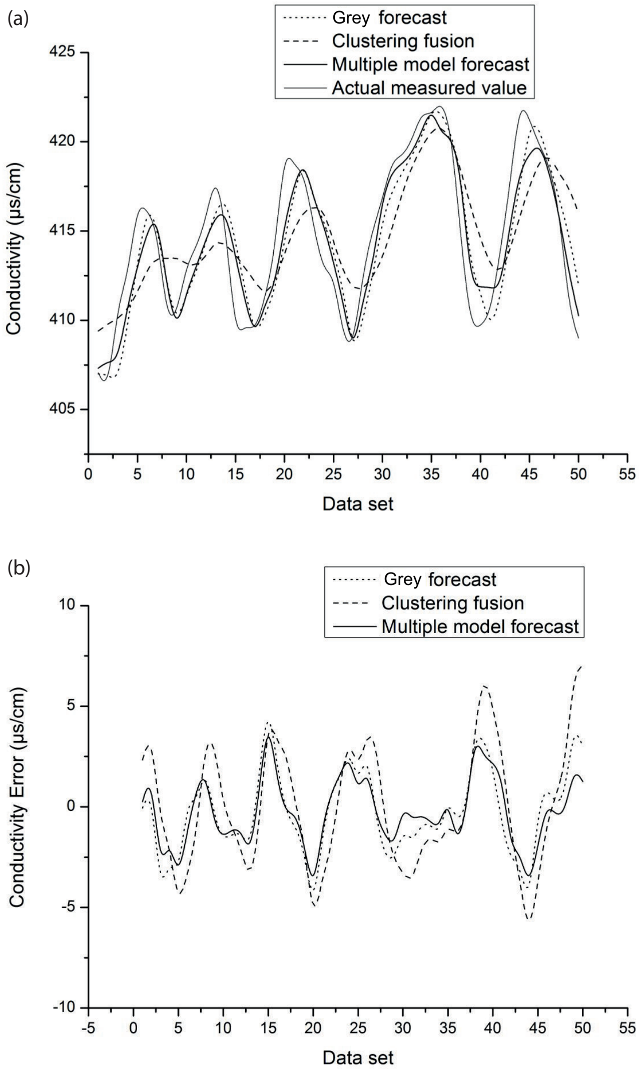

Fig. 3. Forecast error values of electric conductivity: (a) predicted output and actual measured curves and (b) predicted output error curves.

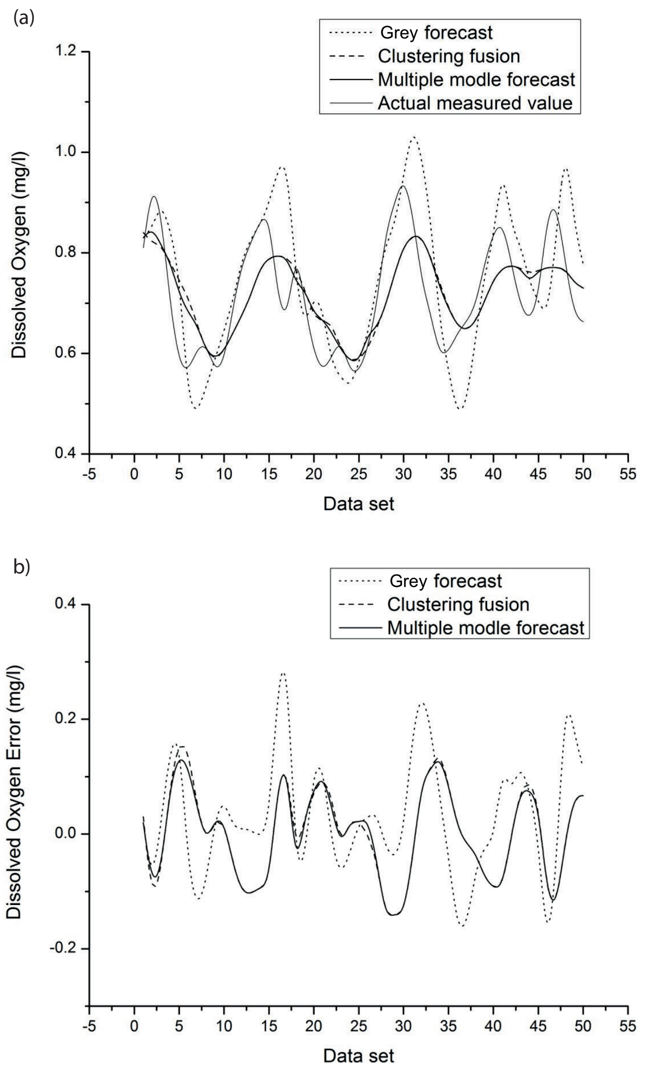

Fig. 4. Forecast error value curves of dissolved oxygen: (a) predicted output and actual measured curves and (b) predicted output error curves.

respectively. The dissolved oxygen predicted accuracy of the grey prediction model based on cluster fusion is better than that of the other prediction models.

3.3. Comparative analysis of prediction models for turbidity

From Fig. 5(a), the change tendency of the predicted output of the three models and the measured value are basically the same. The curve trend of composite model matches the curve trend of single cluster fusion. The cluster fusion method has thus a significant effect on the single grey prediction, weakening the adverse effects of small amplitude fluctuations. The mean absolute difference of single grey forecast, single clustering fusion forecast and multiple model forecast are 0.0114, 0.0089 and 0.0087 NTU, respectively. The mean absolute percentage error of single grey forecast, single clustering fusion forecast and multiple model forecast are 12%, 9.26% and 9.16%, respectively. The turbidity predicted accuracy of the grey prediction model based on cluster fusion is better than that of the single grey prediction and single cluster fusion.

Plots in Figs. 3–5 highlight that measured values and predicted output values of the three parameters have same trend and same amplitude. Therefore, one can conclude that the mean absolute percentage error of the grey prediction model based on cluster fusion is the smallest in the three models, so it has the highest prediction accuracy. In predicting water quality, the grey prediction model based on cluster fusion is advantageous. However, in the process of clustering fusion, the grey prediction model based on cluster fusion uses the six sets of continuous sampling values including the previous time, so its predictive output value is slightly behind the measured value, as shown in Figs. 3–5. To reduce the hysteretic nature, this paper will continue to improve the model by reducing the number of fusion groups and the sampling time appropriately.

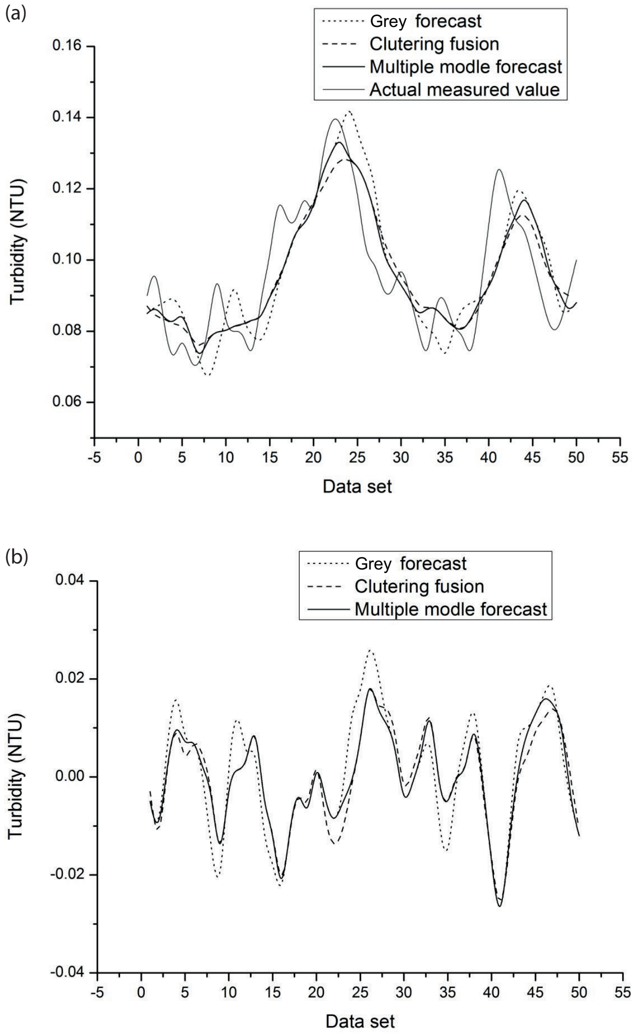

Fig. 5. Forecast error value of turbidity: (a) predicted output and actual measured curves and (b) predicted output error curves.

4. Conclusion

Through the grey prediction model of water quality based on cluster fusion, 150 sets of data have been predicted, and the correctness of the algorithm was verified using 50 sets of data. The results showed that the predicted values of conductivity, dissolved oxygen and turbidity were in good agreement with the measured values, and the mean absolute percentage error were 0.38%, 9.91% and 9.16%, respectively. The predicted output curve of electrical conductivity is in line with the grey prediction curve, and the curves of the predicted output value of the turbidity and dissolved oxygen are more close to the curve of the cluster fusion value.

Practical applications show the feasibility of the application of the clustering fusion method and grey prediction to the control process of water treatment. The method is stable and accurate, and can be applied to electrical conductivity, dissolved oxygen, and turbidity sensor measurements. The prediction performance is better than that of the single grey prediction method. The predictive control method can ensure that when the water quality of raw water is unknown the water quality can be kept relatively stable, and it has a certain practical and reference value.

Acknowledgment

The authors acknowledge the financial support from National Natural Science Fund for youth project (51205288), and the Tianjin Municipal Science and Technology Commission project (13jcybjc15900).

References

[1] Q.W. Qiu, Study of Fuzzy Control in the Wastewater Treatment Processes, Master's Thesis, Zhejiang University, China, 2011.

[2] R.Z. Li, Advance and trend analysis of theoretical methodology for water quality forecast, China, J. HeFei Univ. Technol., 1 (2006) 26–30.

[3] J. Zhang, Improvement of Grey Forecasting Model and Its Application, Master's Thesis, Xi'an University of Technology, China, 2008.

[4] W.R. Shi, Y.X. Wang, Y.J. Tang, M. Fan, Water quality parameter forecast based on grey neural network modeling, J. Comput. Appl., 29 (2009) 1529–1531, 1535.

[5] Y. Zhao, L. Guo, J. Liang, M. Zhang, Seasonal artificial neural network model for water quality prediction via a clustering analysis method in a wastewater treatment plant of China, Desal. Wat. Treat., 57 (2016) 3452–3465.

[6] P.S. Xue, M.Q. Feng, X.P. Xing, Water quality prediction model based on Markov chain improving gray neural network, Eng. J. Wuhan Univ., 3 (2012) 319–324.

[7] G.F. Li, X. Zhang, Z.Y. Wang, Grey prediction control in pretreatment system conductivity control, Power Syst. Eng., 1 (2011) 65–66.

[8] Y. Xu, Study on Activated Carbon Adsorption of Dye Wastewater Based on Grey System Theory, Master's Thesis, South China University of Technology, China, 2011.

[9] P.B. Xu, Research on Automatic Coagulant Dosage in Water Treatment Based on Lonworks and Predictive Control, Master's Thesis, Harbin Polytechnic University, China, 2006.

[10] A. Yan, Z.H. Zou, Y.F. Zhao, Forecasting of dissolved oxygen in the Guanting reservoir using an optimized NGBM (1,1) model, J. Environ. Sci., 3 (2015) 158–164.

[11] J.G. Wang, G.S. Liu, W. Sun, Experiment Research on Fouling Resistance Grey Forecast Model for Circulating Water in Electromagnetic Field, Proc. CSEE, Vol. 11, 2013, pp. 61–68.

[12] J.L. Yang, Study on the Application of Grey Fuzzy PID Algorithm in Flocculating Sedimentation Process Control of Coal Slime, Master's Thesis, Taiyuan University of Technology, China, 2012.

[13] S.P. Wan, Method of clustering fusion for multi-sensors data, Syst. Eng. Theory & Practice, 28 (2008) 131–135.

[14] H. Odeberg, Fusing sensor information using fuzzy measures, Robotica, 12 (1994) 465–472.