Appl Water Sci (2017) 7:1151–1160
DOI 10.1007/s13201-016-0458-4

ORIGINAL ARTICLE

GMDH algorithms applied to turbidity forecasting

Tsung-Min Tsai[1] · **Pei-Hwa Yen**[1]

Received: 11 October 2013 / Accepted: 31 August 2016 / Published online: 12 September 2016

Abstract By applying the group method of data handling algorithm to self-organization networks, we design a turbidity prediction model based on simple input/output observations of daily hydrological data (rainfall, discharge, and turbidity). The data are from a field test site at the Chiahsien Weir and its upper stream in Taiwan, and were recorded from May 2000 to December 2008. The model has a regressive mode that can assess the estimated error, i.e., whether a threshold has been exceeded, and can be adjusted by updating the field input data. Consequently, the model can achieve accurate estimations over long-term periods. Test results demonstrate that the 2006 turbidity prediction model was selected as the best predictive model (RMSE = 5.787 and CC = 0.975) because of its ability to predict turbidity within the acceptable error range and 90 % required confidence interval (50NTU). 70(3,1,1) is the optimum modeling data length and variable combinations.

Keywords GMDH · Turbidity forecast · Nanhua Reservoir · Chiahsien Weir · Over-basin diversion

✉ Tsung-Min Tsai
canl0223@gmail.com

Pei-Hwa Yen
yenph@mail.ncku.edu.tw

[1] Department of Hydraulics and Ocean Engineering, National Cheng Kung University, No. 1, University Road, 70101 Tainan, Taiwan, ROC

Introduction

Water consumption in Taiwan has increased significantly in recent years. The Water Resources Agency and the Taiwan Water Corporation have raised certain issues regarding the quantity and quality of water. According to statistical data, the island's average annual rainfall is approximately 2515 mm. Despite its abundance, rainfall is unevenly distributed in terms of both time and space. Because of the island's steep natural terrain, short river flows, and geological weaknesses, the majority of rainwater flows out to sea before it can be harnessed for public use. Thus, reservoirs are an essential means of realizing effective water usage. From the viewpoint of water resource management, both the availability and quality of water are a concern.

One of the water resources of the Nanhua Reservoir is the discharge of the Cishan River, which is diverted through a tunnel from the Chiahsien Weir. The majority of water diversion occurs during the annual wet period, from June to October, which is also the typhoon season. Because of the adverse effect of soil degradation in the upstream catchment area, heavy rainstorms rapidly and significantly increase the Cishan River discharge; they also increase the sand content and turbidity. If this flow is allowed to persist and enter the Nanhua Reservoir, the level of reservoir sediment will undoubtedly increase, potentially shortening the lifespan of the reservoir and creating problems for the operation of the Nanhua water-treatment plant.

This study examines the relevant hydrological data variables that influence water turbidity in the Chiahsien Weir. A unique group method of data handling (GMDH) multilayer algorithm is used to deduce the relationship between groups of input variables and output functions. The result is combined into a suitable set of higher-order

nonlinear equations that engender a simple turbidity-forecasting model. This enables the prediction of water turbidity, and provides pertinent reference turbidity information for the Chiahsien Weir water diversion operation.

Methodology

The GMDH algorithm introduced by Ivakhnenko (1968) is a heuristic self-organization process that establishes an input–output relationship within a complex system. It utilizes a multilayered conceptual structure, similar to a feed-forward multilayer neural network. Ikeda et al. (1976) added a recursive procedure to the GMDH algorithm to utilize updated observation data and to modify parameters within the nodes of each layer, enabling time-variable modeling. They subsequently applied the enhanced model to the prediction of daily river flows. Tamura and Kondo (1980) utilized the prediction of sum-of-squares or Akaikes's information criterion as parameter selection indicators. Because the algorithm can easily generate high-level nonlinear terms, this nonlinear dynamic system can be well defined; however, its practicality would be seriously reduced. In response, Yoshimura et al. (1982) improved the model with a stepwise regressive procedure, returning the complex final system to a low-level nonlinear system, thereby increasing its applicability.

The GMDH algorithm enables the automatic selection of input variables during model construction, as well as a hierarchical polynomial regression of necessary complexity (Farlow 1984). Specific functional dependence between the input and output variables is unnecessary, as the dependence has been incorporated into the modeling structure. The GMDH algorithm has been applied in various fields, e.g., weather modeling, pattern recognition, physiological experiments, cybernetics, medical science, education, ecology, safety science, economics, and hydraulic field engineering systems (Lebow et al. 1984; Ivakhnenko et al. 1994; Kondo et al. 1999; Chang and Hwang 1999; Sarycheva 2003; Pavel and Miroslav 2003; Hwang et al. 2009; Tsai et al. 2009; Najafzadeh et al. 2013, 2014, 2015; Najafzadeh 2015). Nevertheless, few studies have explored turbidity modeling.

GMDH algorithm

The GMDH algorithm is a kind of feed-forward network, normally classified as a special type of neutral network. The model's underlying concept resembles animal evolution or plant breeding, as it adheres to the principle of natural selection. The multilayer criteria preserve superior networks for successive generations, eventually yielding an optimal network. This network (equation) more closely describes the physical phenomena that the model is intended to simulate. The self-organization algorithm can be classified as GMDH, SGMDH (stepwise regressive GMDH), and recursive/sequential GMDH. These model types are described below.

In the GMDH algorithm, the general connection between input and output variables is expressed by the Volterra functional series of the Kolmogorov–Gabor polynomial (Madala and Ivakhnenko 1994):

$$y(t) = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk} x_i x_j x_k + .., \quad (1)$$

where y(t) is the output variable, $X(x_1,x_2,\ldots,x_m)$ is the vector of input variables, and $A(a_1,a_2,\ldots,a_m)$ gives the vector coefficients or weights.

The GMDH model based on heuristic self-organization was developed to overcome the complexity of large-dimensional problems. It first pairs variables that might affect the system, and sets a default threshold to eliminate variables that cannot achieve a certain level of performance. This procedure describes a self-organization algorithm; it is a fundamental concept of derivative hierarchical multilevel models. The GMDH was built according to the following steps:

Step 1: Divide the original data into training and test sets

The original data are separated into training and test sets. The training data are used to estimate certain characteristics of the nonlinear system, and the test data are then applied to determine the complete set of characteristics.

Step 2: Generate combinations of input variables in each layer

All combinations of r input variables are generated for each layer. The number of combinations is given by:

$$C_r^m = \frac{m!}{r!(m-r)!}, \quad (2)$$

where m is the number of input variables and r is usually set to two (Ivakhnenko 1971).

Step 3: Optimization principle for elements in each layer

Optimum partial descriptions of the nonlinear system are calculated by applying regression analysis to the training data. The optimum standard uses the root mean square (RMS) as an index to screen out underperforming elements in each layer. RMS is defined as:

$$r_i = \left[\frac{\sum_{t=1}^{n} \left(y(t) - Z_i^k(t) \right)^2}{\sum_{t=1}^{n} \left(y(t) \right)^2} \right]^{1/2}, \quad (3)$$

where r_i is the RMS, $t = 1, 2, \ldots n$, n represents the length of the measurement data, $y(t)$ is the measured value at moment t; and $Z_i^k(t)$ is the output value of element i in layer k.

Step 4: Stopping rule for multilayer structure generation

By comparing the index value of the current (competent) layer with that of the next layer to be generated, further layers are prevented from being developed if the index value does not improve or falls below a certain objective default value; otherwise, Steps 2 and 3 are repeated until the value matches the limited condition set above.

After the above steps have been completed, all competent elements in each layer are recombined as an optimum high-level nonlinear equation. This is utilized as the final model for turbidity forecasting.

Stepwise regressive GMDH algorithm

The process of the stepwise regressive GMDH algorithm is very similar to that of the original GMDH algorithm. The key difference is that the least-squares method is replaced by a stepwise regressive procedure in Step 2. This procedure evaluates the optimum forward state, and determines whether it is more accurate than the next variable to be introduced. If so, it is incorporated into the model; otherwise, it is deleted to ensure the most precise simplified system equation. The assessment method employs the F-test for statistical analysis.

Recursive/sequential GMDH algorithm

Because of real-world time-variable characteristics, the system should respond to situations in real time. If the measured input data conceal the errors, or if the system is affected by human or natural factors, model parameters may no longer be applicable to the circumstances. The model forecasts will deviate and affect the overall precision of the model. To resolve this, the forecast model is revised using a recursive structure, thus allowing the system parameters to be modified in real time. This procedure can improve the forecast accuracy. In the GMDH algorithm, each progressive output element is composed of two prior elements with six parameters in a two-dimensional second-order equation. Thus, the system has an n-set of data, and the parameters (θ) of the newly composed equations of each layer are forecast as $Y_n = X_n \times \theta$. When the $n + 1$ data point is added, the system parameter θ can be updated to θ^* according to:

$$\theta_{n+1}^* = (X_{n+1}^t X_{n+1})^{-1} X_{n+1}^t Y_{n+1}. \quad (4)$$

Upon completion of the recursive procedure, the system parameters can be adjusted to ensure model optimality.

Establishment and assessment of the turbidity forecast model

Establishment of a turbidity-forecasting model

We now apply self-organizing nonlinear models for GMDH and SGMDH. The GMDH turbidity forecast model is developed according to the procedure described below. Figure 1 presents a flowchart of turbidity forecasting.

1. Obtain turbidity-related historical data, such as turbidity, rainfall, and discharge, at specific stations.
2. Select the input variables.

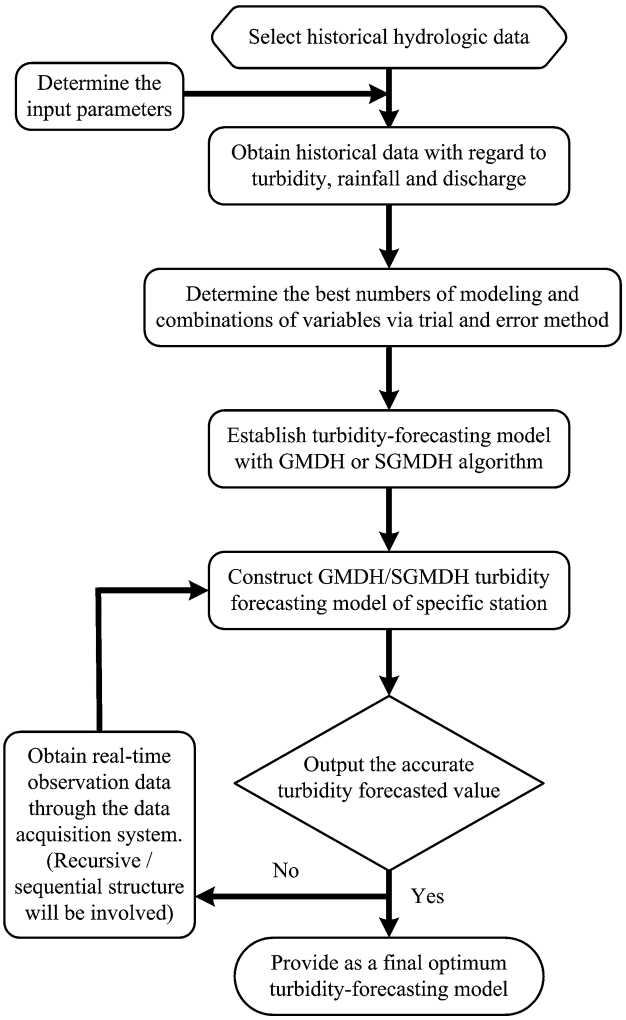

Fig. 1 Flowchart of GMDH/SGMDH turbidity-forecasting model construction

(1) Assume the output variable is Y, which represents the forecast turbidity.
(2) Assume the input variables are X_1, X_2, X_3,…, X_m, which represent turbidity, rainfall, discharge, and so on.
(3) Establish a nonlinear equation $Y = f(X_1, X_2, \ldots, X_m)$.

3. Determine the optimum number of modeling data and variable combinations to establish a forecast model through trial-and-error.
4. Establish an input–output relationship with both the GMDH and SGMDH algorithms; derive the model layer-by-layer until optimality is achieved, and then return, layer-by-layer, to the inertial input layer to establish a GMDH or SGMDH forecast equation.
5. Input the variables and begin model forecasting.
6. Output the forecast results.
7. Consider whether there is a temporal impact. If so, a recursive/sequential structure is necessary.
8. Generate a final optimum turbidity-forecasting model.

Based on the previous step, the output variable of the forecasting model is the turbidity MUD(t) at time t, where t represents the time period. The input variables are the daily turbidity $T(t-1) \sim T(t-m)$ for the period $1 \sim m$, daily rainfall $R(t-1) \sim R(t-n)$ for the period $1 \sim n$, and daily discharge of the Cishan River $Q(t-1) \sim Q(t-k)$ for the period $1 \sim k$. The forecast relation is presented below:

$$\begin{aligned}\mathrm{MUD}(t) = F(&T(t-1), T(t-2), \ldots, T(t-m), R(t-1),\\ &R(t-2), \ldots, R(t-n), Q(t-1),\\ &Q(t-2), \ldots, Q(t-k)).\end{aligned} \tag{5}$$

Model efficiency evaluation

The model can be evaluated by comparing its predictions to the measured values. The efficiency of the model is evaluated using the root mean square error (RMSE) and the coefficient of correlation (CC):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{T=1}^{N}\left(X_T - \hat{X}_T\right)^2}{N}} \tag{6}$$

$$\mathrm{CC} = \frac{\sum_{T=1}^{N}\left(X_T - \bar{X}\right)\left(\hat{X}_T - \bar{\hat{X}}\right)}{\sqrt{\sum_{T=1}^{N}\left(X_T - \bar{X}\right)^2 \sum_{T=1}^{N}\left(\hat{X}_T - \bar{\hat{X}}\right)^2}}, \tag{7}$$

where X_T is the observed value, $\hat{X}_T$ is the predicted value, $\bar{X}$ is the mean observed value, $\bar{\hat{X}}$ is the mean predicted value, and N represents the total number of observations in the data set. RMSE values approaching 0 and CC values approaching 1 signify better forecast performance.

Case studies

In this section, we compare the results given by our forecast model with real-world data. We first describe the study area and the data set used for comparison; then, we present the forecast results and evaluate the model's performance.

Study area description

Nanhua Reservoir is located 40 km northeast of Tainan, Taiwan, and approximately 15 km south of the Tseng-Wen Reservoir. Its catchment area is approximately 104 km^2. Figure 2 illustrates the reservoir location.

Chiahsien Weir is located in Kaohsiung County, near the Cishan River in Jiashian Township, approximately 450 m upstream of the Jiashian Bridge. The weir is part of the over-basin diversion project of Nanhua Reservoir. Figure 3 presents the site layout. Excess water from the Cishan River is mainly diverted into the Nanhua Reservoir during the wet season. According to reports by the Water Resources Agency and the Taiwan Water Corporation, the Nanhua Reservoir is seriously sediment-impacted. Over-basin diversion has been reported to be the most likely cause of increases in the reservoir sediment level.

Selection of research data

This paper explores turbidity changes in the Nanhua Reservoir prior to over-basin diversion (i.e., turbidity changes at the diversion tunnel entrance of the Chiahsien Weir). Numerous variables, such as storms, human activities, and complex natural processes, affect turbidity. These influencing factors closely match the nonlinear structural model of the GMDH algorithm.

Those factors that have the greatest impact on turbidity were utilized as input parameters. Thus, turbidity, rainfall, and discharge were chosen as the domain input parameters. The turbidity at the entrance to the diversion tunnel of the Chiahsien Weir was selected as the main parameter. Rainfall data from the Jiashian rainfall station (the only rainfall station upstream of the diversion channel) and Cishan River discharge data were used as secondary parameters. Using the aforementioned nonlinear system, a predictive turbidity model was built, calibrated, and verified.

GMDH and SGMDH calibrated result comparison

Selection of best algorithm

In the early stages of modeling, the GMDH and SGMDH algorithms were subjected to a trial-and-error procedure. This was intended to select the best algorithm and, finally,

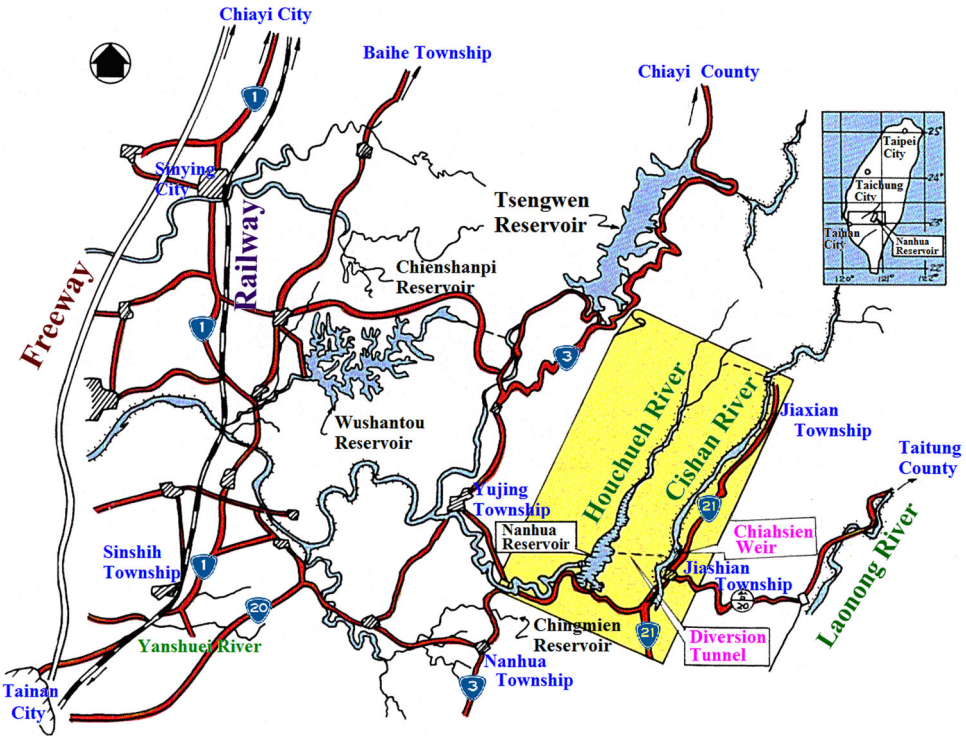

Fig. 2 Position of the study area (Nanhua Reservoir)

to obtain the optimum self-organizing nonlinear system for turbidity forecasting. The best performance results were analyzed by comparing the RMSE and CC given by different data sets and variable combinations. We used historical data from 2000 to 2008, as presented in Table 1. The SGMDH model gave better results in 2000, 2001, 2007, and 2008. The performance in the other years substantiates the assertion that the GMDH model generates better results, and so, this became the preferred model. The algorithm for all hierarchical regression parameters is shown in Table 2. The model was built over four levels, and the variable combinations differed between each level. As can be seen from the table, GMDH remains the best algorithm for computing the average RMSE over an 8-year period. The turbidity data from early 2004 displayed some abnormalities, which led to increased modeling errors. Table 1 presents the modeling results for turbidity data following these abnormal readings; these were not included in the averaging procedure. Eventually, the GMDH algorithm was selected as the most appropriate algorithm for this research.

$$
\begin{aligned}
Z_9^1 &= 16.62866 + 0.5386Q(t-1) - 0.00147R(t-1) + 0.31608Q(t-1)^2 - 0.00088R(t-1)^2 + 0.002Q(t-1)R(t-1)\\
Z_{10}^1 &= 3.60373 + 0.92624T(t-4) - 0.00036T(t-3) + 0.50314T(t-4)^2 + 0.00275T(t-3)^2 - 0.00903T(t-4)T(t-3)\\
Z_5^1 &= 0.29692Q(t-1) + 1.04566Q(t-1)^2 + 0.00148T(t-4)^2 + 0.0117Q(t-1)T(t-4)\\
Z_8^1 &= 24.90101 + 0.08174Q(t-1) - 0.00022T(t-1) + 0.42458Q(t-1)^2 - 0.0026T(t-1)^2 + 0.00479Q(t-1)T(t-1)\\
Z_2^2 &= -4.5464 + 1.64904(Z_9^1) - 0.0072(Z_{10}^1) - 0.8618(Z_9^1)^2 + 0.00757(Z_{10}^1)^2 + 0.0028(Z_9^1)(Z_{10}^1)\\
Z_3^2 &= 6.8366 + 1.20849(Z_9^1) + 0.0015(Z_5^1) - 0.78898(Z_9^1)^2 + 0.01028(Z_5^1)^2 - 0.00896(Z_9^1)(Z_5^1)\\
Z_4^2 &= 20.20312 + 0.22685(Z_8^1) + 0.0034(Z_{10}^1) - 0.22952(Z_8^1)^2 + 0.00681(Z_{10}^1)^2 - 0.00292(Z_8^1)(Z_{10}^1)\\
Z_4^3 &= -4.75467 + 2.95573(Z_4^2) - 0.0161(Z_3^2) - 1.83673(Z_4^2)^2 + 0.00561(Z_3^2)^2 + 0.01(Z_4^2)(Z_3^2)\\
Z_2^3 &= 20.80679 + 1.7062(Z_2^2) - 0.00304(Z_3^2) - 1.54077(Z_2^2)^2 + 0.01046(Z_3^2)^2 - 0.00323(Z_2^2)(Z_3^2)\\
Z_1^4 &= 16.08485 - 1.98537(Z_2^3) + 0.03366(Z_4^3) + 2.35709(Z_2^3)^2 + 0.0026(Z_4^3)^2 - 0.0336(Z_2^3)(Z_4^3)].
\end{aligned}
\tag{8}
$$

Choices of modeling data length and variable combinations

The modeling data length and variable combinations were obtained through a trial-and-error procedure, with a series of combinations of input variables. To develop the model, a sequential data length of 20, 30, 40, 50, 60, and 70 (days) was first introduced (taking into consideration simultaneous data integrity under no residual conditions). The optimum modeling data length differed annually, as can be determined from Table 3 by comparing the aforementioned assessment indicators between the trial-and-error procedures. The optimum data length is 70 in most cases, although it is 60 in 2003 and 2008, and 40 in 2006. According to the analysis results, a modeling data length of 70 is most appropriate for turbidity forecasting. Table 4 shows the results for the optimum modeling data lengths and variable combinations.

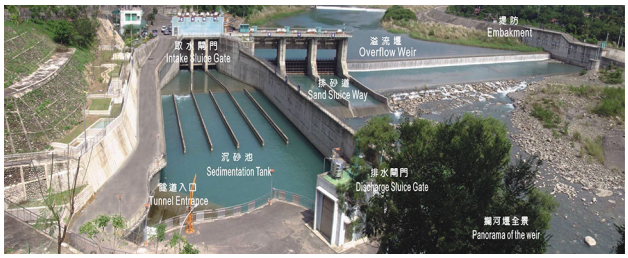

Fig. 3 Photograph of the Chiahsien Weir

Turbidity forecasting

Permissible errors

We adopted the safety concepts applied in general engineering construction projects, allowing a maximum error

Table 1 Comparison of evaluation indicators of GMDH and SGMDH forecast efficiency

Modeling event	GMDH forecast result			SGMDH forecast result		
	Modeling data length (variable combination)	RMSE (NTU)	CC	Modeling data length (variable combination)	RMSE (NTU)	CC
2000	70 (3, 0, 1)	57.920	0.211	70 (4, 3, 1)	52.905	0.126
2001	70 (2, 1, 0)	37.574	0.619	70 (4, 1, 2)	33.149	0.609
2002	70 (3, 1, 1)**	24.598	0.929	70 (3, 1, 1)	32.806	0.891
2003	60 (4, 1, 1)	22.750	0.939	70 (2, 1, 1)	39.962	0.518
2004*	70 (3, 0, 1)	36.395	0.615	70 (3, 1, 1)	67.892	0.107
2005	40 (5, 0, 0)	15.053	0.952	50 (5, 1, 1)	19.142	0.949
2006	70 (3, 1, 1)	5.787	0.975	70 (4, 1, 1)	13.477	0.956
2007	70 (3, 0, 0)	11.026	0.962	70 (3, 0, 0)	7.770	0.965
2008	60 (4, 1, 1)	60.892	0.209	70 (4, 1, 1)	54.773	0.121
Average	–	29.450	0.724	–	31.748	0.642

* Not included in average

** (3,1,1) indicates $T(t-3)$, $T(t-2)$, $T(t-1)$, $R(t-1)$, $Q(t-1)$

Table 2 Regression parameters of all segments by GMDH method

	Export module	a0	a1	a2	a3	a4	a5
First layer	$Z_9^1[Q(t-1), R(t-1)]$	16.62866	0.53860	−0.00147	0.31608	−0.00088	0.00200
	$Z_{10}^1[T(t-4), T(t-3)]$	3.60373	0.92624	−0.00036	0.50314	0.00275	−0.00903
	$Z_5^1[Q(t-1), T(t-4)]$	0.00000	0.29692	0.00000	1.04566	0.00148	0.01170
	$Z_8^1[Q(t-1), T(t-1)]$	24.90101	0.08174	−0.00022	0.42458	−0.00260	0.00479
Second layer	$Z_2^2(Z_9^1, Z_{10}^1)$	−4.54640	1.64904	−0.00720	−0.86180	0.00757	0.00280
	$Z_3^2(Z_9^1, Z_5^1)$	6.83660	1.20849	0.00150	−0.78898	0.01028	−0.00896
	$Z_4^2(Z_8^1, Z_{10}^1)$	20.20312	0.22685	0.00340	−0.22952	0.00681	−0.00292
Third layer	$Z_4^3(Z_4^2, Z_3^2)$	−4.75467	2.95573	−0.01610	−1.83673	0.00561	0.01000
	$Z_2^3(Z_2^2, Z_3^2)$	20.80679	1.70620	−0.00304	−1.54077	0.01046	−0.00323
Fourth layer	$Z_1^4(Z_2^3, Z_4^3)$	16.08485	−1.98537	0.03366	2.35709	0.00260	−0.03360

Table 3 Calibration RMSE results (NTU) of the best annual forecast model given by trial-and-error

Algorithm	Modeling data length	2000	2001	2002	2003	2005	2006	2007	2008
GMDH	20	–	–	–	112.172	241.589	–	63.565	–
	30	–*	112.723	127.826	–	639.366	158.908	31.363	–
	40	68.251	65.930	56.117	52.822	15.053	9.308	29.677	210.145
	50	–	182.696	83.799	–	17.230	7.715	37.756	330.65
	60	62.713	147.584	49.732	22.750	21.551	6.310	39.408	60.892
	70	57.920	37.574	24.598	36.395	24.455	5.787	11.026	91.777
SGMDH	20	60.927	498.712	90.844	251.413	42.926	111.513	69.136	239.910
	30	–	207.838	248.777	173.750	92.572	47.650	27.351	222.800
	40	64.072	98.749	105.783	72.218	20.264	130.401	26.117	108.854
	50	64.087	761.013	122.730	84.200	19.142	18.948	29.234	77.314
	60	53.575	100.265	73.950	43.982	24.871	13.538	33.044	62.374
	70	52.905	33.149	32.806	39.962	19.738	13.477	7.770	54.773

* Shown RMSE value was divergent

Table 4 The best annual input variables

Modeling event	Input variables	Modeling data length	RMSE (NTU)
2000	$T(t-3)$, $T(t-2)$, $T(t-1)$, $Q(t-1)$	70	57.920
2001	$T(t-2)$, $T(t-1)$, $R(t-1)$	70	37.574
2002	$T(t-3)$, $T(t-2)$, $T(t-1)$, $R(t-1)$, $Q(t-1)$	70	24.598
2003	$T(t-4)$, $T(t-3)$, $T(t-2)$, $T(t-1)$, $R(t-1)$, $Q(t-1)$	60	22.750
2005	$T(t-5)$, $T(t-4)$, $T(t-3)$, $T(t-2)$, $T(t-1)$	40	15.053
2006	$T(t-3)$, $T(t-2)$, $T(t-1)$, $R(t-1)$, $Q(t-1)$	70	5.787
2007	$T(t-3)$, $T(t-2)$, $T(t-1)$	70	11.026
2008	$T(t-4)$, $T(t-3)$, $T(t-2)$, $T(t-1)$, $R(t-1)$, $Q(t-1)$	60	60.892

range of only 10 %. The Taiwan Water Corporation is able to treat water with a turbidity of up to 500 NTU. As such, 50 NTU (10 % of 500 NTU) was chosen as the index of turbidity prediction accuracy. According to standard normal distribution and confidence interval calculations, the results for each year were between 51–66 NTU. An error of only 50 NTU is more restrictive, and was, thus, used as the study threshold.

Verification and analysis of forecast results

Errors between the forecast and actual observations were verified using the best yearly forecast models. Figure 4 indicates that almost all errors were within the 50 NTU threshold. The GMDH turbidity prediction model could forecast levels for up to 10 days, after which the prediction became too uncertain to be trusted for modeling data lengths; so, only 20–70 days were selected.

Best forecast model

The best yearly forecast model could be utilized for the overall turbidity forecasting of other years (e.g., the next occurrence of the same type of storm event). Table 5 presents the RMSE range, revealing that the RMSE in 2006 was comparatively small. Thus, the 2006 model was used to forecast water turbidity in 2007. This gave an RMSE value of 93.137 NTU, a slightly higher error. However, when the 2006 model was used to predict

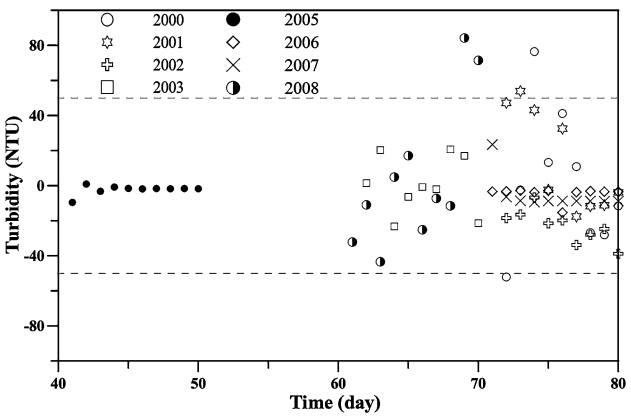

Fig. 4 Predicted errors between observation and forecast values at the Chiahsien Weir

Table 5 Comparison of RMSE (NTU) value of annual forecast

Year	2000	2001	2002	2003	2005	2006	2007	2008
2000	57.920	49.897	62.349	47.164	58.767	43.168	32.435	54.553
2001	32.161	37.574	20.959	22.114	10.090	2.365	19.161	77.564
2002	44.535	38.891	24.598	60.393	28.335	31.021	30.288	31.996
2003	39.588	14.628	27.213	22.749	13.014	14.163	24.257	20.253
2005	71.141	120.520	76.395	116.597	15.053	44.511	61.079	1064.024
2006	41.340	90.357	54.057	95.049	56.392	5.787	60.938	45.879
2007	565.116	114.259	84.570	180.731	42.723	93.137	11.026	72.055
2008	25.018	23.243	20.767	71.414	43.168	31.513	32.836	60.892
Average	109.602	61.171	46.364	77.026	33.442	33.208	34.003	178.402

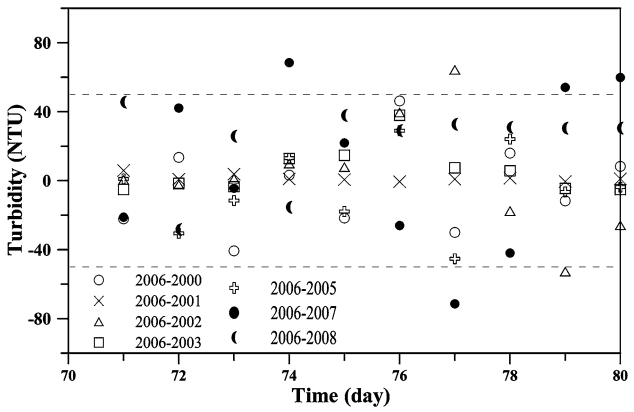

Fig. 5 Predicted errors in the forecasted annual turbidity values using the 2006 model

turbidity for other years, the results were generally within the required confidence interval (Fig. 5). If the predicted values were beyond the error range, a recursive algorithm could be introduced to reduce the prediction error. The 2006 turbidity prediction model was selected as the best predictive model because of its ability to predict turbidity within the acceptable error range and required confidence interval.

Recursive/sequential turbidity forecast model

The recursive/sequential GMDH algorithm incorporates temporal variability once the variance between the predicted and newly observed turbidity exceeds an acceptable range at a certain time. This newly observed value is then added to the model, with previous data being deleted to maintain the same data length. The updated forecasting model then retains its accuracy for later turbidity forecasts. If the updated forecast model does not produce valid output, the steps for adding newly observed values are repeated to enable the system to auto-adjust. Using these procedures, the actual turbidity trend can be observed over any given time period.

In this example, the 2002 model was used to predict the turbidity in 2003, as shown at the top of Fig. 6. Data from the initial 70 days were used to begin model construction, followed by 10-day predictions. However, the results for day 71 already exhibited a large error. Using recursive model calculations, the predicted value for day 71 was discarded, and instead, the measured turbidity value was included. Thus, the original 1st day datum was discarded, the data set was kept at 70 days, and the model was rebuilt to continue 10-day forecast predictions. As shown in the middle of Fig. 6, the predicted values recovered their original accuracy. However, 11 days after the prediction model was rebuilt, the turbidity prediction error for day 81 was excessively large (beyond the acceptable threshold). Thus, GMDH recursive computing was again used to reduce the error. The measured turbidity datum for day 82 was then included for analysis, and its forecast value was discarded. Meanwhile, the original data for the first 12 days were discarded, and the remaining 70 days' data set was utilized for model reconstruction. Again, a recursive structure was applied. As illustrated in the bottom part of Fig. 6, the accuracy of the predicted values was maintained. In principle, using the 10-day prediction as a guide, when the prediction error was within the acceptable range, the forecasts could continue. Whenever an updated turbidity datum was added for recursive calculation, the RMSE value of the prior model decreased, meaning that the deletion of the earliest old datum is more significant than adding the updated one, i.e., no specific recursive computing procedure should be carried out in this step.

Conclusion

Turbidity is the most important index for public water supply. High turbidity inflow causes harassment on treatment of public water supply, even bringing the need to cut off the water supply. To avoid high turbidity water inflow, it is important to strengthen the catchment's conservation,

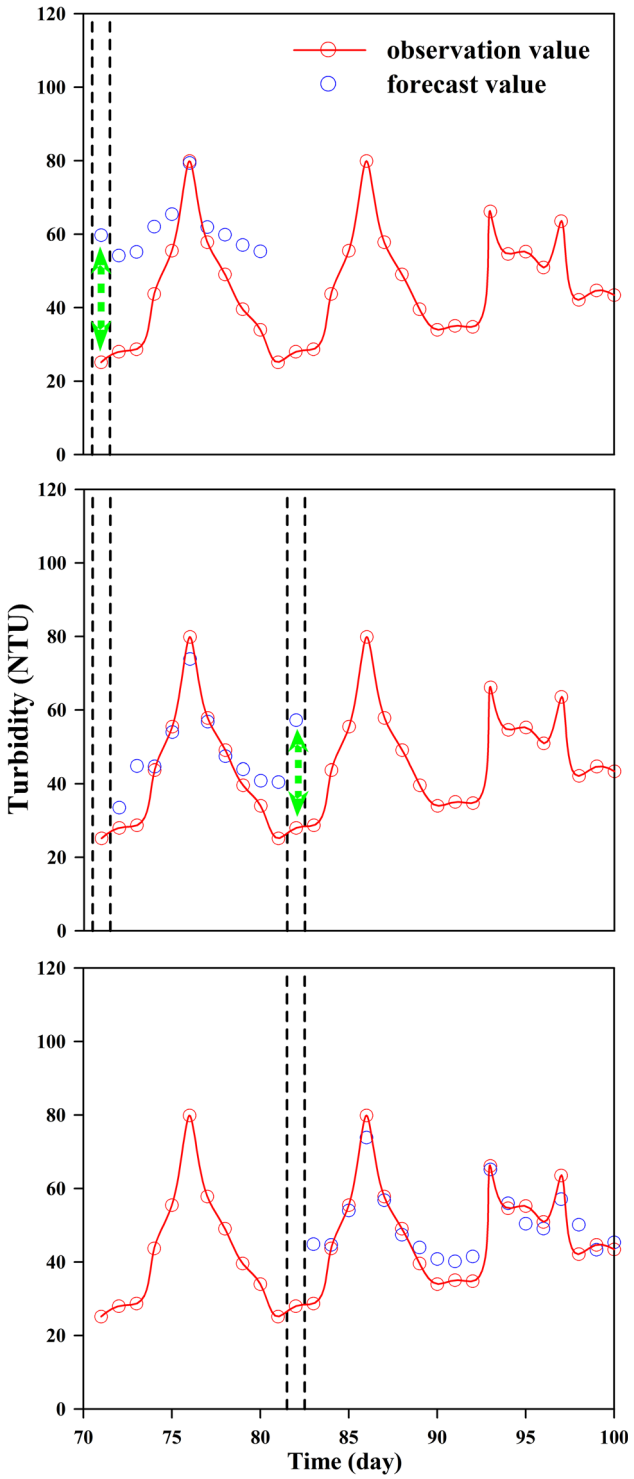

Fig. 6 Forecasted 2003 turbidity values using the regressive 2002 model

protect the water resources territory, and predict the inflow turbidity concentration before the treatment operation.

A local historical turbidity, rainfall, and discharge database was constructed to develop a turbidity prediction model based on the GMDH algorithm. The results from a cross-validation revealed that GMDH was more appropriate than SGMDH for this case study. The majority of predictive turbidity values were within a confidence interval of 90 % or approaching 90 %. Using the recursive GMDH algorithm, the model can be modified to generate better predictions and improve forecast accuracy. The test results indicate that this turbidity prediction model is feasible and reliable for turbidity forecasting. Even with complex environmental factors, the model remains applicable.

Acknowledgments We deeply appreciate the assistance of the Taiwan Water Corporation, which provided us with data for hydrological findings, as well as the generous aid from its engineer Mr. Wang Ying-Ming in finishing our research. In addition, the authors are also indebted to reviewers for their valuable comments and suggestions.

References

Chang FJ, Hwang YY (1999) A self-organization algorithm for real-time flood forecast. Hydrol Process 13(2):123–138

Farlow SJ (1984) Self-organizing methods in modeling: GMDH-type algorithms. Marcel Dekker, New York

Hwang SL, Liang GF, Lin JT, Yau YJ, Yenn TC, Hsu CC, Chuang CF (2009) A real-time warning model for teamwork performance and system safety in nuclear power plants. Saf Sci 47(3):425–435

Ikeda S, Fugishige S, Sawaragi Y (1976) Nonlinear prediction model of river flow by self-organization method. Int J Syst Sci 7(2):165–176

Ivakhnenko AG (1968) Group method of data handling-rival of method of stochastic approximation. Sov Autom Control 13(1):43–55

Ivakhnenko AG (1971) Polynomial theory of complex systems. IEEE Trans Syst Man Cybern SMC 1(4):364–378

Ivakhnenko AG, Ivakhnenko GA, Muller JA (1994) Self-organization of the neural networks with active neurons. Pattern Recognit Image Anal 4(2):177–188

Kondo T, Pandya AS, Zurada JM (1999) GMDH-type neural networks and their application to the medical image recognition of the lungs. In: Proceedings of the 38th SICE Annual Conference, School of Medical Science, Tokushima University 1181–1186

Lebow WM, Mehra RK, Rice H, Tolgalagi PM (1984) Forecasting applications in agricultural and meteorological time series. In: Farrow SJ (ed) Self-organizing methods in modeling: GMDH type algorithms. Marcel Dekker, New York, pp 121–147

Madala HR, Ivakhnenko AG (1994) Inductive learning algorithms for complex systems modeling. CRC Press Inc, Boca Raton

Najafzadeh M (2015) Neurofuzzy-based GMDH-PSO to predict maximum scour depth at equilibrium at culvert outlets. J Pipeline Syst Eng Pract 7(1):06015001

Najafzadeh M, Barani GA, Hazi MA (2013) GMDH to predict scour depth around a pier in cohesive soils. Appl Ocean Res 40:35–41

Najafzadeh M, Barani GA, Hessami Kermani MR (2014) Group method of data handling to predict scour at downstream of a ski-jump bucket spillway. Earth Sci Inform 7(4):231–248

Najafzadeh M, Barani GA, Hessami-Kermani MR (2015) Evaluation of GMDH networks for prediction of local scour depth at bridge abutments in coarse sediments with thinly armored beds. Ocean Eng 104:387–396

Pavel N, Miroslav S (2003) Modeling of student's quality by means of GMDH algorithms. Syst Anal Model Simul (SAMS) 43(10):1415–1426

Sarycheva L (2003) Using GMDH in ecological and socio-economical monitoring problems. Syst Anal Model Simul (SAMS) 43(10):1409–1414

Tamura H, Kondo T (1980) Heuristics free group method data handling algorithm of generating optimal partial polynomials with application to air pollution prediction. Int J Syst Sci 11(9):1095–1111

Tsai TM, Yen PH, Huang TJ (2009) Wave height forecasting using self-organization algorithm model. In: Proceedings of the Nineteenth (2009) International Offshore and Polar Engineering Conference Osaka, Japan, pp 806–812

Yoshimura T, Kiyozumi R, Nishino K, Soeda T (1982) Prediction of air pollutant concentrations by revised GMDH algorithms in Tokushima Prefecture. IEEE Trans Syst Man Cybern SMC 12(1):50–56