http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet

Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (15): 3158-3163, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.3158.3163

A Combination Model of Chaos, Wavelet and Support Vector Machine Predicting Groundwater Levels and its Evaluation Using Three Comprehensive Quantifying Techniques

Jianhua Ping, Yu Qiang and Ma Xixia
School of Water Conservancy and Environment Engineering,Zhengzhou University,
Zhengzhou, Henan, 450001, China

Absract: Groundwater levels prediction is very important to groundwater resources evaluation and management. A combined model of chaos theory, wavelet and support vector machine was develop to overcome the limitations including challenges in determination of orders of nonlinear models and low prediction accuracy which the simulated accuracy is high in groundwater levels foresting. Firstly, groundwater level series were decomposed into different frequency components in application of wavelet analysis. Secondly, phase space was reconstructed using chaotic analysis. Thirdly, support vector machine (SVM) was used to predict each component. Finally, all components were merged into a model to predict groundwater levels. A case study, annual groundwater levels located in the Spallumcheen B aquifer situated in the Fortune Creek watershed surrounded by mountains in semi-arid areas within west interior British Columbia, Canada was employed to examine the combined model. The integrated model was evaluated by qualitative graphical method and three quantitative approaches comprising of NSE, PIBAS and RSR techniques. These evaluation values indicated the combined model high accuracy in groundwater level prediction and the model was valuable and useful for groundwater level forecasting.

Key words: Groundwater level prediction, support vector machine, chaotic analysis, wavelet analysis, combination model, model evaluation

INTRODUCTION

Water is the basic natural resources and strategic economical resources. Groundwater is the important part of water resources. Groundwater is an important part of water resources. Groundwater is the important water supply source of many cities in north China and it plays an important role to the development of economy and society (Currell *et al*., 2012). In the recent 30 years, with the rapid development of economy and society, the exploitation of groundwater are increasing greatly, which induces a series of environmental problems and forms serious intimidate to the sustainable development of social economy in local area (Zhang *et al*., 2008). Groundwater levels regime research is one of the major scientific studies in groundwater discipline and research. Groundwater levels fluctuation indicates that groundwater resource is changing. The immediate reflection of groundwater exploitation is groundwater level decreasing. Groundwater level regime is a complicated nonlinear process. Various methods and models are employed to predict groundwater level.Groundwater level forecasting research is grouped into deterministic and stochastic models in mathematic models. The solving approaches to deterministic models include analytic method, physical simulation and numercial simulation. The stochastics models comprises of regression analysis models, gray models, frequency analysis models, time series models and ect (Maheswaran and Khosa, 2013). Since the 1990's, artificial Neural Network (ANN) (Yoon *et al*., 2011), support Vector Machine (SVM) (Yoon *et al*., 2011), chaotic phase space reconstruction (Trefry *et al*., 2012) and wavelet anaylsis (Adamowski and Chan, 2011) have been used in groundwater level prediction research. Groundwater level forecasting is very complicated only becasue groundwater system is complex, nonlinear, multi-scales and stochastic charcateristics. No single forecating method is applicabe for all kinds groundwater system bceasue the climate and underlying surface varies in differect hydrogeologic units. Chaos theory, wavelet analysis and support vector machine were combined into a forecasting model to predict annual groundwaer levels in this study. First, the complicated nonlinear ground water levels series was decomposed into

Corresponding Author: Jianhua Ping, School of Water Conservancy and Environment Engineering, Zhengzhou University, Zhengzhou, Henan, 450001, China

supperposition of some layers of simple sequence. Second, Phase space was reconstructed based on aturated embedding dimension and built-in dealy time of the series obtained in chaotic analysis. Third, support vector machine with better generalization ability than other nonlinear theory was used to predicted each component. Finally, all components were combined into the predict results.

METHODOLOGY

Phase space reconstruction: Phase space reconstruction is the basis of analysis and prediction in application of Chaotic theory because chaotic verification of groundwater levels series and chaotic analysis should be conducted in phase space reoconstruction of ground water level series (Sekar and Randhir, 2007). τ (Built-in delay time) and m (embedding dimension) are two important paramenters of phase space reconstruction. Given groundwater levels series, $x = \{x_i/i = 1,2,\ldots,N\}$, where, y_i denotes the phase points in m-dimensional phase space, M is the number of phase points, M = N-(m-1) J, So the reconstructed phase space, $i = 1,2,\ldots,M$.

Autorelativity function method and mutual information function method are two approaches to determine the embedding delay time τ. Autorelativity function method is simple, but it is not applicable to all cases mainly because it only describes the degree of linear correction between variables. Mutual information function method utilized in non-linear relationship analysis, but it is unable to avoid numerious calculation and complex space division due to its complex algorithm. False Neighbor method, saturated correlative dimension methoda and C-C method were often used to determine embedding dimension (m) in recent studies.

C-C method (Kim *et al*., 1999) is an effective method for calculating the time delay and the time window of phase space reconstruction. First, it build the statistic via Cross-Correlation Integral of embedded time series. Second, it determines the optimal delay time and embedding window. Third, it determines the embedding dimension with embedding window. Cross-Correlation method was employed to determine τ (Built-in delay time) and m (embedding dimension). The steps of C-C method is described as following:

- Calculating stand deviation (σ) of the given time series, selecting the appropriate sequence length N
- Calculating the following statistics: $\bar{S}(t)$, $\Delta\bar{S}(t)$, $S_{cor}(t)$
- Determining the emberdding delay time τ and delay time window τ_w based on the following plots
- The first zero t of S(t) corresponds to an embedding delay time
- The first minimum t of 6S(t)corresponds to an embedding delay time τ
- The minimum t of $S_{cor}(t)$ corresponds to an delay time bandwidth τ_w
- Calculating embedding dimension

$$m = \mathrm{int}\left(\frac{\tau_w}{\tau} + 1\right) \tag{1}$$

Chaotic identification: Chaotic identification methods consist of qualitative appoaches including Phase Diagram method and power spectrum method and quantitative approaches comprising saturated correlation dimension method and Lyapunov exponent method which is commonly used. Wolf method and small data sets are two main approaches in Lyapunov determination. Wolf method requires time series wihtout noise, the data requiremenined, which takes a long evolutionts are relatively high,and the characteristic parameters of the system is obta time to track .Small data sets is applicable for small sample because its calculation is small and easy to implement based on improvement of Wolf algorithm. Small data sets was employed to figure out the maximum Lyapunov exponent (Liu and Liew, 2003; Ubeyli, 2010). The procuders is presented as following.

Select all of M = N-(m-1) phase point as a reference point,reference phaes point and the nearest phase points in the phase space as a starting point of adjacent track to study the exponential separation of adjacent track.

Orbital distance is the initial distance at the moment of the subscript i:

$$\delta_0^i = \|y_i - y_{ir}\| = \frac{1}{m}\sqrt{\sum_{k=1}^{m}(x_{i-(k-1)\tau} - x_{ir-(k-1)\tau})^2} \tag{2}$$

The orbital separation of chaotic systems with exponential separation characteristics, therefore:

$$\delta_s = \delta_0 e^{\lambda s} \tag{3}$$

Maximum Lyapunov index is adjusted according to the following formula:

$$\lambda = \frac{\ln(\delta_s / \delta_0)}{s} = \frac{\ln\delta_s}{s} - \frac{\ln\delta_0}{s} \tag{4}$$

Taking the overall average from Separation distance which M phase point and near point of its evolutings steps:

$$\bar{\delta}_s = \frac{1}{M}\sum_{i=1}^{M}\delta_s^i \tag{5}$$

Drawing in δ_s~s graph, selecting the linear part of curve to fit a straight line,the slope is the global maximum Lyapunov index.

Support vector machine: Given training data $\{(x_1,y_1), (x_2,y_2), \ldots.., (x_l,y_l)\} \in R_n \times R$, where, x_i denotes input vector, y_i is the output vector of corresponding x_i and l is the number of samples.the basic idea of support vector regression is that the data x_i is mapped to high dimensional feature space F by a nonlinear mapping Φ and carrying out linear regression in this space, that is:

$$f(x = \omega^T\Phi(x)+b) \tag{6}$$

where, w denotes the weigt vector of hyperplane,b is the bias. Solving optimization problems support vector regression in condition Eq. 7:

$$\begin{cases} y_i - [\omega^T\Phi(x_i)+b] \le \varepsilon + \xi_i \\ [\omega^T\Phi(x_i)+b] - y_i \le \varepsilon + \xi_i^* \\ \xi i, \xi_i^* \ge 0, \quad i = 1, 2, \ldots, 1 \end{cases} \tag{7}$$

$$\min_{\omega, b, \xi_i, \xi_i^*} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{1}(\xi_i + \xi_i^*) \tag{8}$$

where, ξ_i, ξ_i(is the slack varibale,denotes the upper and lower limits of the training error in the error ε condition $(|yi-[\omega^T\Phi(x_i)+b]|)$, respectively; ε is the error which is defined by Vapnik-ε not sensitive cost function (David and Sanchez, 2003). Constant C>0, it control the degree of punishment of the sample beyond error.

Wavelet analysis method: Wavelet analysis can be utilizted to decompose complex time series into a couple of detail signal sequence describing the high frequency component and background sigal presenting low frequency components. Annual groundwater levels series are decomposed and reconstructed in application of wavelet should be utilized in discrete wavelet transform because it is discrete series (Sang, 2013).

Multi-resolution analysis (MRA) is commonly used discrete wavelet transform algorithm, which decomposes groundwater levels series into high frequency component and low frequency part and it can be repeated to decompose any resolution of high frequency and low frequency components (Labat, 2008).

In the non-stationary groundwater levels series,the trend component can be seen as long-cycle component which cycle length much longer than the actual series; Random component is caused by the irregular oscillations and random factors,they are small scale high frequency; periodic component is caused by the certainty ,on the spectrum of the groundwater levels ,the frequency range between the trend component and the random component (Karthikeyan and Nagesh , 2013).

Therefore, the analysis of low frequency components and high frequency components were able to identify variations of time series and decompose groundwater levels series into trend term, periodicity term and randomness term.

COMBINATION MODEL

Basic theory: Firstly, the groundwater levels series $\{x_1, x_{2,\ldots,} x_N\}$ are judged if they have chaotic characteristic and then the delay time (embedding residence time) t and the best embedding dimension m are determined using chaotic analysis method. Secondly, the ruoff series are decomposed into background signal and detail signal CDt (t = 1,2, …, J, J is decomposed scales number) which have dimensionality reduction characteristics. Thirdly, At and Dt are obtained from reconstruction of backgroundwater singal CA and detail signal CDt, in which the first n data are employed to train predictor to conduct parameter estimation and topological structure determination of predictor and the rest m of data is used to validate the model. Finally, At and Dt are predicted by SVM predictor separately.

Step (1): Normalization of groundwater levels series $\{x_1, x_{2,\ldots,} x_N\}$ via Eq. 9:

$$x_t = \frac{x_t - \bar{x}}{\bar{\sigma}} \tag{9}$$

where, $\bar{x}$ is groundwater levels' mean value, σ is groundwater levels' mean square error.

Step (2): C-C method is employed to determine the bulit-in delay time (τ) and embedding dimension (m)

Step (3): Groundwater levels series is identified chaotic characteristics

Step (1): The wavelet decomposition series {cAt, cD1, cD2,…cDt} which are decomposed by the db3 function are separately reconstructed into the series {At, D1, D2,…Dt} based on the Mallat decomposition method If the groundwater levels series $\{x_1, x_{2,\ldots,} x_N\}$ are chaotic series. Training sequence and verification sequence are formed based on the relationship of input X and output Y using built-in delay time τ and embedding dimension m for each series At, D1, D2, ... Dt

$$X=\begin{bmatrix} x_1 & x_{1+t} & \cdots & x_{1+(m-1)t} \\ x_2 & x_{2+t} & \cdots & x_{2+(m-1)t} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n-1-(m-1)t} & x_{n-1-(m-1)t+t} & \cdots & x_{n-1} \end{bmatrix} Y=\begin{bmatrix} x_{2+(m-1)t} \\ x_{3+(m-1)t} \\ \vdots \\ x_n \end{bmatrix} \quad (10)$$

Step (5): First, regression support vector machine prediction model is established separately based on series At, D1, D2,... Dt and then the mapping relationship of input and output is obtained via training learning sample and determining model parameters. Second, The predictive value of each sequence is obtained via putting verification sequence into each prediction model. Finally, the combination of predictive value are forecast results

Step (6): Inversing normalization is conducted for the predicted results

Step (7): Model calibration and validation

Various approaches including qualitative graphic methods and quantitative statistical methods can be employed to calibrate and validate the model (Moriasi *et al*., 2007).

The quantitative methods comprise of standard regression statistics methods, nondimensional methods and error exponential methods. The graphical method provide a visual comparison of computed and observed data and a first overall view of model performance. Coefficient of determination included in standard regression statistics methods describing the proportion of the variance in observed data explained by the model, Nash-Sutcliffe efficiency (NSE) (Gupta *et al*., 1999) which is a normalized statistic that determines the relative magnitude of the residual variance compared to the observed data variance and suggests how well the plot of observed versus computed data fits the 1:1 line included in dimensionless techniques, Percent bias (PBIAS) (Legates and McCabe, 1999) included in error index techniques measuring the average tendency of the modeled data to be larger or smaller than their measured counterparts, RMSE-observations standard deviation ratio (RSR) (Singh *et al*., 2005) incorporating the benefits of error index statistics and including a scaling/ normalization factor were employed to evaluate the combination model.

NSE is estimated as shown in Eq. 11:

$$\text{NSE}=1-\left[\frac{\sum_{i=1}^{n}\left(X_i^{obs}-X_i^{com}\right)^2}{\sum_{i=1}^{n}\left(X_i^{obs}-X^{mean}\right)^2}\right] \quad (11)$$

where, X_i^{obs} means the ith observation for the constituent being evaluated, X_i^{com} is the ith computed value for the constituent being evaluated, X^{mean} is the mean of measured data for the constituent being evaluated and n represents the total number of measurements. NSE covers a range from -8 to 1, with NSE =1 being the optimal value. The values are less than zero suggests unacceptable model performance.

PBIAS is obtained with Eq. 12:

$$\text{PBIAS}=\left[\frac{\sum_{i=1}^{n}\left(X_i^{obs}-X_i^{com}\right)\times 100}{\sum_{i=1}^{n}\left(X_i^{obs}\right)}\right] \quad (12)$$

The lower magnitude values of PBIAS indicates the more accurate model modeling. Zero is the optimal value. Negative values indicate model overestimation bias and positive values indicate model underestimation bias. RSR is calculated by Eq. 13:

$$\text{RSR}=\frac{\text{RMSE}}{\text{STDEV}_{obs}}=\frac{\left[\sqrt{\sum_{i=1}^{n}\left(X_i^{obs}-X_i^{com}\right)^2}\right]}{\left[\sqrt{\sum_{i=1}^{n}\left(X_i^{obs}-X^{mean}\right)^2}\right]} \quad (13)$$

where, STDEV_{obs} means standard deviation of observation data. RSR ranges from the optimal value of 0, which suggests zero RMSE or residual variation and therefore perfect model simulation, to a large positive magnitude. The higher RSR, the higher RMSE and the worse the model simulation performance.

CASE STUDY

Study area description: The Okanagan valley is situated in the semi-arid southern interior of British Columbia (BC). Due to population growth, increasing water demand and the possible impacts of climate change, there are growing concerns over shortages of water resources to meet the needs of future economic and social development in the valley. Concerns have been raised over the quantity and quality of shallow and moderate groundwater resources and how groundwater pumping may affect surface stream flows and deep groundwater. Shallow and moderate Groundwater pumping is the major resource for the water supply in Fortune Creek watershed located in the Northern Okanagan. There is an artisan well drilled in the regional deep aquifer Spallumcheen B represents the deep groundwater level monitored by BC Ministry of Environment since 1971. The groundwater levels fluctuations of the artisan well is important to understand the deep groundwater resources variation. Prediction of groundwater level of the artesian well is valuable to the water supply in Fortune Creek watershed and the development of the society. The historical annual groundwater level series between 1971-2000 was employed to develop and calibrate the combination model and determine the model parameter, while the groundwater level series data between 2001 to 2008 was utilized to validate the model forecasting accuracy.

RESULTS AND DISCUSSION

Qualitative graphical modeling results during calibration Fig. 1 and validation Fig. 2 indicated adequate calibration and validation over the range of groundwater level variation, although the computed results in calibration showed a better match than the simulated results in validation. The trends of groundwater level variation of simulated and observed groundwater levels in calibration and validation were similar. The results of four quantitative statistic techniques for the combined model evaluation in calibration and validation were presented in Table 1.

NSE values being 0.88 and 0.77 which were nearly 1.0 in model calibration and validation respectively indicated that the model was close to optimal and the model performance in calibration was better than what in validation. The model calibration and validation PBIAS values being 0.19 and 0.26 suggested that the combined model was near accurate and the underestimation bias. RSR values being 0.15 and 0.29 which were near the optimal value of 0 indicated that the combined model performance was closely perfect. The values of PBIAS and RSR in validation being larger than

Table 1: NES, PBIAS and RSR values of the combined model for the case study

Item	Calibration	Validation
NSE	0.88	0.77
PBIAS	0.19	0.26
RSR	0.15	0.29

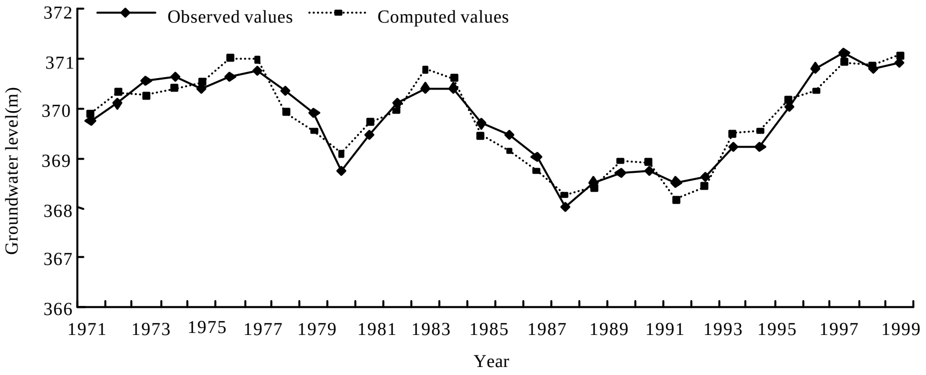

Fig. 1: Comparison of observed and computed groundwater levels in calibration

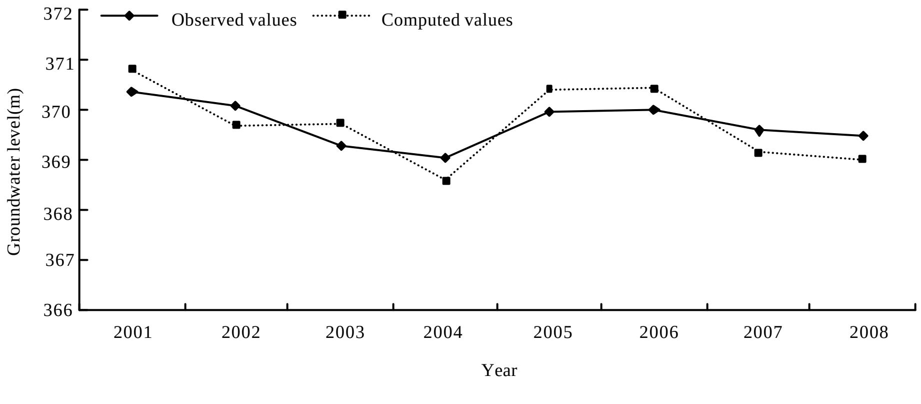

Fig. 2: Comparison of observed and computed groundwater levels in validation

what in calibration indicated the model performance in validation working less than what in calibration.

CONCLUSION

The combined model integrated chaotic theory and wavelet analysis with support vector machine fully applying phase space reconstruction of chaos theory, multi-resolution ability of wavelet and ability of nonliner approach of support vector machine, which was built using annual groundwater levels located in the deep aquifer named Spallumcheen B in Fortune Creek watershed situated in British Columbia, Canada, overcoming the challenges of the nonlinear model exponent number determination, prediction accuracy low with high accuracy of model simulated in groundwater level forecasting in this paper is applicable, useful and applicable for other time series prediction.

ACKNOWLEDGEMENTS

The authors would like to thank for the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, China. The authors also are grateful for the work supported by Supported by Canadian Water Network and Natural Sciences and Engineering Research Council (NSERC) of Canada Foundations The authors are grateful for the study supported by the Talent Project of Zhengzhou University as well.

REFERENCES

Adamowski, J. and H.F. Chan, 2011. A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol., 407: 28-40.

Currell, M.J., D. Han, Z. Chen and I. Cartwright, 2012. Sustainability of groundwater usage in northern China: Dependence on palaeowaters and effects on water quality, quantity and ecosystem health. Hydrol. Proc., 26: 4050-4066.

Gupta, H., S. Sorooshian and P. Yapo, 1999. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. J. Hydrol. Eng., 4: 135-143.

Karthikeyan, L. and D.N. Kumar, 2013. Predictability of nonstationary time series using wavelet and EMD based ARMA models. J. Hydrol., 502: 103-119.

Kim, H.S., R. Eykholt and J.D. Salas, 1999. Nonlinear dynamics, delay times and embedding windows. Physica D: Nonlinear Phenomena, 127: 48-60.

Labat, D., 2008. Wavelet analysis of the annual discharge records of the world's largest rivers. Adv. Water Resources, 31: 109-117.

Legates, D.R. and G.J. McCabe, 1999. Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. Water Resources Res., 35: 233-241.

Liu, X.B. and K.M. Liew, 2003. The Lyapunov exponent for a codimension two bifurcation system that is driven by a real noise. Int. J. Non-Linear Mech., 38: 1495-1511.

Maheswaran, R. and R. Khosa, 2013. Long term forecasting of groundwater levels with evidence of non-stationary and nonlinear characteristics. Comput. Geosci., 52: 422-436.

Moriasi, D.N., J.G. Arnold, M.W. VanLiew, R.L. Bingner, R.D. Harmel and T.L. Veith, 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE, 50: 885-900.

David, V. and A. Sanchez, 2003. Advanced support vector machines and kernel methods. Neurocomputing, 55: 5-20.

Sang, Y.F., 2013. A review on the applications of wavelet transform in hydrology time series analysis. Atmospheric Res., 122: 8-15.

Sekar, I. and T.O. Randhir, 2007. Spatial assessment of conjunctive water harvesting potential in watershed system. J. Hydrol., 334: 39-52.

Singh, J., H.V. Knapp, J.G. Arnold and M. Demissie, 2005. Hydrologic modeling of the Iroquois river watershed using HSPF and SWAT. J. Am. Water Resources Assoc., 41: 343-360.

Trefry, M.G., D.R. Lester, G. Metcalfe, A. Ord and K. Regenauer-Lieb, 2012. Toward enhanced subsurface intervention methods using chaotic advection. J. Contaminant Hydrol., 127: 15-29.

Ubeyli, E.D., 2010. Lyapunov exponents/probabilistic neural networks for a nalysis of EEG signals. Exp. Syst. Appl., 37: 985-992.

Yoon, H., S.C. Jun, Y. Hyun, G.O. Bae and K.K. Lee, 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J. Hydrol., 396: 128-138.

Zhang, L., J. Wang, J. Huang and S. RoZelle, 2008. Development of groundwater markets in China: A glimpse into progress to date. World Dev., 36: 707-726.