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

Abstract: 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 forecasting. 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.

Keywords: 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 numerical simulation. The stochastics models comprises of regression analysis models, gray models, frequency analysis models, time series models and etc (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 analysis (Adamowski and Chan, 2011) have been used in groundwater level prediction research. Groundwater level forecasting is very complicated only because groundwater system is complex, nonlinear, multi-scales and stochastic characteristics. No single forecasting method is applicaba for all kinds groundwater system because the climate and underlying surface varies in different hydrogeologic units. Chaos theory, wavelet analysis and support vector machine were combined into a forecasting model to predict annual groundwater 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 reconstruction of ground water level series (Sekar and Randhir, 2007). τ (Built-in delay time) and m (embedding dimension) are two important parameters of phase space reconstruction. Given groundwater levels series, \( x = \{x_{i}, i = 1,2,...,N\} \), where, \( x_{i} \) denotes the phase points in m-dimensional phase space, \( M \) is the number of phase points, \( M = N-(m-1) \) I, So the reconstructed phase space, \( i = 1,2,...,M \).

Autocorrelativity function method and mutual
information function method are two approaches to determine the embedding delay time \( \tau \). Autocorrelativity 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 numerous calculation and complex space division due to its complex algorithm. False Neibor method, saturated correlational 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 \( \tau \) (Built-in delay time) and m (embedding dimension). The steps of C-C method is described as following:

- Calculating stand deviation (\( \sigma \)) of the given time
  series, selecting the appropriate sequence length \( N \)
- Calculating the following statistics: \( S(t), \Delta S(t), S_m(t) \)
- Determining the embedding delay time \( \tau \) and delay
time window \( \tau_w \) 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 \( \tau \)
- The minimum t of \( S_m(t) \) corresponds to a delay
time bandwidth \( \tau_w \)
- Calculating embedding dimension

\[
m - \text{int}\left(\frac{\tau}{\tau} + 1\right)
\]  

(1)

Chaotic identification: Chaotic identification methods consist of qualitative approaches 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's method and small data sets are two main approaches in Lyapunov determination. Wolf's method requires time series without noise, the data requirement which takes a long evolution can be relatively high, and the characteristic parameters of the system is obtainable to track. Small data sets are applicable for small sample because its calculation is small and easy to implement based on improvement of Wolf's algorithm. Small data sets was employed to figure out the maximum Lyapunov exponent (Liu and Liew, 2003; Ubeyli, 2010). The procedures is presented as follows.

Select all of \( M = N-(m-1) \) phase point as a reference point, reference phase 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 1:

\[
\delta_1 = \|y_1 - y_0\| = \frac{1}{m} \sum_{i=0}^{m} \left( x_{i+1} - x_{i}\right) \left( x_{i+1} - x_{i}\right)
\]

(2)

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

\[
\delta_s = \delta_1 e^{\lambda t}
\]

(3)

Maximum Lyapunov index is adjusted according to the following formula:

\[
\lambda = \frac{\ln(\delta_s/\delta_1)}{s} = \frac{\ln\delta_s - \ln\delta_1}{s}
\]

(4)

Taking the overall average from Separation distance which M phase point and near point of its evolutions steps:
\[ \delta_t = \frac{1}{M} \sum_{i} \delta_i \]  

(5)

Drawing on the \( \delta_t \)-s graph, selecting the linear part of the curve to fit a straight line, the slope is the global maximum Lyapunov index.

**Support vector machine:** Given training data \{ \( (x_i,y_i) \), \( (x_2,y_2) \), ..., \( (x_n,y_n) \) \} \( \in \mathbb{R}^n \times \mathbb{R} \), where \( x_i \) denotes input vector, \( y_i \) is the output vector of corresponding \( x_i \), and \( M \) 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 \( \Phi \) and carrying out linear regression in this space, that is:

\[ f(x) = w^T \Phi(x) + b \]  

(6)

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

\[
\begin{align*}
    \{y_i - [\omega^T \Phi(x_i) + b] \leq \varepsilon + \xi_i \\
    \{\omega^T \Phi(x_i) + b\} - y_i \leq \varepsilon + \xi_i' \\
    \xi_i, \xi_i' \geq 0, \quad i = 1,2,...,M
\end{align*}
\]  

(7)

\[
\min_{\omega, b, \xi_i, \xi_i'} \frac{1}{2} \omega^T \omega + C \sum_{i} (\xi_i + \xi_i')
\]  

(8)

where, \( \xi_i, \xi_i' \) is the slack variable, denotes the upper and lower limits of the training error in the error \( \varepsilon \) condition \( \{|y_i - [\omega^T \Phi(x_i) + b]|\} \), respectively, \( \varepsilon \) is the error which is defined by Vapnik-\( \varepsilon \) not sensitive cost function (David and Sanchez, 2003). Constant \( C \geq 0 \), it control the degree of punishment of the sample beyond error.

**Wavelet analysis method:** Wavelet analysis can be utilized to decompose complex time series into a couple of detail signal sequence describing the high frequency component and background signal 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 \{ \( x_1, x_2, ..., x_n \) \} 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 rawoff series are decomposed into background signal and detail signal \( CD_t (t = 1,2, ..., J, J \) is decomposed scales number) which have dimensionality reduction characteristics. Thirdly, \( CDT \) and \( DT \) are obtained from reconstruction of background signal \( CA \) and detail signal \( CD_t \), 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, \( CDT \) and \( DT \) are predicted by SVM predictor separately.

**Step (1):** Normalization of groundwater levels series \{ \( x_1, x_2, ..., x_n \) \} via Eq. 9:

\[ x_i = \frac{x_i - \bar{x}}{\sigma} \]

(9)

where, \( \bar{x} \) is groundwater levels’ mean value, \( \sigma \) is groundwater levels’ mean square error.

**Step (2):** C-C method is employed to determine the built-in delay time \( \tau \) and embedding dimension \( m \)

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

**Step (1):** The wavelet decomposition series \{ \( cDA_1, cDA_2, ..., cDA_T \) \} which are decomposed by the db3 function are separately reconstructed into the series \{ \( A_1, D_1, D_2, ..., D_T \) \} based on the Mallat decomposition method. If the groundwater levels series \{ \( x_1, x_2, ..., x_n \) \} 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 \( \tau \) and embedding dimension \( m \) for each series \( A_t, D_1, D_2, \ldots, D_t \)

\[
X = \begin{bmatrix}
  x_1 & x_{1+m} & \cdots & x_{1+(m-1)} \\
  x_2 & x_{2+m} & \cdots & x_{2+(m-1)} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n-(m-1)} & x_{n-(m-1)+m} & \cdots & x_n
\end{bmatrix}
\]

\[Y = \begin{bmatrix}
  x_{1+n-m} \\
  x_{1+n-m+1} \\
  \vdots \\
  x_{n}
\end{bmatrix}
\]  \( (10) \)

**Step (5):** First, regression support vector machine prediction model is established separately based on series \( A_t, D_1, D_2, \ldots, D_t \) 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:

\[
\text{NSE} = 1 - \frac{\sum_{i=1}^{n} (x_i^{\text{obs}} - x_i^{\text{sim}})^2}{\sum_{i=1}^{n} (x_i^{\text{obs}} - \bar{x}_i)^2}
\]  \( (11) \)

where, \( x_i^{\text{obs}} \) means the ith observation for the constituent being evaluated, \( x_i^{\text{sim}} \) is the ith computed value for the constituent being evaluated, \( \bar{x}_i \) 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:

\[
\text{PBIAS} = \frac{\sum_{i=1}^{n} (x_i^{\text{obs}} - x_i^{\text{sim}}) \times 100}{\sum_{i=1}^{n} (x_i^{\text{sim}})}
\]  \( (12) \)

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:

\[
\text{RSR} = \frac{\text{RMSE}}{\text{STDEV}_{\text{obs}}} = \frac{\sqrt{\frac{\sum_{i=1}^{n} (x_i^{\text{obs}} - x_i^{\text{sim}})^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n} (x_i^{\text{obs}} - \bar{x}_i)^2}{n}}}
\]  \( (13) \)

where, \( \text{STDEV}_{\text{obs}} \) 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>
<thead>
<tr>
<th>Item</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSE</td>
<td>0.88</td>
<td>0.77</td>
</tr>
<tr>
<td>PBIAS</td>
<td>0.19</td>
<td>0.26</td>
</tr>
<tr>
<td>RSR</td>
<td>0.15</td>
<td>0.29</td>
</tr>
</tbody>
</table>

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

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

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

3162
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


