Hydrol. Earth Syst. Sci. Discuss., 7, 3691–3731, 2010
www.hydrol-earth-syst-sci-discuss.net/7/3691/2010/
doi:10.5194/hessd-7-3691-2010

Hydrology and Earth System Sciences Discussions

A hybrid least squares support vector machines and GMDH approach for river flow forecasting

R. Samsudin[1], **P. Saad**[1], **and A. Shabri**[2]

[1]Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia
[2]Faculty of Science, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia

Received: 27 April 2010 – Accepted: 3 May 2010 – Published: 23 June 2010

Correspondence to: R. Samsudin (ruhaidah@utm.my)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

This paper proposes a novel hybrid forecasting model, which combines the group method of data handling (GMDH) and the least squares support vector machine (LSSVM), known as GLSSVM. The GMDH is used to determine the useful input variables for LSSVM model and the LSSVM model which works as time series forecasting. In this study the application of GLSSVM for monthly river flow forecasting of Selangor and Bernam River are investigated. The results of the proposed GLSSVM approach are compared with the conventional artificial neural network (ANN) models, Autoregressive Integrated Moving Average (ARIMA) model, GMDH and LSSVM models using the long term observations of monthly river flow discharge. The standard statistical, the root mean square error (RMSE) and coefficient of correlation (R) are employed to evaluate the performance of various models developed. Experiment result indicates that the hybrid model was powerful tools to model discharge time series and can be applied successfully in complex hydrological modeling.

1 Introduction

River flow forecasting is an essential procedure that is necessary for proper reservoir system controls and successful planning and management of water resources. Accurate forecasting of river flow has been one of the most important issues in hydrological research. Due to river flow forecasting involves a rather complex nonlinear and chaotic data pattern; several techniques have been proposed in the literature to improve the forecasting accuracy. The most comprehensive of all popular and widely known statistical models which have been utilized in the last four decades for river flow forecasting are autoregressive moving average (ARMA) models. The popularity of the ARIMA model is due to its statistical properties as well as the well known Box-Jenkins methodology. In the literature, extensive applications and reviews of ARIMA models proposed for modeling of water resources time series were reported (Yurekli et al.,

HESSD

7, 3691–3731, 2010

2004; Muhamad and Hassan, 2005; Huang et al., 2004; Modarres, 2007; Fernandez and Vega, 2009; Wang et al., 2009). However, the ARIMA model is only a class of linear model and thus it can only capture linear feature of data time series. But many water resources time series are often full of nonlinearity and chaotic.

More advanced nonlinear methods such as neural networks have been frequently applied in nonlinear time series modeling and chaotic time series modeling in recent years (Karunasinghe and Liong, 2006, Rojas et al., 2008; Camastra and Colla, 1999, Han and Wang, 2009, and Abraham and Nath, 2001). ANNs provide an attractive alternative tool for forecasting researchers and have shown their nonlinear modeling capability in data time series forecasting. In the field chaotic time series modeling, the most popular neural network model is the feed-forward neural network with the back propagation (BP) algorithm (Ye, 2007). In the last decade, ANN have been widely extensively to model many nonlinear hydrologic processes such as in river flow (Firat, 2008; Shrestha et al., 2005; Shamseldin et al., 2002; Dolling and Varas, 2003; Muhamad and Hassan, 2005; Kisi, 2008; Wang et al., 2009, Keskin and Taylan, 2009), rainfall (Hung et al., 2009; de Vos and Rientjes, 2005) and ground water (Affandi and Watanabe, 2007; Birkinshaw et al., 2008).

More advanced AI is support vector machine (SVM) is proposed by Vapnik and his co-workers in 1995 through statistical learning theory. The SVM is a powerful methodology and has become a hot topic of intensive study due to its successful employed to solve most non-linear regression and time series problem and becoming increasingly in the modeling and forecasting of hydrological and water resource processes. Several studies have been carried out using SVM in hydrological modelling such as stream flow forecasting (Wang et al., 2009, Asefa et al., 2006; Lin et al., 2006), rainfall runoff modeling (Dibike et al.,2002) and flood stage forecasting (Liong and Sivapragasam, 2002; Yu et al., 2006). The standard SVM is solved using quadratic programming methods. However, this method is often time consuming and has higher computational burden because of the required constrained optimization programming.

HESSD

7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

Least squares support vector machines (LSSVM), as a modification of SVM which was introduced by Suykens (1999). The method uses equality constraints instead of inequality constraints and adopts the least squares linear system as its loss function, which is computationally attractive. LSSVM also has good convergence and high precision. Hence, this method is easier to use than quadratic programming solvers in SVM method. The major advantage of LS-SVM is that it is computationally very cheap while it still possesses some important properties of the SVM. In the water resource, the LSSVM method has received very little attention literature and only a few applications of LSSVM to modeling of environmental and ecological systems such as water quality prediction (Yunrong and Liangzhong, 2009).

One sub-model of ANN is a group method of data handling (GMDH) algorithm was first developed by Ivakhnenko (1971) as a multivariate analysis method for modeling and identification of complex systems. The main idea of GMDH is to build an analytical function in a feed-forward network based on a quadratic node transfer function whose coefficients obtained by using a regression technique. This model has been successfully used to deal with uncertainty, linear or nonlinearity of systems in a wide range of disciplines such as engineering, science, economy, medical diagnostics, signal processing and control systems (Tamura and Kondo, 1980; Ivakhnenko, 1995; Voss and Feng, 2002). In the water resource, the GMDH method has received very little attention literature and only a few applications of GMDH to modeling of environmental and ecological systems (Chang and Hwang, 1999; Onwubolu et al., 2007; Wang et al., 2005).

There have been several studies suggesting hybrid models, combining the ARIMA and ANN model (Zhang, 2003; Jain and Kumar, 2006; Su et al., 1997; Wang et al., 2005), the GMDH and ANN model (Wang et al., 2005), GMDH and differential evolution (Onwubolu, 2008), ARIMA and support vector machine (SVM) (Chen and Wang, 2007), ANN and Fuzzy system (Yang et al., 2006). Their results showed that the hybrid model can be an effective way to improving predictions achieved by either of the models used separately.

HESSD
7, 3691–3731, 2010

In this paper, a novel hybrid GMDH-type algorithm is proposed by integrating simple GMDH with LSSVM to forecast river flow time series data. The hybrid model combines GMDH and LSSVM into one methodology, known as GLSSVM. To verify the application of this approach, the hybrid model was compared with ARIMA, ANN, GMDH and LSSVM models using the monthly river flow of Selangor and Bernam rivers located in Selangor of Malaysia.

2 Individual forecasting models

This section presents the ARIMA, ANN, GMDH and LSSVM models used for modeling time series. The reason for choosing these models in this study were because these methods have been widely and successfully used in forecasting time series.

2.1 The autoregressive integrated moving average model

The ARIMA models were introduced by Box and Jenkins (1970), have been one of the most popular approaches to the analysis of the time series and prediction. The general ARIMA models are compound of a seasonal and non-seasonal part are represented by the following way:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D x_t = \theta_q(B)\Theta_Q(B^s)a_t \quad (1)$$

where $\phi(B)$and $\theta(B)$ are polynomials of order p and q, respectively; $\Phi(B^s)$ and $\Theta(B^s)$ are polynomials in B^s of degrees $P$and Q, respectively; p order of non-seasonal auto regression; d number of regular differencing; q order of the non-seasonal moving average; P order of seasonal auto regression; D number of seasonal differencing; Q order of seasonal moving average; and s length of season. Random errors, $a_t$are assumed to be independently and identically distributed with a mean of zero and a constant variance of σ^2. The order of an ARIMA model is represented by ARIMA (p, d, q) and the order of an SARIMA model is represented by ARIMA(p, d, q) × (P,D,Q)$_s$. The term (p,

HESSD
7, 3691–3731, 2010

d, q) is the order of the non-seasonal part and $(P,D,Q)_s$ is the order of the seasonal part.

The Box-Jenkins methodology is basically divided into four steps: identification, estimation, diagnostic checking and forecasting. In the identification step, transformation is often needed to make time series stationary. The behavior of the autocorrelation (ACF) and partial autocorrelation function (PACF) is used to see whether the series is stationary or not, seasonal or non-seasonal. The next step is choosing a tentative model by matching both ACF and PACF of the stationary series. Once a tentative model is identified, the parameters of the model are estimated. The last step of model building is the diagnostic checking of model adequacy, basically to check if the model assumptions about the error, a_t are satisfied. Diagnostic checking using the ACF and PACF of residuals was carried out, which can be referred to Brockwell and Davis (2002). If the model is not adequate, a new tentative model should be identified followed by the steps of parameter estimation and model verification. The process is repeated several times until a satisfactory model is finally selected. The forecasting model was then used to compute the fitted values and forecasts values.

The Akaike's Information Criterion (AIC) is used to evaluate the goodness of fit with smaller values indicating a better fitting and more parsimonious model than larger values (Akaike, 1974). Mathematical formulation of AIC is defined as

$$\mathrm{AIC} = \ln\left(\frac{\sum_{t=1}^{n} e_t^2}{n}\right) + \frac{2p}{n} \tag{2}$$

where p the number of parameters and n the periods of data.

2.2 The artificial neural network model

The ANN are flexible computing has been extensively studied and used for time series forecasting in many areas of science and engineering since early 1990. An ANN is a mathematical model which has a highly connected structure similar to brain cells.

The model has the capability of a complex mapping between input and output that enables the network to approximate nonlinear functions. Single hidden layer feed forward network is the most widely used model form for time series modeling and forecasting (Zhang et al., 1998). The model usually consists of three layers: the first layer is the input layer where the data are introduced to the network, the second layer is the hidden layer where data are processed and the last layer is the output layer where the results of given input are produced. The structure of a feed-forward ANN is shown in Fig. 1.

The output of the ANN assuming a linear output neuron j, a single hidden layer with h sigmoid hidden nodes and an input variable (x_t) is given by

$$x_t = g\left(\sum_{j=1}^{h} w_j f(s_j) + b_k\right) \tag{3}$$

where $g(\cdot)$ is the linear transfer function of the output neuron k and b_k is its bias, w_j is the connection weights between hidden layers and output units, $f(\cdot)$ is the transfer function of the hidden layer (Coulibaly and Evora, 2007). The transfer functions can take several forms and the most widely used transfer functions are

Log-sigmoid: $f(s_i) = \mathrm{logsig}(s_i) = \frac{1}{1+\exp(-s_i)}$

Linear: $f(s_i) = \mathrm{purelin}(s_i) = s_i$

Hyperbolic tangent sigmoid: $f(s_i) = \mathrm{tansig}(s_i) = \frac{2}{1+\exp(-2s_i)} - 1$

where $s_i = \sum_{i=1}^{n} w_i x_i$ is the input signal referred to as the weighted sum of incoming information.

For a univariate time series forecasting problem, the inputs of the network are the past lagged observations $(x_{t-1}, x_{t-2}, .., x_{t-p})$ and the output is the predicted value (x_t)

HESSD
7, 3691–3731, 2010

(Zhang et al., 2001). Hence the ANN of Eq. (2) can be written as

$$x_t = g(x_{t-1}, x_{t-2}, \ldots, x_{t-p}, w) + \varepsilon_t \quad (4)$$

where $w$is a vector of all parameters and $g(.)$ is a function determined by the network structure and connection weights. Thus, in some senses, the ANN model is equivalent to a nonlinear autoregressive (NAR) model.

Several optimization algorithms can be used to train the ANN. Among the several training algorithms available, back-propagation has been the most popular and most widely used (Zou et al., 2007). In a back-propagation network, the weighted connections feed activations only in the forward direction from an input layer to the output layer. Theses interconnections are adjusted using an error convergence technique so that the network's response best matches are desire response.

2.3 The Least Square Support Vector Machines Model

The LSSVM is a new technique for regression. The LSSVM predictor is trained using a set of time series historic values as inputs and a single output as the target value. In the following, we briefly introduce LSSVM, which can be used for time series forecasting.

Consider a given training set of n data points $\{x_i, y_i\}_{i=1}^{n}$ with input data $x_i \in R^n$, $p$is the total number of data patterns and output$y_i \in R$. SVM approximate the function in the following form

$$y(x) = w^T \phi(x) + b \quad (5)$$

where $\phi(x)$represents the high dimensional feature spaces, which is nonlinearly mapped from the input space x. In LSSVM for function estimation, the optimization problem is formulated (Suykens et al., 2002):

$$\min J w, e = \frac{1}{2} w^T w + \frac{Y}{2} \mathrm{sum}_{i=1}^{n} e_i{}^2 \quad (6)$$

HESSD
7, 3691–3731, 2010
A hybrid least squares support vector machines and GMDH approach
R. Samsudin et al.

Subject to the equality constraints

$$y(x) = w^T \phi(x_i) + b + e_i \quad i = 1, 2, \ldots, n$$

The solution is obtained after constructing the Lagrange

With Lagrange multipliers α_i. The conditions for optimality are given by

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i \phi(x_i),$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0,$$

$$\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i,$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow w^T \phi(x_i) + b + e_i - y_i = 0,$$

for $i = 1, 2, \ldots, n$. After elimination of e_i and w the solution is given by the following set of linear equations:

where $y = [y_1; \ldots; y_n]$, $1 = [1; \ldots; 1]$, $\alpha = [\alpha_1; \ldots; \alpha_n]$. According to Mercer's condition, the kernel function can be defined as

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j), \quad i, j = 1, 2, \ldots, n \tag{7}$$

This finally leads to the following LSSVM model for function estimation:

$$y(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) + b \tag{8}$$

where α_i, b are the solution to the linear system. Any function that satisfies Mercer's condition can be used as the kernel function. The choice of the kernel function $K(.,.)$

HESSD
7, 3691–3731, 2010

has several possibilities. $K(x_i, x_j)$ is defined as the kernel function. The value of the kernel is equal to the inner product of two vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ in the feature space $\phi(x_i)$and $\phi(x_j)$, that is, $K(x_i, x_j) = \phi(x_i) * \phi(x_j)$. The structure of a SVM is shown in Fig. 2.

The typical examples of the kernel function are as follows:

Linear: $K(x_i, x_j) = x_i^T x_j$

Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$

Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \quad \gamma > 0$

$$\text{Radial basis function (RBF)}: K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \quad \gamma > 0 \tag{9}$$

Here γ, r and d are kernel parameters. The kernel parameters should be carefully chosen as they implicitly define the structure of the high dimensional feature space $\phi(x)$and thus control the complexity of the final solution.

2.4 The group method of data handling model

The algorithm of Group Method of Data Handling (GMDH) was introduced by Ivakhnenko in early 1970 as a multivariate analysis method for modeling and identification of complex systems. The GMDH method was originally formulated to solve higher order regression polynomials specially for solving modeling and classification problems. General connection between inputs and output variables can be expressed by a complicated polynomial series in the form of the Volterra series, known as the Kolmogorov-Gabor polynomial (Ivakhnenko, 1971):

$$y = a_0 + \sum_{i=1}^{M} a_i x_i + \sum_{i=1}^{M} \sum_{j=1}^{M} a_{ij} x_i x_j + \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{k=1}^{M} a_{ijk} x_i x_j x_k + \ldots \tag{10}$$

where x is the input to the system, M is the number of inputs and $a$are coefficients or weights. However, for most application of the quadratic forms are called as partial descriptions (PD) where only two variables are used in the form

$$y = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \tag{11}$$

to predict the output. To obtain the value of the coefficients a for each m models, a system of Gauss normal equation is solved. The coefficient a_i of nodes in each layer are expressed in the form of

$$\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

where $\mathbf{Y} = [y_1\, y_2 \ldots y_M]^T$, $\mathbf{A} = [a_0, a_1,\, a_2, a_3, a_4, a_5]$,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1p} & x_{1q} & x_{1p}x_{1q} & x_{1p}^2 & x_{1q}^2 \\ 1 & x_{2p} & x_{2q} & x_{2p}x_{2q} & x_{2p}^2 & x_{2q}^2 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{Mp} & x_{Mq} & x_{Mp}x_{Mq} & x_{Mp}^2 & x_{Mq}^2 \end{bmatrix}$$

and M is the number of observations in the training set.

The main function of GMDH is based on the forward propagation of signal through nodes of the net similar to the principal used in classical neural nets. Every layer consists of simple nodes each of which performs its own polynomial transfer function and passes its output to nodes in the next layer. The basic steps involved in the conventional GMDH modeling (Nariman-Zadeh et al., 2002) are as follows:

Step 1: Select normalized data X=$\{x_1, x_2, \ldots, x_M\}$ as input variables. Divide the available data into training and testing data sets.

Step 2: Construct ${}^M C_2 = M(M-1)/2$ new variables, in the training data set and construct the regression polynomial for first layer by forming the quadratic expression which approximates the output y in Eq. (11).

Step 3: Identify the contributing nodes at each hidden layer according to the value of mean root square error (RMSE). Eliminate the least effective variable by replacing the columns of X (old columns) with the new columns of Z.

Step 4: The GMDH algorithm is carried out by repeating steps 2 and 3 of the algorithm. When the errors of the test data in each layer stop decreasing, the iterative computation is terminated.

The configuration of the conventional GMDH structure is shown in Fig. 3.

2.5 The hybrid model

In the proposed method, the combination of GMDH and LSSVM (GLSSVM) are applied to enhance the capability of hybrid model. As input variables are selected by the decision made by GMDH and LSSVM model is used as time series forecasting. The hybrid model procedure is carried out in the following step:

Step 1: The normalized data are separated into the training and testing sets data.

Step 2: All combinations of two input variables $(x_i, x_j)$are generated in each layer. The number of input variables are ${}^{M}C_2 = \frac{M!}{(M-2)!2!}$. Construct the regression polynomial for this layer by forming the quadratic expression which approximates the output y in Eq. (11). The coefficient vector of the PD is determined by the least square estimation approach.

Step 3: Determine new input variables for the next layer. The output x' variable which give the smallest of root mean square error (RMSE) for the train data set is combined with the input variables $\{x_1, x_2, ..., x_M, x'\}$ with $M = M+1$. The new input $\{x_1, x_2, ..., x_M, x'\}$ of the neurons in the hidden layers are use as input for the LSSVM model.

Step 4: The GLSSVM algorithm is carried out by repeating steps 2 to 4 until $k = 5$ iteration. The GLSSVM model with the minimum value of the RMSE is selected as the output model. The configuration of the GLSSVM structure is shown in Fig. 4.

HESSD
7, 3691–3731, 2010

3 Case study

In this study, monthly flow data from Selangor River and Bernam River in Selangor, Malaysia are selected as a study site. The location of the Selangor and Bernam rivers are shown in Fig. 5. The Bernam River located between the Malaysian states of Perak and Selangor, demarcating the border of the two states. The Selangor River is a major river in Selangor, Malaysia. It runs from Kuala Kubu Bharu in the east and empties into the Straits of Malacca at Kuala Selangor in the west.

The catchment area at Selangor site (3.24°, 101.26°) is 1450 km^2 and the mean elevation is 8 m. Meanwhile the catchment area at Bernam site (3.48°, 101.21°) is 1090 km^2 with the mean elevation is 19 m. Both rivers basin have quite significant effect on drinking water supply, irrigation and aquaculture activities such as the cultivation of fresh water fishes for human consumption.

The observed data are within 47 years (564 months) long with an observation period between January 1962 and December 2008 for Selangor River and 43 years (516 months) from January 1966 to December 2008 for Bernam River. The training dataset of 504 monthly records (January 1962 to December 2004) for Selangor River and 456 monthly records (January 1966 to December 2004) were used to train the network to obtain parameters model. Another dataset consisting of 60 monthly (January 2005 to December 2008) records was used as testing dataset for both stations (Fig. 6).

Before starting the training, the collected data were normalized within the range 0 to 1 using the following formula

$$x_t = 0.1 + \frac{y_t}{1.2\max(y_t)}$$

where x_t is the normalized value, y_t is the actual value and $\max(y_t)$ is the maximum value in the collected data.

The performances of each model for both training and forecasting data are evaluated according to the root-mean-square error (RMSE) and correlation coefficient (R) which

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

are widely used for evaluating results of time series forecasting. RMSE and MAE are the commonly used error index statistics. The RMSE and R are defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_i - o_i)^2}$$

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(o_i - \bar{o})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - \bar{o})}}$$

where o_i and y_i are the observed and forecasted values at data point i, respectively, $\bar{o}$ is the mean of the observed values, and N is the number of data points. The criteria to judge the best model are relatively small of RMSE in the training and testing. Correlation coefficient measures how well the flows predicted correlate with the flows observed. Clearly, the R value close to unity indicates a satisfactory result, while a low value or close to zero implies an inadequate result.

4 Result and discussion

4.1 Fitting the ARIMA models to the data

The sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF) for Selangor River and Bernam River series are plotted in Figs. 7 and 8, respectively. The ACFs curve for monthly flow data of two study sites decayed with mixture of sine wave pattern and exponential curve that reflects the random periodicity of the data and indicates the need for seasonal MA terms in the model. For PACF, there were significant lag at spikes from lag 1 to 5, which suggest an AR process. In the PACF there were significant spikes present near lag 12 and 24. Therefore the series need for seasonal AR process. The identification of best model for river flow series based on

HESSD
7, 3691–3731, 2010

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

minimum AIC is shown in Table 1. The criteria to judge the best model based on AIC show that ARIMA$(1,0,0)\times(1,0,1)_{12}$ was selected as the best model for Selangor River where the ARIMA $(2,0,0)\times(2,0,2)_{12}$ were relatively best model for Bernam River.

Since the ARIMA $(1,0,0)\times(1,0,1)_{12}$ was the best model for Selangor River and ARIMA $(2,0,0)\times(2,0,2)_{12}$ for Bernam River, then the model was used to identify the input structures. The ARIMA $(2,0,0)\times(2,0,2)_{12}$ model can be written as

$$(1-0.3515B-0.1351B^2)(1-0.7014B^{12}-0.2933B^{24})x_t=(1-0.5802B^{12}-0.3720B^{24})$$

$$a_t x_t=0.3515x_{t-1}+0.1351x_{t-2}+0.7014x_{t-12}-0.2465x_{t-13}-0.0948x_{t-14}$$

$$+0.2933x_{t-24}$$

$$-0.1031x_{t-25}-0.0396x_{t-26}-0.5802a_{t-12}-0.3720a_{t-24}+a_t$$

and the ARIMA $(1,0,0)\mathrm{x}(1,0,1)_{12}$ model can be written as

$$(1-0.4013B)(1-0.9956B^{12})x_t=(1-0.9460B)a_t$$

$$x_t=0.4013x_{t-1}+0.9956x_{t-12}-0.3995x_{t-13}-0.9460a_{t-12}+a_t$$

the above equation for Selangor River can be rewritten as

$$x_t=f(x_{t-1},x_{t-12},x_{t-13},a_{t-12}) \tag{12}$$

and for Bernam River as

$$x_t=f(x_{t-1},x_{t-2},x_{t-12},x_{t-13},x_{t-14},x_{t-24},x_{t-25},x_{t-26},a_{t-12},a_{t-24}) \tag{13}$$

4.2 Fitting ANN to the data

One of the most important steps in developing a satisfactory forecasting model such as ANN and LSSVM models is the selection of the input variables. In this study, the six input structures which have various input variables were trained and tested by LSSVM

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

and ANN. Four approaches were used to identify the input structures, the first three inputs were accomplished by setting the input layer nodes equal to the number of the lagged variables from river flow data, $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$ where p is the time delay. The time delay was taken 2, 6 and 12 months. The second, third and forth approaches were identified using correlation analysis, stepwise regression analysis and ARIMA model, respectively. The input structures of forecasting models are shown in Tables 2 and 3.

In this study, a typical three-layer feed-forward ANN model was constructed for forecasting monthly river flow time series. The training and testing data were normalized in the range zero to one. From the input layer to the hidden layer, the hyperbolic tangent sigmoid transfer function that has been commonly used in hydrology was applied. From the hidden layer to the output layer, a linear function was employed as the transfer function because the linear function is known to be robust for a continuous output variable.

The network was trained for 5000 epochs using the conjugate gradient descent backpropagation algorithm with a learning rate of 0.001 and a momentum coefficient of 0.9. The six models (M1-M6) having various input structures were trained and tested by ANN models and the optimal number of neuron in the hidden layer was identified using several practical guidelines. These include using I/2 (Kang, 1991), 2I (Wong, 1991) and 2I+1 (Lipmann, 1987), where I is the number of input. The effect of changing the number of hidden neurons on the RMSE, MAE and R of the data set is shown in Table 4.

Table 4 shows the performance of ANN varying with the number of neurons in the hidden layer.

For Selangor River, in the training phase, the M3 model with the number of hidden neurons 2I+1 obtained the best RMSE, MAE and R statistics of 0.098, 0.07 and 0.66, respectively. While in testing phase, the M6 model with 2I + 1 numbers of hidden neurons is the best RMSE, MAE and R statistics of 0.112, 0.079 and 0.594, respectively.

For the Bernam River, in the training phase, the M6 model with the number of hidden neurons are I/2 obtained the best RMSE, MAE and R statistics of 0.0608, 0.0474, and

0.9137, respectively. While in testing phase, the M6 model with I/2 numbers of hidden neurons was the smallest RMSE and MAE of 0.0714 and 0.0506, respectively; while the M6 model with 2I numbers of hidden neurons obtained the highest of R statistics of 0.8479.

Hence, according to their performances indices, ANN (4,9,1) was selected for appropriate ANN model in Selangor River whereas ANN (10,5,1) was selected for appropriate ANN model in Bernam River.

4.3 Fitting LSSVM to the data

There is no theory that can use to guide the selection the optimal number of input nodes of the LSSVM model. In the training and testing of LSSVM model, the same input structures of the data set (M1–M6) were used. The precision and convergence of LSSVM were also affected by (γ, σ^2). There is no structured way to choose the optimal parameters of LSSVM. In order to obtain the optimal model parameters of the LSSVM, a grid search algorithm was employed in the parameter space. Cross-validation is a popular technique for estimating generalization performance. To get good generalization ability, we conducted a validation process to decide parameters. In order to better evaluate the performance of the proposed approach, we considered a grid search of (γ, σ^2) with γ in the range 10 to 1000 and σ^2 in the range 0.01 to 1.0. For each hyperparameter pair (γ, σ^2) in the search space, 5-fold cross validation on the training set was performed to predict the prediction error. The best fit model structure for each model was determined according to criteria of performance evaluation. In the study, the LSSVM model was implemented with software package LS-SVMlab1.5 (Pelckmans et al., 2003) using MATLAB. The LSSVM method was employed, so a kernel function had to be selected from the qualified function. Many works on the use LSSVM in time series modeling and forecasting have demonstrated the favorable performance of the RBF (Liu and Wang, 2008, Yu et al., 2006; Gencoglu and Ulyar, 2009). Therefore, the RBF, which has a parameter γ as in Eq. (8), was adopted in this work. Table 5 shows the performance results obtained in the training and testing period

HESSD
7, 3691–3731, 2010

of the LSSVM approach.

As can be seen from Table 5, the LSSVM models were evaluated based on their performance in the training and testing sets. For Selangor River, in the training phase, the best value of the RMSE and R statistics 0.09 and 0.69 (in M6), respectively. While in testing phase, the lowest value of the RMSE was 0.106 (in M3) and the highest value of the R was 0.63 (in M5).

4.4 Fitting GMDH and GLSSVM to the data

In designing the GMDH and GLSSVM model, one must determine the following variables: the number of input nodes and the number of layers. The selection of the number of input corresponds to the number of variables play important roles for many successful applications of GMDH.

GMDH works by building successive layers with complex connections that were created by using second-order polynomial function. The first layer created was made by computing regressions of the input variables. The second layer was created was computing regressions of the output value. Only the best was chosen at each layer and this process continued until a pre-specified selection criterion was found.

The proposed hybrid learning architecture was composed of two stages. In the first stage, GMDH was used to determine the useful inputs for LSSVM method. The estimated output values x' was used as the feedback value and it was combined with the input variables $\{x_1, x_2, \ldots, x_M\}$ in the next loop calculations. The second stage the LSSVM mapping the combination inputs variables $\{x_1, x_2, \ldots, x_M, x'\}$ to seek optimal solutions for determining the best output for forecasting.

To make the GMDH and GLSSVM models simple and reduce some computational burden, only six levels of input nodes (M1–M6) and five hidden layers (k) from 1 to 5 were selected for experiment.

In LSSVM model, parameter values for γ and σ^2 needed to be first specified. The LSSVM parameters were selected by grid searching with γin the range 10 to 1000 and σ^2 in the range 0.01 to 1.0. For each parameter pair(γ, σ^2) in the search space, 5-fold

HESSD
7, 3691–3731, 2010

cross validation on the training set was performed to predict the prediction error. The performances of GMDH and GLSSVM for time series forecasting models are given in Table 5.

For the GMDH model, in the training phase, the best value of the RMSE and R statistics was 0.096 and 0.68 (in M5), respectively. Analyzing the result during testing, the best value of the RMSE and R statistics was 0.103 and 0.59 (in M3), respectively.

In the training phase, GLSSVM model obtained the best RMSE and R statistics of 0.069 and 0.844 (in M2), respectively. While in testing phase, the lowest value of the RMSE was 0.101 (in M3) and the highest value of the R was 0.64 (in M5).

For Bernam River, in the training and testing phase, the best value of RMSE, MAE and R for LSSM, GMDH and GLSSVM models were obtained using M6.

The model that performed best in testing was chosen as final model for forecasting of sixty monthly flows. As can seen from Table 5, for Selangor River, the model input M5 gave the best performance for LSSVM and GLSSVM models, and M3 for GMDH model. While for Bernam River, the model input M6 gave the best performance for LSSVM, GMDH and GLSSVM models and hence, these model inputs were chosen as the final model.

4.5 Comparisons of forecasting models

For further analysis, the error statistics of the optimum ARIMA, ANN, GMDH, LSSVM and GLSSVM were compared. The performances of all methods for training and testing data set are given in Table 6.

Comparing performances of ARIMA, ANN, GMDH, LSSVM and GLSSVM models, for Selangor and Bernam rivers in training, the lowest RMSE and the largest R were calculated for GLSSVM model, respectively. For testing data, the best value of MAE and R were found for GLSSVM model. However, the lowest RMSE were observed for GMDH model for Selangor River and LSSVM model for Bernam River. From the Table 6, it is evident that the GLSSVM performed better than the ARIMA, ANN, GMDH and LSSVM models in training and testing process.

HESSD
7, 3691–3731, 2010

Figures 9 and 10 show the comparison of time series and scatter plots between modeled results by the five models and actual data for the last sixty months during testing stage for Selangor and Bernam rivers, respectively. All five models gave close approximations of the actual observations, suggesting that these approaches are applicable for modeling river flow time series data. However, the tested line generated from GLSSVM is much closer to the actual value line than tested line generated from other models. Similar to R and fit line equation coefficients, the GLSSVM is slightly superior to the other models. The results obtained in this study indicate that the GLSSVM model is powerful tools to model the river flow time series and can give good prediction performance than ARIMA, ANN, GMDH and LSSVM time series approaches. The results indicate that the best performance can be obtained by GLSSVM model followed by LSSVM, GMDH, ANN and ARIMA models.

5 Conclusions

Monthly river flow estimation is vital in hydrological practices. There are plenty of methods used to predict river flows. In this paper, we have demonstrated how the monthly river flow could be well represented by the hybrid models, combining the GMDH and LSSVM models. To illustrate the capability of the LSSVM model, Selangor River and Bernam River, located in Selangor, Malaysia was chosen as a case study. The river flow forecasting models having various input structures were trained and tested to investigate the applicability of GLSSVM compared with ARIMA, ANN, GMDH and LSSVM models. One of the most important in developing a satisfactory forecasting model such as ANN, GMDH and LSSVM models is the selection of the input variables. Empirical results on the two data sets using five different models clearly reveal the efficiency of the hybrid model. In terms of RMSE and R values, for both data sets, hybrid model has the best in training. In testing, high correlation coefficient (R) was achieved by using the hybrid model for both data sets. However, the lowest value of RMSE were achieved using the GMDH for Selangor River and LSSVM for Bernam River. These results show that the hybrid model provides a robust modeling capable of

HESSD
7, 3691–3731, 2010

capturing the nonlinear nature of the complex river flow time series and thus producing more accurate forecasts.

Acknowledgements. The authors would like to thank the Ministry of Science, Technology and Innovation (MOSTI), Malaysia for funding this research with grant number 79346 and Muda Agricultural Development Authority (MADA), Kedah, Malaysia for providing the data of rice yield.

References

Abraham, A. and Nath, B.: Hybrid intelligent systems de-sign – A review of a decade of research," Technical Report (5/2000), Gippsland School of Computing and Information Technology, Monash University, Australia, 2001.

Abraham, A. and Nath, B.: A neuro-fuzzy approach for modeling electricity demand in Victoria, Appl. Soft Comput., 1(2), 127–138, 2001.

Affandi, A. K. and Watanabe, K.: Daily groundwater level fluctuation forecasting using soft computing technique, Nature and Science, 5(2), 1–10, 2007.

Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automat. Control, 19, 716–723, 1974.

Asefa.: Multi-time scale stream flow prediction: The support vector machines approach, J. Hydrol., v318, 7–16, 2006.

Birkinshaw, S. J., Parkin, G., and Rao, Z.: A hybrid neural networks and numerical models approach for predicting groundwater abstraction impacts, J. Hydroinform., 10(2), 127–137, 2008.

Box, G. E. P. and Jenkins, G.: Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco, CA, 1970.

Brockwell, P. J. and Davis, R. A.: Introduction to Time Series and Forecasting, Springer, Berlin, Germany, Introduction to Time Series and Forecasting, 2nd edn., Springer-Verlag, Inc. New York, NY, 2002.

Camastra, F. and Colla, A.: Mneural short-term rediction based on dynamics reconstruction, ACM, 9(1), 45–52, 1999.

Chang, F.J. and Hwang, Y. Y.: A self-organization algorithm for real-time flood forecast, Hydrol. Process., 13, 123–138, 1999.

Chen, K. Y. and Wang, C. H.: A hybrid ARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan, Expert Syst. Appl., 32, 254–264, 2007.
Coulibaly, P. and Evora, N. D.: Comparison of neural network methods for infilling missing daily weather records, J. Hydrol., 341, 27–41, 2007.
de Vos, N. J. and Rientjes, T. H. M.: Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation, Hydrol. Earth Syst. Sci., 9, 111–126, doi:10.5194/hess-9-111-2005, 2005.
Dibike, Y. B., Velickov, S., Solomatine, D. P., and Abbott, M. B.: Model induction with support vector machines: introduction and applications, ASCE Journal of Computing in Civil Engineering, 15(3), 208–216, 2001.
Dolling, O. R. and Varas, E. A.: Artificial neural networks for streamflow prediction, J. Hydraul. Res., 40(5), 547–554, 2003.
Fernandez, C. and Vega, J. A.: Streamflow drought time series forecasting: a case study in a small watershed in north west spain, Stoch. Environ. Res. Risk Assess, 23, 1063–1070, 2009.
Firat, M.: Comparison of Artificial Intelligence Techniques for river flow forecasting, Hydrol. Earth Syst. Sci., 12, 123–139, doi:10.5194/hess-12-123-2008, 2008.
Gencoglu, M. T. and Uyar, M.: Prediction of flashover voltage of insulators using least square support vector machines, Expert Syst. Appl., 36, 10789–10798, 2009.
Han,M. and Wang, M.: Analysis and modeling of multivariate chaotic time series based on neural network, Expert Syst. Appl., 2(36), 1280–1290, 2009.
Hung, N. Q., Babel, M. S., Weesakul, S., and Tripathi, N. K.: An artificial neural network model for rainfall forecasting in Bangkok, Thailand, Hydrol. Earth Syst. Sci., 13, 1413–1425, doi:10.5194/hess-13-1413-2009, 2009.
Huang, W., Bing Xu, B., and Hilton, A.: Forecasting flow in apalachicola river using neural networks, Hydrol. Process., 18, 2545–2564, 2004.
Ivanenko, A. G.: Polynomial theory of complex system, IEEE Trans. Syst., Man Cybern. SMCI-1, 1, 364–378, 1971.
Ivakheneko A. G. and Ivakheneko G. A.: A review of problems solved by algorithms of the GMDH, Pattern Recognition and Image Analysis, 5(4), 527–535, 1995.
Jain, A. and Kumar, A.: An evaluation of artificial neural network technique for the determination of infiltration model parameters, Appl. Soft Comput., 6, 272–282, 2006.

Kang, S.: An Investigation of the Use of Feedforward Neural Network for Forecasting, Ph.D. Thesis, Kent State University, 1991.
Karunasinghe, D. S. K. and Liong, S. Y.: Chaotic time series prediction with a global model: Artificial neural network, J. Hydrol., 323, 92–105, 2006.
Keskin, M. E. and Taylan, D.: Artifical models for interbasin flow prediction in southern turkey, J. Hydrol. Eng., 14(8), 752–758, 2009.
Kisi, O.: River flow forecasting and estimation using different artificial neural network technique, Hydrol. Res., 39(1), 27–40, 2008.
Kondo, T. and Ueno, J.: Revised gmdh-type neural network algorithm with a feedback loop identifying sigmoid function neural network, Int. J. Innov. Comput. I., 2(5), 985–996, 2006.
Lin, J. Y., Cheng, C. T., and Chau, K. W.: Using support vector machines for long-term discharge prediction, Hydrolog. Sci. J., 51(4), 599–612, 2006.
Lippmann, R. P.: An introduction to computing with neural nets, IEEE ASSP Magazine, April, 4–22, 1987.
Liong S. Y. and Sivapragasam, C.: Flood stage forecasting with support vector machines, J. Am. Water Resour. As., 38(1), 173–196, 2002.
Liu, L. and Wang, W.: Exchange rates forecasting with least squares support vector machines, International Conference on Computer Science and Software Engineering, 1017–1019, 2008.
Maier, H. R. and Dandy, G. C.: Neural networks for the production and forecasting of water resource variables: a review and modelling issues and application, Environmental Modelling and Software, 15, 101–124, 2000.
Modarres, R.: Streamflow drought time series forecasting, Stoch. Environ. Res. Risk Assess., 21, 223–233, 2007.
Muhamad, J. R. and Hassan, J. N.: Khabur River flow using artificial neural networks, Al_Rafidain Engineering, 13(2), 33–42, 2005.
Onwubolu, G. C.: Design of hybrid differential evolution and group method of data handling networks for modeling and prediction, Inform. Sciences, 178, 3616–3634, 2008.
Onwubolu, G.C., Buryan, P., Garimella, S., Ramachandran V., Buadromo, V., and Abraham, A.: Self-organizing data mining for weather forecasting, IADIS European Conference Data Ming, 81–88, 2007.
Pelckmans, K., Suykens, J., Van, G., de Brabanter, J., Lukas, L., Hanmers, B., De Moor, B., and Vandewalle, J.: LS-SVMlab: A MATLAB/C Toolbox for Least Square Support Vector

Machines, available at: www.esat.kuleuven.ac.be/sista/lssvmlab, 2003.

Rojas, I., Valenzuela, O., Rojas, F., Guillen, A., Herrera, L. J., Pomares, H., Marquez, L., and Pasadas, M.: Soft-computing techniques and ARMA model for time series prediction, Neurocomputing, 71, 4–6, 519–537, 2008.

Shamseldin, A. Y., Nasr, A. E., and O'Connor, K. M.: Comparison of different forms of the Multi-layer Feed-Forward Neural Network method used for river flow forecasting, Hydrol. Earth Syst. Sci., 6, 671–684, doi:10.5194/hess-6-671-2002, 2002.

Shrestha, R. R., Theobald, S., and Nestmann, F.: Simulation of flood flow in a river system using artificial neural networks, Hydrol. Earth Syst. Sci., 9, 313–321, doi:10.5194/hess-9-313-2005, 2005.

Su, C. T., Tong, L. I., and Leou, C. M.: Combination of time series and neural network for reliability forecasting modeling, Journal of Chinese Industrial Engineering, 14, 419–429, 1997.

Suykens, J. A. K. and Vandewalle, J.: Least squares support vector machine classifiers, Neural Process. Lett, 9(3), 293–300, 1999.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., and Vandewalle, J.: Least squares support vector machines, World Scientific, 2002, Singapore, 2002.

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

Vapnik, V.: The nature of Statistical Learning Theory, Springer Verlag, Berlin, 1995.

Voss, M. S. and Feng, X.: A new methodology for emergent system identification using particle swarm optimization (PSO) and the group method data handling (GMDH), GECCO 2002, 1227–1232, 2002.

Wang, W. C., Chau, K. W., Cheng, C. T., and Qiu, L.: A Comparison of Performance of Several Artificial Intelligence Methods for Forecasting Monthly Discharge Time Series, J. Hydrol., 374, 294–306, 2009.

Wang, W., Van, P. Gelder, P., and Vrijling, J.K.: Forecasting daily stream flow using hybrid ANN models, J. Hydrol., 324, 383–399, 2006.

Wang, X., Li, L., Lockington, D., Pullar, D., and Jeng, D. S.: Self-organizing polynomial neural network for modeling complex hydrological processes, Research Report. No. R861:1-29, 2005.

Wong, F. S.: Time series forecasting using backpropagation neural network, Neurocomputing, 2, 147–159, 1991.

Yang, Q., Lincang Ju, L., Ge, S., Shi, R., and Yuanli Cai, Y.: Hybrid fuzzy neural network control for complex industrial process, International conference on intelligent computing, Kunming , CHINA, 533–538, 2006.

Yu, P. S., Chen, S. T., and Chang, I. F.: Support vector regression for real-time flood stage forecasting, J. Hydrol., 328(3–4), 704–716, 2006.

Yunrong, X. and Liangzhong, J.: Water quality prediction using LS-SVM with particle swarm optimization, Second International Workshop on Knowledge Discovery and Data Mining, 900–904, 2009.

Yurekli, K., Kurunc, A., and Simsek, H.: Prediction of Daily Streamflow Based on Stochastic Approaches, Journal of Spatial Hydrology, 4(2), 1–12, 2004.

Zadeh, N. N., Darvizeh, A., Felezi, M. E., and Gharababaei, H.: Polynomial modelling of explosive process of metalic powders using GMDH-type neural networks and singular value decomposition, Model. Simul. Mater. Sc., 10, 727–744, 2002.

Zou, H. F., Xia, G. P., Yang, F. T. and Wang, H. Y.: An investigation and comparison of artificial neural network and time series models for chinese food grain price forecasting, Neurocomputing, 70, 2913–2923, 2007.

Zhang, G. P.: Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing, 50, 159–175, 2003.

Zhang, G., Patuwo, B. E., and Hu, M. Y.: Forecasting with artificial neural networks: the state of the art, Int. J. Forecasting., 14, 35–62, 1998.

Zhang, G. P., Patuwo, B. E., and Hu, M. Y.: A simulation study of artificial neural networks for nonlinear time-series forecasting, Computers and Operations Research, 28(4), 381–396, 2001.

HESSD

7, 3691–3731, 2010

Table 1. Comparison of ARIMA models' Statistical Results for Selangor and Bernam Rivers.

Selangor River		Bernam River	
ARIMA Model	AIC	ARIMA Model	AIC
$\mathbf{(1,0,0)}\times\mathbf{(1,0,1)}_{12}$	**−4.765**	$(1,0,0)\times(1,0,1)_{12}$	−4.458
$(1,0,0)\times(3,0,0)_{12}$	−4.620	$(5,0,0)\times(2,0,2)_{12}$	−4.251
$(1,0,0)\times(1,0,0)_{12}$	−4.514	$(3,0,0)\times(2,0,1)_{12}$	−4.459
$(1,0,1)\times(3,0,0)_{12}$	−4.614	$(2,0,0)\times(1,0,1)_{12}$	−4.466
$(1,0,1)\times(1,0,1)_{12}$	−4.757	$\mathbf{(2,0,0)}\times\mathbf{(2,0,2)}_{12}$	**−4.467**

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

Table 2. The Input Structure of the Models for Forecasting of Selangor River.

Model	Input Structure
M1	$x_t = f(x_{t-1}, x_{t-2})$
M2	$x_t = f(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-6})$
M3	$x_t = f(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}, x_{t-9}, x_{t-10}, x_{t-11}, x_{t-12})$
M4	$x_t = f(x_{t-1}, x_{t-2}, x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}, x_{t-10}, x_{t-11}, x_{t-12})$
M5	$x_t = f(x_{t-1}, x_{t-2}, x_{t-4}, x_{t-5}, x_{t-7}, x_{t-10}, x_{t-12})$
M6	$x_t = f(x_{t-1}, x_{t-2}, x_{t-12}, x_{t-13}, x_{t-14}, x_{t-24}, x_{t-25}, x_{t-26}, a_{t-12}, a_{t-24})$

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

Table 3. The Input Structure of the Models for Forecasting of Bernam River.

Model	Input Structure
M1	$x_t = f(x_{t-1}, x_{t-2})$
M2	$x_t = f(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-6})$
M3	$x_t = f(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}, x_{t-9}, x_{t-10}, x_{t-11}, x_{t-12})$
M4	$x_t = f(x_{t-1}, x_{t-2}, , x_{t-4}, x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}, x_{t-10}, x_{t-11}, x_{t-12})$
M5	$x_t = f(x_{t-1}, x_{t-2}, x_{t-4}, x_{t-5}, x_{t-7}, x_{t-10}, x_{t-12})$
M6	$x_t = f(x_{t-1}, x_{t-2}, x_{t-12}, x_{t-13}, x_{t-14}, x_{t-24}, x_{t-25}, x_{t-26}, a_{t-12}, a_{t-24})$

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

Table 4. Comparison of ANN structures for Selangor and Bernam River.

		Selangor River				Bernam River			
Model	Hidden	Training		Testing		Training		Testing	
Input	Layer	RMSE	R	RMSE	R	RMSE	R	RMSE	R
M1	I/2	0.1149	0.4569	0.1222	0.5143	0.1259	0.5377	0.1096	0.5084
	2I	0.1094	0.5311	0.1251	0.4521	0.1257	0.5398	0.1113	0.4939
	2I + 1	0.1093	0.5328	0.1237	0.4754	0.1263	0.5333	0.1110	0.4928
M2	I/2	0.1055	0.5804	0.1256	0.4562	0.1411	0.3363	0.1200	0.3306
	2I	0.1050	0.5859	0.1270	0.4398	0.1197	0.6013	0.1087	0.5368
	2I + 1	0.1050	0.5855	0.1232	0.4864	0.1232	0.5684	0.1112	0.4850
M3	I/2	0.1042	0.5969	0.1160	0.5640	0.1143	0.6410	0.1050	0.5738
	2I	0.1012	0.6276	0.1240	0.4905	0.1105	0.6705	0.1088	0.5543
	2I + 1	**0.0976**	**0.6599**	0.1238	0.5137	0.1107	0.6686	0.1039	0.5785
M4	I/2	0.1019	0.6199	0.1178	0.5547	0.1199	0.5931	0.1047	0.5692
	2I	0.1003	0.6355	0.1205	0.5251	0.1128	0.6525	0.1057	0.5701
	2I + 1	0.1023	0.6171	0.1181	0.5632	0.1123	0.6563	0.1089	0.5382
M5	I/2	0.1007	0.6315	0.1200	0.5256	0.1150	0.6352	0.1034	0.5863
	2I	0.1005	0.6342	0.1175	0.5409	0.1134	0.6476	0.1086	0.5397
	2I + 1	0.1014	0.6251	0.1253	0.5150	0.1127	0.6535	0.1048	0.5736
M6	I/2	0.1072	0.5672	0.1158	0.5690	**0.0608**	**0.9137**	**0.0714**	0.8464
	2I	0.1020	0.6186	0.1217	0.5027	0.0678	0.8898	0.0768	**0.8479**
	2I + 1	0.1044	0.5950	**0.1145**	**0.5682**	0.0648	0.8999	0.0781	0.8186

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

Table 5. The RMSE, MAE and R statistics of LSSVM Model for Selangor and Bernam River.

		Selangor River				Bernam River			
	Model	Training		Testing		Training		Training	
Model	Input	RMSE	R	RMSE	R	RMSE	R	RMSE	R
LSSVM	M1	0.1053	0.5792	0.1196	0.5280	0.1244	0.5530	0.1080	0.5263
	M2	0.1035	0.0505	0.1216	0.5110	0.1035	0.6033	0.1216	0.511
	M3	0.1108	0.6809	**0.1055**	0.5572	0.1108	0.6809	0.1055	0.5572
	M4	0.0997	0.6422	0.1163	0.5738	0.1044	0.6037	0.1031	0.6037
	M5	0.0961	0.6747	0.1126	**0.6269**	0.1021	0.7294	0.1009	0.6118
	M6	**0.0938**	**0.6932**	0.1119	0.5971	**0.0579**	**0.9319**	**0.0621**	**0.8727**
GMDH	M1	0.1079	0.5491	0.1251	0.4557	0.1235	0.5611	0.1072	0.5376
	M2	0.1025	0.6114	0.1199	0.5353	0.1025	0.6114	0.1199	0.5353
	M3	**0.0955**	**0.6776**	**0.1144**	**0.6052**	0.1101	0.6733	0.1034	0.5850
	M4	0.0973	0.6621	0.1176	0.5742	0.1142	0.6411	0.1008	0.6085
	M5	0.0956	0.6750	0.1164	0.5797	0.1119	0.6598	0.0992	0.6244
	M6	0.1065	0.5729	0.1224	0.5023	**0.0578**	**0.9216**	**0.0853**	**0.8387**
GLSSVM	M1	0.0908	0.7107	0.1127	0.5907	0.1180	0.6207	0.1044	0.5701
	M2	**0.0694**	**0.8441**	0.1187	0.5458	0.0694	0.8441	0.1187	0.5458
	M3	0.1006	0.7408	**0.1014**	0.6137	0.0900	0.7968	0.1046	0.5996
	M4	0.0698	0.8432	0.1511	0.5875	0.0783	0.8508	0.1002	0.6402
	M5	0.0853	0.7544	0.1123	**0.6398**	0.1039	0.7164	0.1010	0.6136
	M6	0.0920	0.7076	0.1138	0.6008	**0.0290**	**0.9808**	**0.0642**	**0.8761**

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

Table 6. Forecasting performance indices of models for Selangor and Bernam River.

	Selangor River				Bernam River			
	Training		Testing		Training		Testing	
Model	RMSE	R	RMSE	R	RMSE	R	RMSE	R
ARIMA	0.0914	0.7055	0.1226	0.5487	0.1049	0.7098	0.1042	0.5842
ANN	0.1044	0.5950	0.1145	0.5682	0.0608	0.9137	0.0714	0.8464
GMDH	0.1101	0.6733	**0.1034**	0.5850	0.0578	0.9216	0.0853	0.8387
LSSVM	0.0961	0.6747	0.1126	0.6269	0.0579	0.9319	**0.0621**	0.8727
GLSSVM	**0.0853**	**0.7544**	0.1123	**0.6398**	**0.0290**	**0.9808**	0.0642	**0.8761**

HESSD

7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

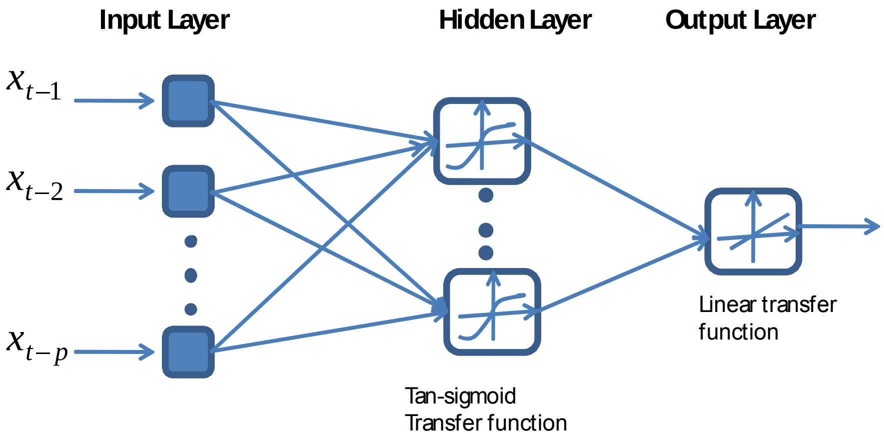

Fig. 1. Architecture of three layers feed-forward back-propagation ANN.

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

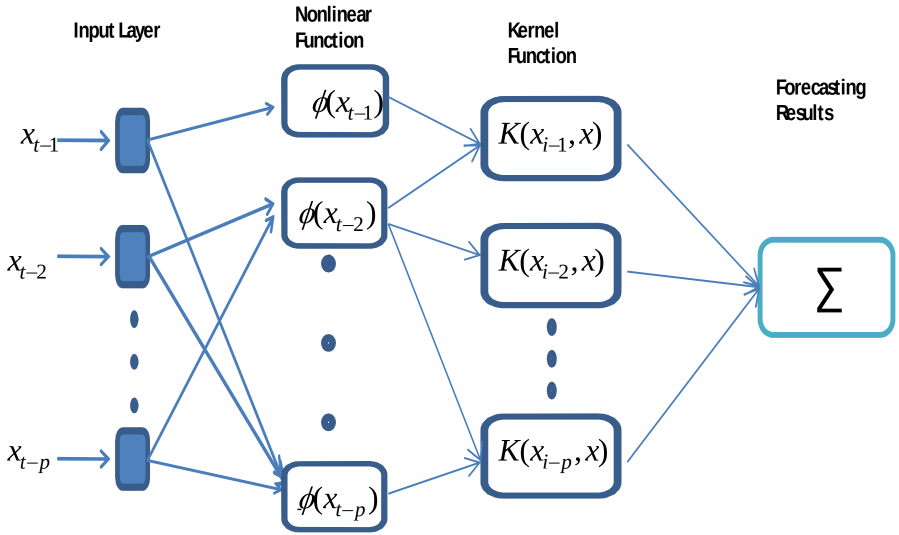

Fig. 2. The structure of a SVM.

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

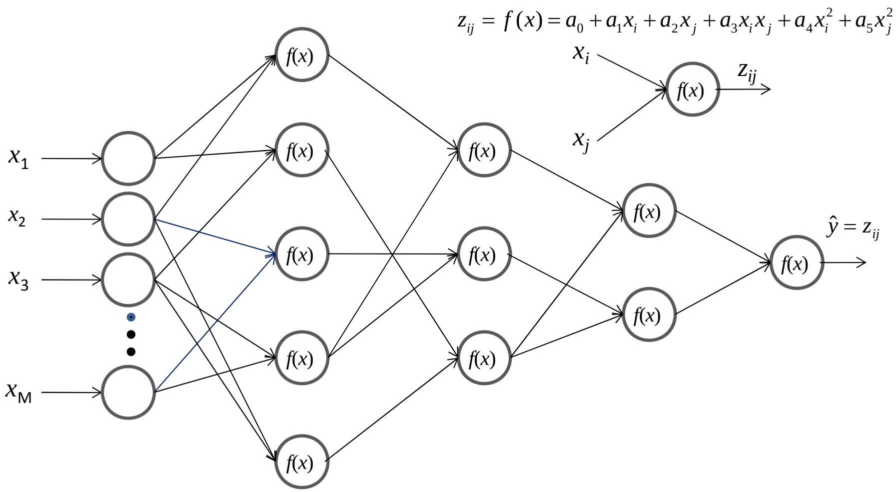

Fig. 3. Basic Structure of Conventional GMDH.

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

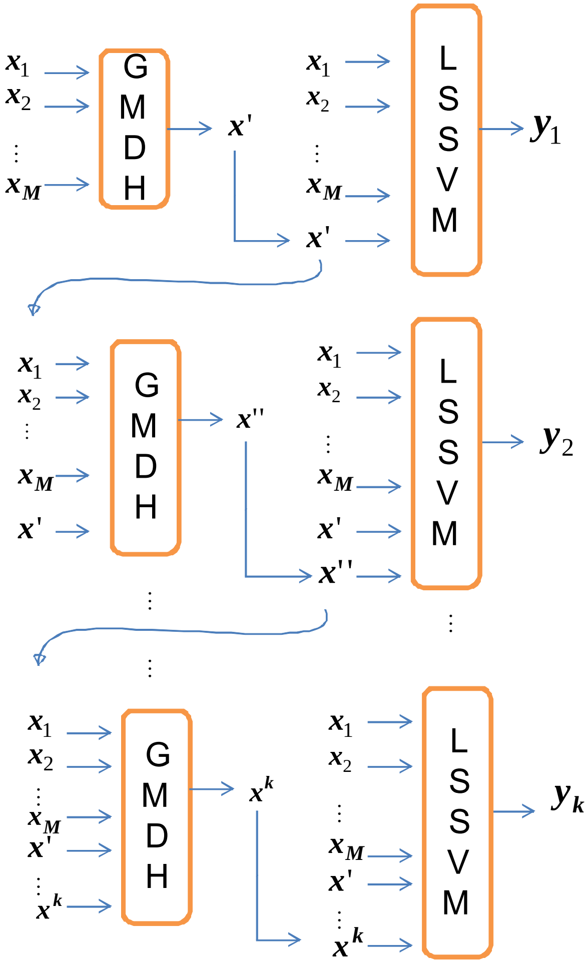

Fig. 4. The structure of the GLSSVM model for time series forecasting.

HESSD
7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

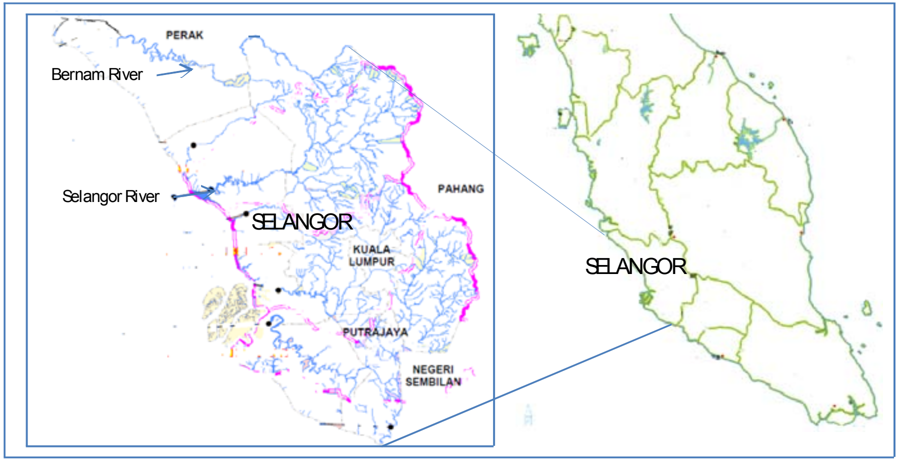

Fig. 5. Location of the Study sites.

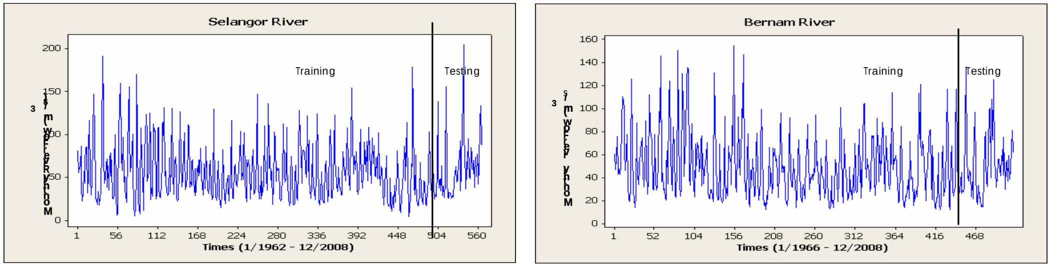

Fig. 6. Time Series of Monthly River Flow of Selangor and Bernam River.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD

7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

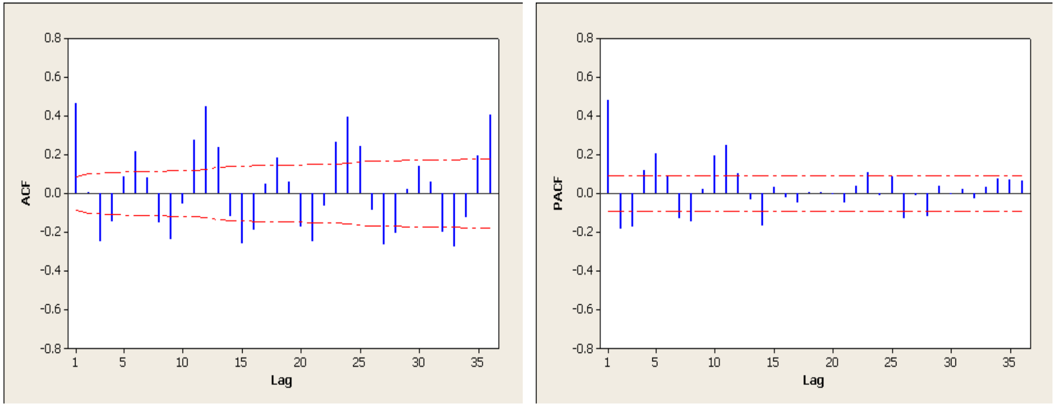

Fig. 7. ACF and PACF of flow series in Selangor River.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD

7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

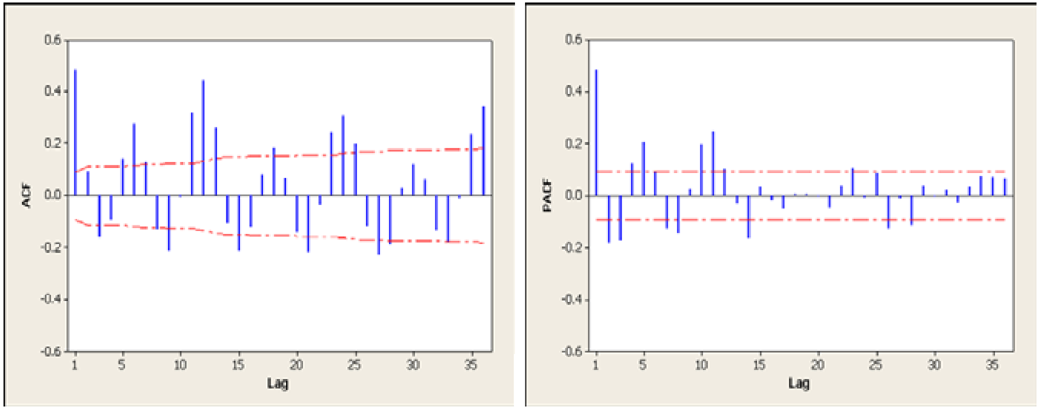

Fig. 8. ACF and PACF of flow series in Bernam River.

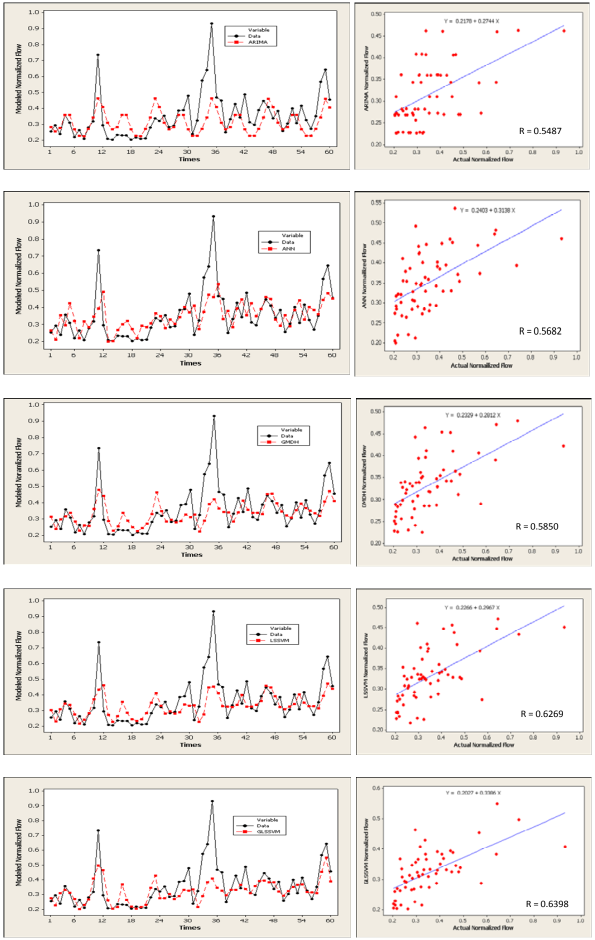

Fig. 9. Time Series and Scatter Plot of the Modeled and Actual Montly Flows for Selangor River (Testing Stage).

HESSD

7, 3691–3731, 2010

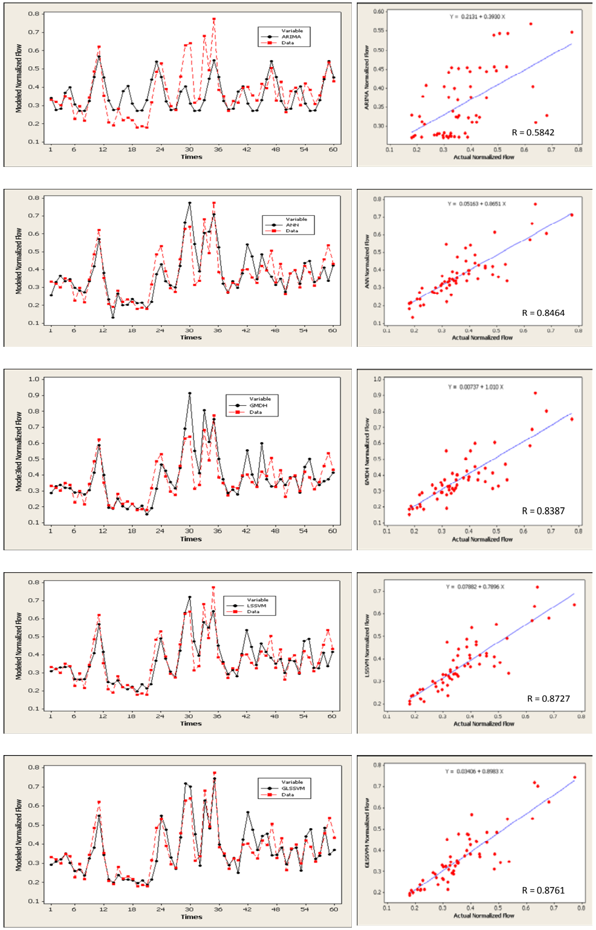

Fig. 10. Time Series and Scatter Plot of the Modeled and Actual Montly Flows for Bernam River (Testing Stage).

HESSD

7, 3691–3731, 2010

A hybrid least squares support vector machines and GMDH approach

R. Samsudin et al.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion