06JEI00063
1726-2135/1684-8799

www.iseis.org/jei

Journal of Environmental Informatics 7 (1) 14-23 (2006)

Support Vector Machines for Environmental Informatics: Application to Modelling the Nitrogen Removal Processes in Wastewater Treatment Systems

Y. H. Yang[1], A. Guergachi[2*] and G. Khan[3]

[1]Citizens' Environment Watch, Toronto, ON, Canada
[2]School of Information Technology, Ryerson University, Toronto, ON, Canada
[3]Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada

ABSTRACT. In order to meet the new stringent environmental regulations, it is necessary to investigate the adaptive and optimal control strategies for the biological wastewater treatment processes. Nitrogen removal is one of the essential concerns in wastewater treatment. Nitrogen removal is a nonlinear, dynamic, and time variant complex process as complicated activities of microbial metabolism are involved. The mechanistic models for nitrogen removal are complicated and still uncertain to some extent. A new machine learning approach, Support Vector Machine (SVM) was proposed as black-box modeling technique to model the biological wastewater treatment processes. LS-SVM, a simplified formulation of SVM, has been applied in this study to predict the concentration of nitrate and nitrite (NO) in the Mixed Liquor (ML) of wastewater treatment plant. Nonlinear Autoregressive model with Exogenous inputs (NARX model) can be employed with LS-SVM to extract useful information and improve the prediction performance. In this paper, the premium wastewater treatment plant simulation and optimization software, GPS-X, is used to create virtual plant layout and simulated data. The simulation results indicate that the proposed method has good generalization performance, especially when the input is fluctuated without a usual pattern. We conclude that LS-SVM with NARX modelling could be used as an alternative approach to predict the behaviour of wastewater treatment systems by further studying some essential issues such as the tuning of memory order and training data size.

Keywords: GPS-X, least-square SVMs, NARX model, nitrogen removal, support vector machines, wastewater treatment

1. Introduction

As the regulations for effluent quality are getting more and more stringent in North America, the advanced biological wastewater treatment techniques, such as the conversion of ammonia nitrogen to nitrate by biological nitrification and the removal of nitrate by biological denitrification, have become essential. To accommodate plant influent fluctuations and other disturbances, there is a need to investigate the development and implementation of adaptive process control strategies, so that more precise and timely controls are achieved for the aforementioned techniques. As the basis of the development of the adaptive controllers, estimating the dynamics of the concentrations of some important trace elements in the effluent is of primary consideration.

Conventionally, mechanistic models have been the most commonly used method for predicting the process dynamics so as to estimate the concentrations of various elements. Many mechanistic models and various control strategies have been incorporated in different software packages to address practical wastewater treatment problems. However, mechanistic models suffer from several fundamental deficiencies that have been discussed extensively by several researchers (Beck, 1986; Marsili-Libelli, 1989; Juppseon, 1996; Beck et al., 1997). A recent synthesis and consolidation of these deficiencies can be found in the article by Guergachi and Patry (Guergachi and Patry, 2003b), and can be recapped in the following fact: mechanistic models are not able to deal with uncertainty in an effective manner. The rationale behind this statement is described next.

Despite the usage of the adjective "*mechanistic*" in the name "mechanistic models" and all the claims that can be made with regard to the inclusion of the first principles of physics and bio-chemistry in the development of the mechanistic models, the latter remain a rough approximation of the dynamic behaviour of the wastewater treatment systems, as many parts of the foregoing models are still highly empirical. While the mechanistic modelling approach aims at implementing the Newtonian thinking in dealing with system complexity, there is so far no possible comparison between the performance of the mechanistic models that have resulted from this approach for wastewater treatment systems and that of the Newton's laws for gravitation or Maxwell's equations for electromagnetism, for example. The question of how far (or close) the predictions of mechanistic models are from reality is still an open-ended one, and the mechanistic modelling approach provides no formal method for addressing it.

One of the main long-term goals of our research is to de-

* Corresponding author: a2guerga@ryerson.ca

velop a framework that allows us to address the deficiencies of the mechanistic modelling approach (Guergachi, 2003; Guergachi and Patry, 2003a). A model is viewed in this framework as a learning machine that acquires knowledge not only from first principles but also from other sources of information such as data and empirical laws. The mechanistic modelling approach becomes a particular case within this framework. Based on this idea, a novel modelling technique has been developed (Guergachi and Patry, 2004) to allow the modellers to make use of the mechanistic thinking and, at the same time, be able to manage the uncertainty that underlies the system under study. This technique aims in developing *not* one single mechanistic model, but a series of nested mechanistic models M_i (i = 1, 2, 3, …) such that $M_1 \subset M_2 \subset ... \subset M_i \subset M_{i+1} \subset ...$

To carry out this work, statistical learning theory (Vapnik, 1998) has been used as the main logical basis for putting the various pieces of the framework together. However, developing novel modelling techniques that are able to handle system uncertainty is not the only objective of building this framework. Another objective is to integrate existing modelling technologies (not only the mechanistic modelling approach) under one single over-arching umbrella that takes advantage of their strengths and minimizes the weaknesses of each modelling technology (Guergachi, 2003). While we are aware of the strengths and weaknesses of many of the exiting technologies such as neural networks, fuzzy logic and knowledge-based systems, there is a novel system modelling approach that has emerged recently as a natural consequence of the results of statistical learning theory and that needs more investigation before one can start thinking about integrating it with other technologies: it is the support vector machine (SVM) approach (Vapnik, 1998). Many researchers and authors have acknowledged this approach, when it is applied to pattern recognition, is a powerful one, and delivers better results than the other traditional and competing approaches such as neural networks do (Terrillon et al., 2000; Scholkopf, 1997). There is, however, a need to carry out more investigations (empirical first and then theoretical) of the performance of the support vector machines (SVMs) in the case of continuous and complex systems such as biological wastewater treatment plants. It is the intention of this paper to present an empirical investigation of SVMs by applying them to the modelling of nitrogen removal in wastewater treatment systems.

The SVM concept was initially introduced by Vapnik (1998), but many variations of SVMs have been developed by other researchers to leverage the strength while overcoming the difficulties in applications of the initial SVM concept. One of these variations known as the least squares SVM (LS-SVM) was introduced by Suykens and co-workers (Suykens et al., 2002). It is this variation that will be investigated in this study. An advantage of the LS-SVM is its simplicity in terms of memory requirements and algorithmic implementation, and the fact that LS-SVM can be used for applications where adaptive and online learning is needed.

In a previous paper (Yang et al., 2004); a simulation study was carried out using 'toy' continuous functions and systems to investigate the performance of LS-SVM for the time-series prediction has shown satisfactory results. However, the nitrogen removal processes in wastewater treatment systems are not as simple as 'toy' systems, because the output depends on many inputs and control variables, as well as on the previous values of the output itself. We proposed to make use of the NARX (Nonlinear Autoregressive model with Exogenous Inputs) modelling concept to handle the time series aspect of the output concentrations. Extensive simulation work using simulated data generated from the wastewater treatment software package GPS-X (Hydromantis, 2004) is carried out to examine how LS-SVM can perform on predicting nitrogen concentrations in treated wastewater. NARX model implemented within the LS-SVMlab MATLAB/C toolbox is used in the estimation of NO concentration.

In section 2 brief introductions of SVM, LS-SVM and NARX are given to provide the basic knowledge. The simulation settings and results are demonstrated in section 3 while section 4 concludes the paper.

2. Support Vector Machines (SVMs)

2.1. Introduction

Machine learning is an artificial intelligence approach to establish and train a model to recognize the pattern or the underlying mapping of a system based on a set of training examples consisting of input and output patterns. SVM is a similar machine learning approach. The simplest support vector machines were developed for binary classification. With continuous extension and advancement, SVMs were applied to functional approximation and time series prediction. Linear learning machines are the fundamental formulations of SVMs. The objective of the linear learning machine is to find the linear function that minimizes the generalization error from a set of functions, which can approximate the underlying mapping between the input and output data. Generalization error is the distance between the true and estimated values on the data point outside the training data set. According to statistical learning theory (Vapnik, 1998), the generalization (test) error can be upper bounded in terms of training error and a confidence term as shown in equation (1):

$$R(\theta) \le R_{emp}(\theta) + \sqrt{\frac{h(\ln(2N/h)+1) - \ln(\eta/4)}{N}} \tag{1}$$

The term on left side represents generalization error. The first term on right side is empirical error calculated from the training data and the second term is called confidence term which is associated with the VC dimension h of the learning machine. VC dimension is used to describe the complexity of the learning system (Vapnik, 1998). The relationship between these three items is illustrated in Figure 1.

Unlike the principle of Empirical Risk Minimization (ERM) applied in neural network which aims to minimize the training error, SVMs implemented Structural Risk Minimization (SRM) in their formulations. SRM principle takes both the

training error and the complexity of the model into account, as finding the minimum of the generalization error shown in Figure 1. SRM can be seen as a nested structure of the learning system. Each element of the structure represents a subset of functions (Suykens et al., 2002). However, most of the practical problems are nonlinear instead of simple linear ones. Kernel functions extended the power of linear learning machine by mapping the input data into a high dimensional feature space. A linear learning machine can be employed in the feature space to solve the original non-linear problem. Kernel functions satisfying Mercer condition not only enable implicit mapping of data from input space to feature space but also ensure the convexity of the cost function which leads to the unique optimum. There are several typical choices of kernels such as linear, polynomial, MLP and RBF kernel. Mercer condition states that a continuous symmetric function *K(x, z)* must be positive semi-definite to be a kernel function which can be written as inner product between the data pairs as in equation (2) (Cristianini and Shawe-Taylor, 2003):

$$K(x,z)=\sum_{j=1}^{\infty}\lambda_j\varphi_j(x)\ \varphi_j(z)=\ [\phi(x),\ \phi(z)] \tag{2}$$

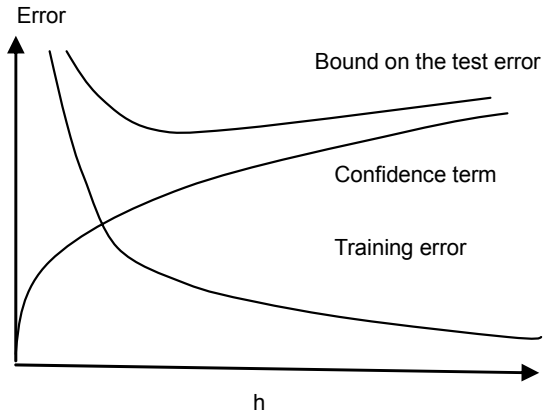

Figure 1. Upper bounded generalization error.

In order to minimize a cost function which takes into account both empirical error and complexity of the learning machine, SVM formulations were established as a constrained optimization problem shown in the formulation (3):

$$\min_{\omega,b,\xi,\xi^*} J_P(\omega,\xi,\xi^*)=\frac{1}{2}\omega^T\omega+c\sum_{k=1}^{N}(\xi_k+\xi_k^*)$$

$$\begin{aligned} &y_k-\omega^T\varphi(x_k)-b\le\varepsilon+\xi_k, k=1,...,N\\ &\omega^T\varphi(x_k)+b-y_k\le\varepsilon+\xi_k^*, k=1,...,N\\ &\xi_k,\xi_k^*\ge 0, k=1,...,N. \end{aligned} \tag{3}$$

where C is a regularization parameter which adjusts the balance of the training error and the complexity of the system. This formulation is named the Vapnik's ε-insensitive cost function due to ignoring the error within a band ε around the target function. The ξ_k and ξ_k^* are slack variables; ω is a weight vector of the linear learning machine and follows from the solution of the optimization problem in (3); $\varphi(x)$ is the mapping from original input space to feature space. The target function is expressed in equation (4):

$$f(x)=w^T\varphi(x)+b \tag{4}$$

where b can be calculated using the complementary conditions (Suykens et al., 2002)

This formulation is in primal weight space as a constrained optimization problem. Lagrangian technique is used to solve this problem into dual space of Lagrangian multipliers. The dual problem is described in the formulation (5):

$$\begin{aligned} \max_{\alpha,\alpha^*} J_D(\alpha,\alpha^*)=&-\frac{1}{2}\sum_{k,l=1}^{N}(\alpha_k-\alpha_k^*)(\alpha_l-\alpha_l^*)K(x_k,x_l)\\ &-\varepsilon\sum_{k=1}^{N}(\alpha_k+\alpha_k^*)+\sum_{k=1}^{N}y_k(\alpha_k-\alpha_k^*) \end{aligned}$$

$$\sum_{k=1}^{N}(\alpha_k-\alpha_k^*)=0$$

$$\alpha_k,\alpha_k^*\in[0,c] \tag{5}$$

The non-zero Lagrangian multipliers α_k are called support vectors. Only support vectors are involved in the expression of target function as in equation (6):

$$f(x)=\sum_{k=1}^{N}(\alpha_k-\alpha_k^*)K(x,x_k)+b \tag{6}$$

The advantage of using the dual representation is derived from the fact that in this representation the number of tuneable parameters (equals the dimension of weight vector) does not depend on the number of dimensions of the input space. In the formulation of Lagrangian, the training examples never appear isolated but always in the form of inner products between pairs of examples. By replacing the inner product of input data pairs with an appropriately chosen 'kernel' function, one can avoid to explicitly establish a non-linear mapping to a high dimensional feature space. In the feature space, the size of the model is determined by the number of support vectors and the difficulty that was caused by infinite dimension of the input space is avoided.

SVMs possess the properties of global optimum, sparseness of support vectors and bounded generalization risk, which can be derived directly from the solution of the optimization problem or the statistical learning theory (Cristianini and Shawe-Taylor, 2003). These properties make SVMs advanta-

geous compared to Neural Networks that suffer the local optimum and over-fitting problem due to using ERM principle.

2.2. Least-Squares SVM (LS-SVM)

Standard SVMs have been developed for classification and regression, while least squares support vector machines (LS-SVMs) can tackle a wider range of problems (including kernel ridge regression, classification (kernel Fisher discriminant analysis), kernel PCA, kernel CCA, kernel PLS, recurrent networks, optimal control, and other) which is possible via the equality constraints instead of inequality constraints and the use of a simpler loss function (Suykens et al., 2002). LS-SVM is a variation of Vapnik's SVM and is expressed as in the formulation (7) (Suykens and Vandewalle, 1999):

$$\min_{\omega,b,e} J_P(\omega,e) = \frac{1}{2}\omega^T\omega + \gamma\frac{1}{2}\sum_{k=1}^{N}(e_k^2)$$

Such that

$$y_k = \omega^T\varphi(x_k) + b + e_k, k = 1, \ ..., N \tag{7}$$

In fact, this is a ridge regression (Cristianini and Shawe-Taylor, 2003) cost function formulated in the feature space. γ plays the same role as the regularization parameter C in SVM formulation. This LS-SVM formulation modifies Vapnik's SVM at two points. First, LS-SVM takes equality constraints instead of inequality constraints. Second, the error variable e_k was introduced in the sense of least-square minimization. These error variables play similar role as the slack variables in SVM formulation such that relatively small errors can be tolerated (Suykens et al., 2002).

In the case of linear function approximation one could easily solve the primal problem, but in general ω might become infinite dimensional and difficult to solve. The solution is to derive the dual problem by constructing Lagrangian for this primal problem. Taking the condition for optimality of the Lagrangian yields a set of linear equations shown in equation set (8):

$$\begin{aligned}
&\frac{\partial L}{\partial \omega} = 0 \rightarrow \omega = \sum_{k=1}^{N}\alpha_k\varphi(x_k) \\
&\frac{\partial L}{\partial b} = 0 \rightarrow \omega = \sum_{k=1}^{N}\alpha_k = 0 \\
&\frac{\partial L}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k, \ k = 1, \ 2, ..., \ N \\
&\frac{\partial L}{\partial \alpha_k} = 0 \rightarrow \omega^T\varphi(x_k) + b + e_k - y_k = 0, \ k = 1, \ 2, ..., \ N
\end{aligned} \tag{8}$$

Solving this set of linear equations in α, b, the resulting LS-SVM model for function approximation becomes equation (9):

$$y(x) = \sum_{k=1}^{N}\alpha_k K(x, x_k) + b \tag{9}$$

As it was shown in the previous section, SVMs solve the nonlinear regression problems by means of convex quadratic programs (QP). The use of least squares and equality constraints for the models leads to solving a set of linear equations, which is easier to use than QP solvers. But on the other hand it has potential drawbacks such as the lack of sparseness which is clear from the condition $\alpha_k = \gamma\ e_k$ in equation set (8) since the error would not be zero for most of data points. One can overcome the drawbacks using special pruning techniques for sparse approximation (Suykens et al., 2002).

2.3. NARX Model

Basically, SVM is used for classification or function approximation, which involves mapping of multi-dimensional input and output. Time series prediction is to predict one or more variables in the future point in time. From the inspective of machine learning, time series prediction is a special case of function estimation and equivalent to find the underlying functional relationship between previous values and the next value. Thus, SVMs can be used in time series prediction in the form as given in equation (10):

$$\hat{y}_{k+1} = f(y_k, y_{k-1}, \ ..., y_{k-p}) \tag{10}$$

where p is referred as embedding dimension or memory order in time series. Note that the value of p determines the dimension of inputs of the SVM model.

One of the deficiencies of time series prediction is that it is unable to accommodate other meaningful input variables in system identification or dynamic modelling. An important and useful class of discrete-time nonlinear model is the NARX model (Nonlinear Auto Regressive model with Exogenous Inputs) (Lin et al., 1997).

$$y(t) = f[u(t - Du), \ ..., \ u(t-1), \ u(t), \ y(t - Dy), \ ..., \ y(t-1)] \tag{11}$$

where $u(t)$ and $y(t)$ represent input and output of the model at time t, D_u and D_y are the input-memory and output-memory order respectively, and the function f is a nonlinear function.

NARX model combines the power of function approximation and time series prediction. The embedded memory of the input and output variables plays an important role in the learning capability and the generalization performance through incorporating the historical information. The selection of input-memory and output-memory is critical for the forecasting performance. "The problem of choosing the proper memory architecture corresponds to giving a good representation of input data. A good representation can make useful information explicit and easy to extract" (Lin et al., 1996, 1997).

3. Simulations

3.1. Simulation Settings

The target problem of our study is to predict the concentration of indicator variables in nitrogen removal processes, such as Ammonia, Nitrate and Nitrite (NO), in a short term. When the concentration of an indicator variable is predicted to exceed the standards for discharging treated wastewater, the system acts and adjusts the input of chemicals or operation conditions in order to optimally control the biological wastewater treatment processes.

As discussed in section 2, in terms of simplicity and memory requirement, LS-SVM has more advantages than standard SVM. LS-SVM can be implemented by adaptive and on-line algorithm. For these reasons, we proposed to use LS-SVM combined with NARX model to predict the target variable. LS-SVMlab MATLAB/C toolbox was used to train the sample data and predict the future output (Suykens et al. 2002). NARX model was employed to transform the input and output into more suitable state space in order to extract the information effectively.

3.2. Data Collection

In order to obtain the training and testing data, we selected to use COST simulation benchmark. COST simulation benchmark is a comprehensive description in terms of simulation and evaluation procedures including plant layout, simulation models and model parameters, a detailed description of disturbance to be applied during testing. A complete description of the COST benchmark and how to use the simulation benchmark layout can be found in literature (Copp et al., 2002). COST simulation benchmark has been implemented in GPS-X. The 'simulation benchmark' plant design comprises five reactors in series with 10-layer secondary settling tank. The first two anoxic tanks, the following three aerobic tanks as well as the internal recycle from the fifth to the first tank are designed in order to realize nitrification. We focused on predicting the Nitrate and Nitrite (NO) concentration in the effluent of the reactor (MLSS). The plant layout in GPS-X is shown in Figure 2.

From the knowledge of the nitrogen removal process in wastewater treatment plant, the input variables are selected as influent flow rate i.e. the concentration of TSS, COD, TKN and TN in the influent. As the plant is operated with dissolved oxygen (DO) controller and nitrate controller, the oxygen transfer coefficient (KLa) in the final tank and the controlled nitrate value in the second anoxic tank are also used as input. The output is selected as the concentration of NO in MLSS.

Three different influent files are included in the COST simulation benchmark implemented in GPS-X and each is meant to be representative of dry, rain or storm weather condition respectively. The influent files include the data of influent flow rate and influent composition. The concentrations of TSS, COD, TKN and TN in the influent can be obtained as composite variables in GPS-X. Each of the influent file contains 14 days (2 weeks) of influent data at 15-minute intervals. In general, these files depict expected diurnal variations in influent flow. Additionally, expected trends in weekly data have been incorporated. This means, much lower peak flows are depicted in the 'weekend' data, which is consistent with normal load behaviour at a municipal treatment facility (Copp et al., 2002). The influent flow rates under three weather conditions are illustrated in Figures 3(a) to (c). In the first week, all of the three files contain dry weather data. In Figure 3(a), the dry weather influent flow rate depicts what is considered to be normal diurnal variations in flow. Figure 3(c) is a variation on the dry weather with the incorporation of two storm events in the second week. The first storm event is of high intensity and short duration. The peak flow for both storms is the same while

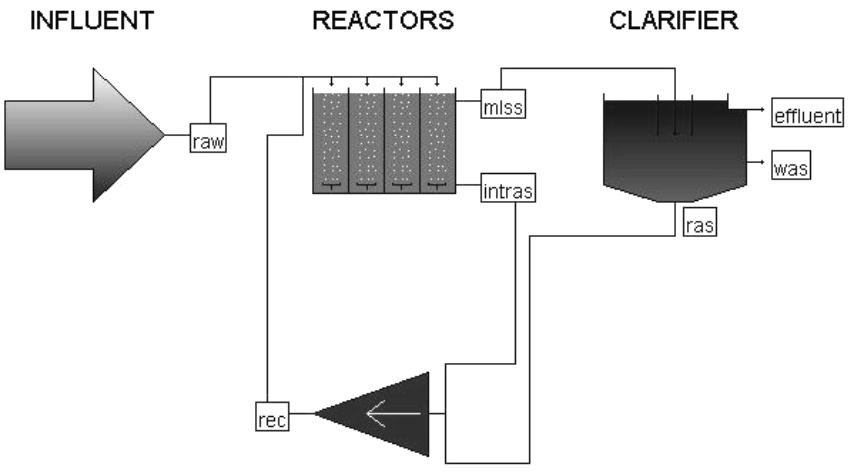

Figure 2. COST simulation benchmark plant layout.

the peak flow of the second storm is maintained over a longer period of time. Figure 3(b) represents a long rain event occurring in the second week. The influent flow during this rain event does not reach the level attained during the storm events, but the increased flow is sustained for a much longer period of time. What we need to do is to simply execute the simulation benchmark using three influent disturbances and record the corresponding inputs and outputs. The recorded output, the concentrations of NO under three weather conditions are depicted in Figures 4 (a) to (c). It indicates that the concentration of NO is mainly maintained around 10-15 gN/m^3. It can be observed that the concentration of NO under rain or storm event is reduced because the wastewater is diluted by the excess water flow due to the rain or storm.

3.3. Simulation Results

Now that we have obtained the simulated input and output data, we can use this data as the training and testing data for the LS-SVM prediction. The simulation approach of the LS-SVM includes five steps: input selection, model selection, training, prediction and result visualization. LS-SVMlab MATLAB/C toolbox provides a rich class of functions to realize these steps. In this paper, we applied LS-SVM combined with NARX model to the prediction of NO concentration in MLSS. The input variables have been selected according to the domain knowledge as given in section 3.2. RBF kernel is used and the kernel parameters can be tuned using cross-validation in LS-SVM toolbox.

Table 1. Prediction Error of NO Concentration under Dry Weather (g N m^{-3})

Output memory order	Input memory order			
	0	1	2	3
0	0.5204	0.3624	0.3104	0.2827
1	0.2793	0.1550	0.1208	0.1049
2	0.3338	0.1924	0.1423	0.1112
3	0.2473	0.1693	0.1389	0.1166

The emphasis of this paper is on exploring the effects of the input and output memory order of NARX model on the performance of LS-SVM prediction under different weather conditions. A total of 672 training data examples over the first week (representing dry weather) was used to train the LS-SVM with RBF kernel and predict the next 672 values of NO concentration over the second week for three different weather conditions (representing dry weather, rain event and storm event respectively). The Mean Square Error (MSE) was used to measure the prediction accuracy. In Table 1-3, the prediction errors with different input and output memory orders are listed for three weather conditions. These results indicate that the LS-SVM using input memory order-3 and output memory order-1 has the best performance for predicting NO concentration under the dry weather. Under rain or storm event, the optimal input and output memory order are both 1.

Table 2. Prediction Error of NO Concentration under Rain Event (g N m^{-3})

Output memory order	Input memory order			
	0	1	2	3
0	8.0022	7.2378	6.5409	6.1016
1	4.6539	1.7146	2.5735	3.3245
2	6.0985	2.2491	2.0145	2.3592
3	6.0974	2.7482	2.4547	2.4989

Table 3. Prediction Error of NO Concentration under Storm Event (g N m^{-3})

Output memory order	Input memory order			
	0	1	2	3
0	5.0235	4.3416	4.0412	3.9068
1	2.6405	2.0437	2.6062	3.1143
2	3.532	2.5308	2.6585	3.0345
3	3.7025	2.8422	2.9338	3.1949

Figures 5(a) to (c) show the comparison of predicted and actual NO concentrations using the optimal input and output memory order under three weather conditions. The solid line represents the actual concentration and the dashed line the predicted value.

4. Conclusions

In a short term period, the LS-SVM model with RBF kernel and optimized parameters provides reasonably accurate prediction of the NO concentration in ML. From the simulation results, the generalization ability of LS-SVM in combination with NARX model is evident. We use only one week data representing dry weather condition as the training data. Since NARX model takes account of both the historical data of the input and output and the current influent disturbance, the NO concentration in MLSS under various weather conditions can be predicted, given appropriately selected parameters. As we can see from the simulation results, given the influent disturbance, the response of dry, rain and storm weather condition can be predicted using the same LS-SVM model trained by one-week dry weather data.

In this paper, NARX model was proposed and experimented to be effective in transforming the input and output into new state space in order to extract useful information. The simulation results are consistent with the embedding theory stating that forecasting performance could be seriously defi-

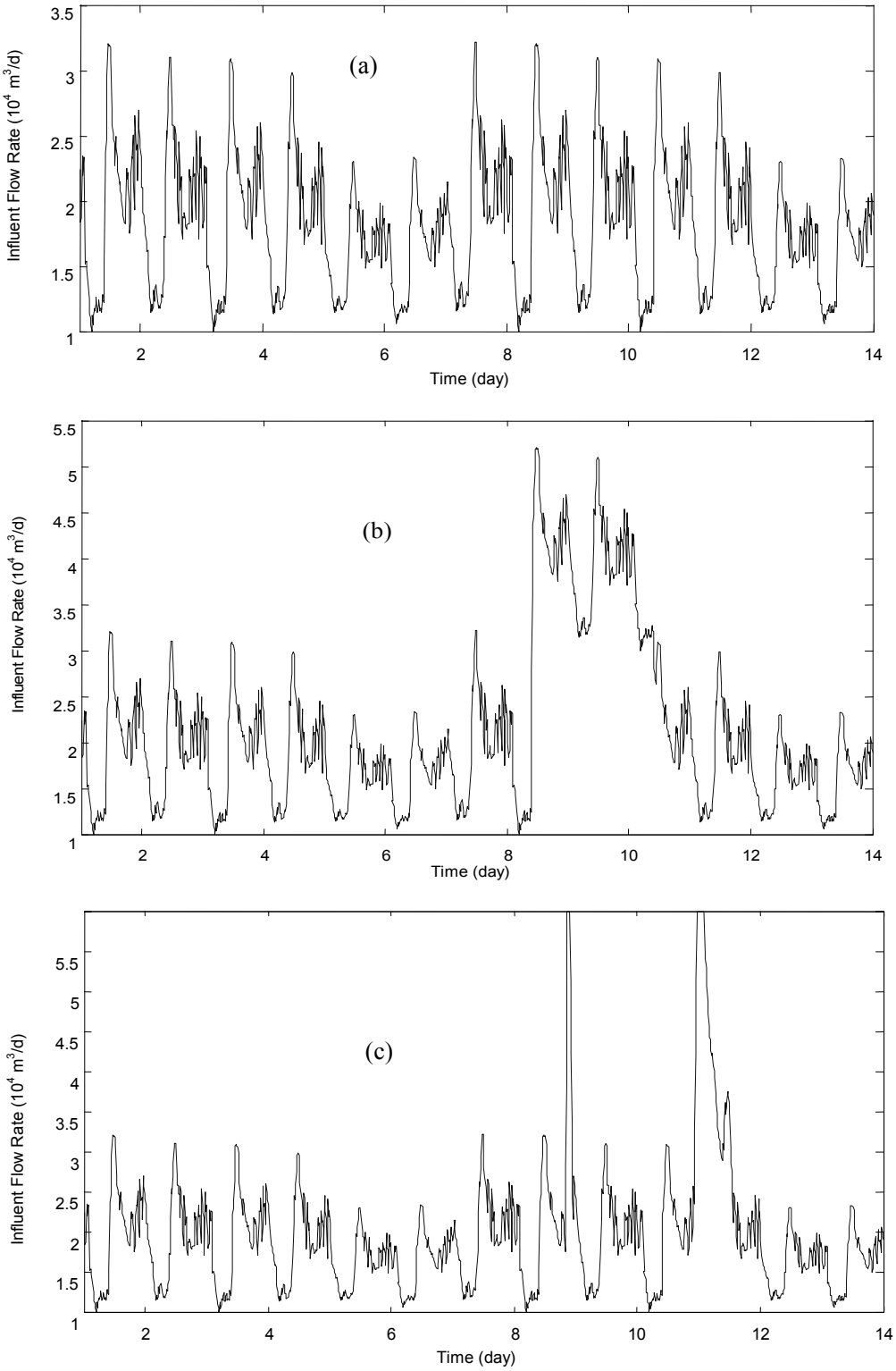

Figure 3. Influent flow rate under (a) dry weather, (b) rain event, and (c) storm event (m^3/d).

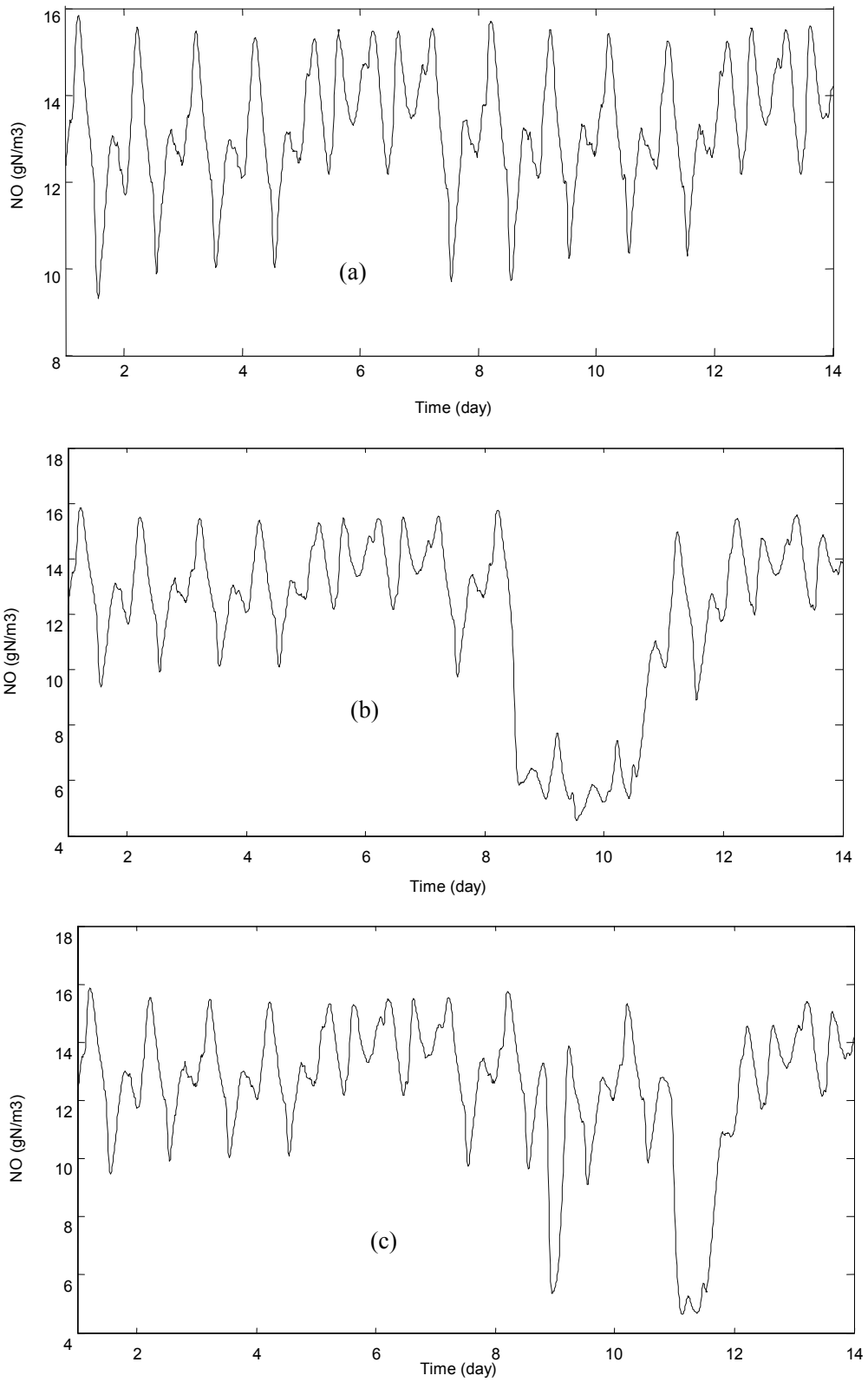

Figure 4. NO concentration in MLSS under (a) dry weather, (b) rain event, and (c) storm event (g N m^{-3}).

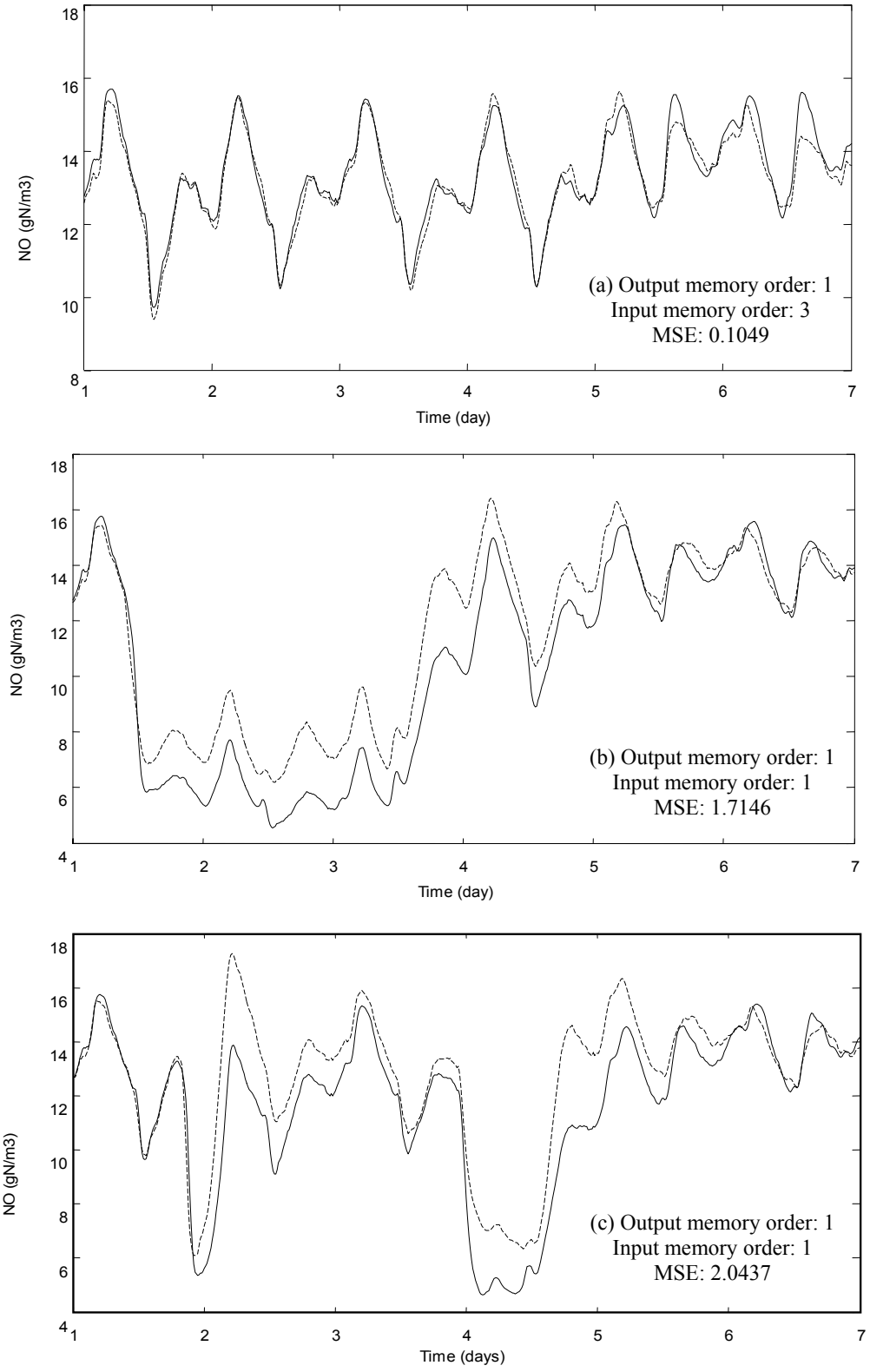

Figure 5. Comparison of predicted (dashed line) and actual (solid line) NO concentrations in MLSS under (a) dry weather, (b) rain event, and (c) storm event (g N m^{-3}).

cient if a model's memory order is either too little or too large. Therefore, choosing the appropriate memory architectures for a given task is a critical issue in NARX models. Our next study will investigate how to determine the appropriate memory order automatically. The prediction can be extended to other elements such as ammonia in further studies.

Acknowledgments. This research has been financially supported by research and equipment grants from NSERC and CFI respectively. This financial support is highly acknowledged. The helpful discussions with J.B. Copp and O. Schraa in Hydromantis Inc. are also gratefully appreciated. The comments from the anonymous reviewers are acknowledged that has improved this paper.

References

Beck, M.B. (1986). Identification, estimation and control of biological wastewater treatment processes. *Proc. IEE*, 133, 254-264.

Beck, M.B., Ravetz, J.R., Mulkey, L.A. and Barnwell, T.O. (1997). On the problem of model validation for predictive exposure assessments. *Stochastic Hydrol. Hydraul.*, 11, 229-254.

Copp, J.B. (2002). *The COST Simulation Benchmark: Description and Simulator Manual*, A Product of COST Action 624 and COST Action 682, Directorate-General for Research.

Cristianini, N. and Shawe-Taylor, J. (2003). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, Cambridge.

Guergachi, A. and Patry, G.G. (2002). Statistical learning theory, model identification and system information content. *Int. J. Gen. Syst.*, 31(4), 343-357.

Guergachi, A. and Patry, G.G. (2003a). Using statistical learning theory to rationalize system model identification and validation, Part I: Mathematical foundations. *Complex Syst. J.*, 14(1), 63-90(c).

Guergachi, A. and Patry, G.G. (2003b). Identification, verification and validation of process models in wastewater engineering: A critical review. *J. Hydro-inf.*, 5(3), 181-188.

Guergachi, A. (2003). Integrating domain knowledge and machine learning theoretic methods for modelling complex environmental systems: methodology and rationales, in *Proc. of the ISEIS 2003 International Conference on Environmental Informatics*, Regina, Canada, pp. 386-393.

Guergachi, A. And Patry, G.G. (2004). Constructing a model hierarchy with background knowledge for structural risk minimization. *IEEE Trans. Syst., Man Cybern., Part A*, (accepted).

Hydromantis (2004). GPS-X4.1.2, a modular, multi-purpose computer program for the modelling and simulation of large-scale wastewater treatment plants. http://www.hydromantis.com.

Jeppsson, U. (1996). *Modelling Aspects of Wastewater Treatment Processes*, Ph.D. Dissertation, Department of Industrial Electrical Engineering and Automation, Lund Institute of Technology, Lund, Sweden.

Lin, T.N., Horne, B.G., Tino, P. and Giles, C.L. (1996). Learning long-term dependencies in NARX recurrent neural networks. *IEEE Trans. Neural Networks*, 7(6), 1329-1338.

Lin, T.N., Giles, C.L., Horne, B.G. and Kung, S.Y. (1997). A delay damage model selection algorithm for NARX neural networks. *IEEE Trans. Signal Process.*, 45(11), 2719-2730.

Marsili-Libelli, S. (1989). Computer Control of the Activated Sludge Process. *Encyclopedia Environ. Control Technol.-Wastewater Treat. Technol.*, 3, 229-270.

Metcalf and Eddy, Inc. (1991). *Wastewater Engineering: Treatment, Disposal, and Reuse*, 3rd Edition, McGraw-Hill, New York, USA.

Scholkopf, B., Sung, K., Burges, C.J.C., Girosi, F., Niyogi, P., Poggio, T. and Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.*, 45(11), 2758-2765.

Suykens, J.A.K., Van Gestel, T., Brabanter, J.D., Moor, B.D. and Vandewalle, J. (2002). *Least Squares Support Vector Machines*, World Scientific, Singapore.

Suykens, J.A.K. (2001). Support Vector Machines: a nonlinear modelling and control prospective. *Eur. J. Control*, Special Issue on Fundamental Issues in Control, 7(2-3), 311-327.

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

Terrillon, T.J., Shirazi, M.N., Sadek, M., Fukamachi, H. and Akamatsu, S. (2000). Invariant face detection with support vector machines, in *Proc. of Fifteenth International Conference on Pattern Recognition*, Barcelona, 4, pp. 210-217.

Van Gestel, T., Suykens, J.A.K., Baestaens, D.E., Lambrechts, A., Lanckriet, G., Vandaele, B., De Moor, B. and Vandewalle, J. (2001). Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Trans. Neural Networks*, 12(4), 809-821.

Vapnik, V.N. (1998). *Statistical Learning Theory*, John Wiley & Sons, New York, USA.

Yang, Y.H., Guergachi, A.A. and Khan, G.N. (2004). Explore the performance of time series prediction using least square support vector machines, in *Proc. of third International Conference for Upcoming Engineer*, Toronto, ON, Canada.