Comment on 'Rainfall and runoff forecasting with SSA-SVM approach'

B. Sivakumar

B. Sivakumar
Department of Land, Air & Water Resources,
University of California, Davis,
Davis,
California 95616,
USA
Tel.: +1-530-752-8577
Fax: +1-530-752-5262
E-mail: *sbellie@ucdavis.edu*

INTRODUCTION

Applications of machine learning techniques, such as nonlinear dynamical methods (e.g. chaos theory), artificial intelligence methods (e.g. artificial neural networks or ANNs) and evolutionary algorithmic methods (e.g. genetic programming or GP), have been gaining considerable attention in hydrology. Studies that have employed such methods yielded encouraging results on the characterization, modeling and forecasting of a variety of hydrological phenomena (see, for instance, Sivakumar (2000) and ASCE Task Committee (2000) for reviews of such studies). The study by Sivapragasam *et al.* (2001) attempts the application of another new method, the Support Vector Machine (SVM), for forecasting two hydrological time series: (1) the daily rainfall data observed in one of the stations (Station No. 23) in Singapore; and (2) the daily runoff data observed at the Tryggevaelde catchment in Denmark. The forecasts are made for both the raw data and the pre-processed data (using the Singular Spectrum Analysis (SSA)). These forecast results are then compared with those achieved (for the raw data) using the nonlinear prediction (NLP) method, which uses the concept of phase-space reconstruction. Based on such a comparison, Sivapragasam *et al.* (2001) report that the SSA-SVM method is a significantly better forecasting method than the NLP method. The purpose of the present comment is to point out some of the possible inconsistencies in their study and, hence, the results reported and conclusions drawn, and also to suggest possible ways to avoid such inconsistencies.

DISCUSSION OF RESULTS OF SIVAPRAGASAM *et al.* (2001)

As clearly mentioned in the Abstract and elsewhere by Sivapragasam *et al.* (2001), the essence of their study is the proposal of a simple and efficient prediction technique based on SSA (i.e. for pre-processing the data) coupled with SVM method. As reported by Sivapragasam *et al.* (2001), the SVM forecasts achieved for the pre-processed rainfall and runoff data are significantly better than those achieved for the raw data (see Tables 1 and 2 and Figure 4). In view of this, the validity of one of the conclusions drawn regarding the usefulness of pre-processing of the data in the SVM method is acknowledged and appreciated. However, based on the forecast results achieved, Sivapragasam *et al.* (2001) also provide another important conclusion regarding the better forecasting method between the SVM and the NLP methods. They claim that the SSA-SVM method is significantly better than the NLP method for rainfall and runoff forecasting. The validity of such a conclusion is questionable for the reasons discussed below.

The discussion herein on the possible inconsistencies in the analysis carried out and the results reported by Sivapragasam *et al.* (2001) is based essentially on three aspects: (1) the types of data and forecast results used for comparing the SVM and the NLP methods; (2) the pre-processing of the raw data (using SSA) only in the SVM method; and (3) the forecasting approach adopted in the

NLP method. In addition to these, some secondary issues are also discussed.

Types of data and forecast results used for comparing the SVM and NLP methods

A conspicuous inconsistency in the analysis carried out by Sivapragasam *et al.* (2001) is the data used for comparing the forecast results achieved using the SVM and the NLP methods, and towards identifying the better method. The importance of this point lies in the fact that such a comparison indeed forms the basis for one of the major conclusions drawn therein. In case of the SVM method, forecasts are made for both the raw data and the pre-processed data (using SSA), whereas in the NLP method forecasts are made only for the raw data. However, the conclusion on the better forecasting method is made based on the comparison of the forecast results achieved for the pre-processed data in the SVM method and those achieved for the raw data in the NLP method. Comparing the forecast results achieved for one type of data (e.g. raw data) in one method (e.g. NLP) and another type of data (e.g. pre-processed data) in another method (e.g. SVM) and deriving conclusions based on such a comparison is clearly misleading and probably wrong. The right approach should be to compare the forecast results achieved for the same type of data, i.e. *either* raw data *or* pre-processed data, in both the methods. Such a comparison, for instance, of forecasts for the raw data in SVM against forecasts for the raw data in NLP (Tables 1 and 2 and Figure 4(a) and (c)) reveals an entirely different picture, as the forecasts achieved for the raw data using the NLP method are significantly better than (in case of rainfall) or almost similar to (in case of runoff) the forecasts achieved for the raw data using the SVM method (see Tables 1 and 2 for details). It should be noted that a similar observation (i.e. the better performance of the NLP method compared to the SVM method) might also be possible if the pre-processed data were used for forecasting in the NLP method. Such observations (could) certainly alter the conclusions drawn by Sivapragasam *et al.* (2001) regarding the better forecasting method.

Pre-processing of the raw data (using SSA) only in the SVM method

Sivapragasam *et al.* (2001) use the SSA (usually seen as an adaptive noise-reduction algorithm) as an efficient pre-processing algorithm for modifying the representation of the input vectors. The SSA is first used to pre-process the raw rainfall and runoff data. The pre-processed data are then used (only) in the SVM method for forecasting purposes.

It is relevant to note that the SSA is a general data pre-processing procedure and, therefore, may be coupled with any forecasting method. In view of this, the coupling of SSA *only* with the SVM method for pre-processing and subsequent forecasting raises a serious concern. The importance of such a concern lies in the fact that, as reported by Sivapragasam *et al.* (2001), the forecast results for the raw data using the NLP method are significantly better than those using the SVM method (see Tables 1 and 2 and Figure 4). In other words, the coupling of SSA with the NLP method for pre-processing and forecasting could in all probability provide better forecasts than the forecasts achieved using the SSA-SVM method. However, Sivapragasam *et al.* (2001), by not attempting the coupling of SSA with the NLP method, fail to look into the possibility of achieving still better forecast results for the (pre-processed) rainfall and runoff data than those achieved using the SSA-SVM method.

The reasons for not using the SSA pre-processed data for forecasting in the NLP method are not clear. The failure to attempt such a task raises serious concerns, particularly when the pre-processed data are already available and used in the SVM method and, therefore, the above task can be easily carried out. As the main objective of the study by Sivapragasam *et al.* (2001) is the proposal of a simple and efficient forecasting technique by coupling SSA and SVM method, the forecasting of the raw data as well as the SSA pre-processed data using the SVM method and their comparison are understandable. However, there is no basis for comparing the forecast results achieved for the SSA pre-processed data using the SVM method and the results achieved for the raw data using the NLP method, as the data under study in the two methods are not consistent with each other. Sivapragasam *et al.* (2001)

may argue that the phase-space reconstruction employed in the NLP method may also be considered as a pre-processing procedure. If this is the case, then the same procedure may also be employed in the SVM method so that the procedures in both the methods are consistent, except the forecasting algorithm. In short, the study by Sivapragasam *et al.* (2001) neither recognizes the inconsistency of the data used in the two methods, nor provides any explanation for using different data for comparing the forecast results achieved using the SSA-SVM and the NLP methods. All these observations seem to suggest a possible *a priori* assumption on the part of Sivapragasam *et al.* (2001) that the SSA-SVM method is certainly better than the NLP method for forecasting purposes (see below in the section 'Secondary issues' for details). The reasons for such a possible *a priori* assumption, if any, are unfounded for the reasons stated above (and also below); in fact, the results reported by Sivapragasam *et al.* (2001) indicate that the opposite could be true.

Forecasting approach adopted in the NLP method

The usual procedure in the local approximation approach, proposed by Farmer & Sidorowich (1987), reportedly employed by Sivapragasam *et al.* (2001) in the NLP method is as follows. The entire time series available is divided into two parts: (1) training or learning set; and (2) verification or testing set. The phase-spaces are reconstructed using the values in the first part (i.e. training set), by embedding the values in different embedding dimensions, and then forecasts are made for the second part (i.e. verification set). Forecasts are made by identifying the nearest neighbor(s) to the vector of interest (i.e. the last vector in the training set) in the phase-space and averaging the evolutions of such neighbor(s). In other words, forecasts are made *only* for the verification set (not for the training set) and the forecast accuracy is evaluated by comparing the forecasted values with the original values in the verification set. In regards to this, the forecast results presented by Sivapragasam *et al.* (2001) seem to be inconsistent with the local approximation approach (Farmer & Sidorowich 1987) employed therein, as they present forecast accuracy for both the training set and the verification set (see Tables 1 and 2).

In view of the above, either of the following two possibilities emerges: (1) the local approximation approach employed by Sivapragasam *et al.* (2001) is not the same as the one proposed by Farmer & Sidorowich (1987); and (2) the training set is further divided into two sub-parts (one sub-part for training and the other for verification). However, neither of these two possibilities seems to be the case, as such details are not made available by Sivapragasam *et al.* (2001). These observations raise further questions on the reported results.

Another important observation that can be made from the results achieved using the NLP method is the clear (one-day) lag between the forecasted and the observed values throughout the forecasting period (or verification set), as can be seen from the direct time series comparison presented in Figure 6. Such a consistent lag between the forecasted and the observed values does not seem to be a possible result of the forecasting approach (see, for instance, Sivakumar *et al.*, 2000 (Figure 2) and Sivakumar *et al.*, 2001 (Figure 5)); rather it seems to be an error resulting from possible mis-handling of the training and the verification samples. Solving this problem could lead to significant improvements in the forecast accuracies (Correlation coefficient and RMSE values) achieved using the NLP method than those reported (Tables 1 and 2) by Sivapragasam *et al.* (2001). A possible implication of such improvements is that the forecasts from the NLP method could be significantly better than (in case of rainfall) or closely comparable to (in case of runoff) the forecasts from the SVM method than as presently reported by Sivapragasam *et al.* (2001).

Secondary issues

Sivapragasam *et al.* (2001) report (pp. 142 and 144) that the SVM method generally provides better results than the NLP and the ANN methods in forecasting. To support their claim, they quote the studies by, for instance, Mukherjee *et al.* (1997), Babovic *et al.* (2000), and Liong & Sivapragasam (2002). This could be one possible reason for their inclination to prove that the SVM method is better than the NLP method and hence, the use of SSA only in the SVM method (as mentioned above on the

section 'Pre-processing of the raw data (using SSA) only in the SVM method'). However, this may not be the case always, as the performance of these techniques depends on the characteristics of the time series (or the system) under investigation. This can be supported from the forecast results reported by Sivapragasam *et al.* (2001) for the raw rainfall and runoff data, as the SVM method yields significantly worse results (for rainfall) than the NLP method. Even though Sivapragasam *et al.* (2001) may argue that the presence of zeros in the rainfall data significantly affects the forecasts using the SVM method (and hence the use of SSA for pre-processing), the same argument applies to NLP method as well. Whatever the reasons may be, a correlation coefficient in the order of 0.18 for the training set and 0.10 for the verification set does not seem to suggest the appropriateness of the method (SVM) employed for the data studied. In fact, the results show that NLP method (with correlation coefficients of 0.57 and 0.51 for the training and verification sets, respectively) is more appropriate than the SVM method for the rainfall data studied. It is the author's opinion, based on personal experience with Singapore rainfall data (and also others), that the use of ANN could certainly provide better results than those of the SVM method for the raw rainfall data studied. Also, on the other hand, the forecast results achieved for the raw runoff data using the SVM method are only very marginally better than those using the NLP method (and possibly using the ANN as well). For these reasons, the contention by Sivapragasam *et al.* (2001) that the SVM method is better than the NLP method (and the ANN) is questionable. On the other hand, the contention that the coupling of SSA (for pre-processing) with the SVM method may provide significantly better results in forecasting is very well received, but this could be true for any other forecasting method (e.g. NLP and ANN) as well.

Sivapragasam *et al.* (2001) refer to a significant amount of literature relevant to the applications of SVM and NLP methods to hydrological time series. It is surprising, however, to note that they fail to refer to some of the studies that attempted forecasting of one of the data sets studied in their work, i.e. the Singapore rainfall data (e.g. Sivakumar *et al.* 1999*a*, *b*; Sivakumar 2000). The significance of these studies lies in the fact that such studies attempted forecasts of not only the raw (rainfall) data but also the pre-processed data (through a nonlinear noise-reduction procedure). The inclusion of such references would certainly result in a more accurate and complete presentation of the work by Sivapragasam *et al.* (2001).

CLOSING REMARKS

This comment was aimed at pointing out some of the possible inconsistencies in the study by Sivapragasam *et al.* (2001), which proposed the coupling of SSA and SVM for forecasting purposes and reported that the SSA-SVM method was significantly better than the NLP method. The inconsistencies were identified based essentially on three aspects: (1) the types of data and forecast results used for comparing the two methods; (2) the pre-processing of the raw data only in the SVM method; and (3) the forecasting approach used in the NLP method. A discussion of the analysis carried out and the results reported by Sivapragasam *et al.* (2001) with respect to these inconsistencies, and also others, raises serious concerns on one of their major conclusions regarding the better forecasting method, i.e. the SSA-SVM method is significantly better than the NLP method. It was also argued, based on the same results reported by Sivapragasam *et al.* (2001), that the opposite could be true, i.e. the NLP method is better than the SVM method. In view of these, a consistent procedure to forecasting, i.e. pre-processing of the data followed by forecasting, in both the methods is suggested, to verify the results reported and conclusions drawn by Sivapragasam *et al.* (2001).

REFERENCES

ASCE Task Committee. 2000 Artificial neural networks in hydrology, II: hydrologic applications. *J. Hydrol. Engng.* **5**(2), 124–137.

Babovic, V., Keijzer, M. & Bundzel, M. 2000 From global to local modeling: a case study in error correction of deterministic models. *Hydroinformatics 2000*, Iowa Institute of Hydraulic Research, Iowa, USA (CD-ROM).

Farmer, J. D. & Sidorowich, J. J. 1987 Predicting chaotic time series. *Phys. Rev. Lett.* **59**(8), 845–848.

Liong, S. Y. & Sivapragasam, C. 2002 Flood stage forecasting with SVM. *J. Am. Water Res. Assoc.* **38**(1), 173–186.

Mukherjee, S., Osuna, E. & Girosi, F. 1997 Nonlinear prediction of chaotic time series using support vector machines. *Neural Networks for Signal Processing—Proceedings of the IEEE Workshop*. IEEE, Piscataway, NJ, pp. 511–520.

Sivakumar, B. 2000 Chaos theory in hydrology: important issues and interpretations. *J. Hydrol.* **227**(1–4), 1–20.

Sivakumar, B., Liong, S. Y., Liaw, C. Y. & Phoon, K. K. 1999*a* Singapore rainfall behavior: Chaotic? *J. Hydrol. Engng.* **4**(1), 38–48.

Sivakumar, B., Phoon, K. K., Liong, S. Y. & Liaw, C. Y. 1999*b* A systematic approach to noise reduction in chaotic hydrological time series. *J. Hydrol.* **219**(3/4), 103–135.

Sivakumar, B., Berndtsson, R., Olsson, J., Jinno, K. & Kawamura, A. 2000 Dynamics of monthly rainfall-runoff process at the Göta basin: A search for chaos. *Hydrology and Earth System Sciences* **4**(3), 407–417.

Sivakumar, B., Berndtsson, R. & Persson, M. 2001 Monthly runoff prediction using phase space reconstruction. *Hydrol. Sci. J.* **46**(3), 377–387.

Sivapragasam, C., Liong, S. Y. & Pasha, M. F. K. 2001 Rainfall and runoff forecasting with SSA-SVM approach. *J. Hydroinformatics* **3**(3), 141–152.