A coupled SSA-SVM technique for stochastic short-term rainfall forecasting

N. Simões[1,2,*], L. Wang[1], S. Ochoa[1], J. P. Leitão[3], R. Pina[4], C. Onof[1], A. Sá Marques[2], Č. Maksimović[1], R. Carvalho[2], L. David[3]

[1]*Department of Civil and Environmental Engineering, Imperial College London, Skempton Building, South Kensington Campus, London, SW7 2AZ, United Kingdom (nsimoes@imperial.ac.uk, nunocs@dec.uc.pt)*
[2]*Departamento de Engenharia Civil, Universidade de Coimbra, Rua Luís Reis Santos, Pólo II da Universidade, 3030-788 Coimbra, Portugal*
[3] *LNEC - Laboratório Nacional de Engenharia Civil, Portugal, Av. do Brasil 101, 1700-066 Lisboa, Portugal.*
[4] *AC, Águas de Coimbra, E.E.M., Rua da Alegria n8 111, 3000-018 Coimbra, Portugal.*
**Corresponding author, e-mail nsimoes@imperial.ac.uk, nunocs@dec.uc.pt*

ABSTRACT
Short-term surface flood modelling requires reliable estimation of the distribution of floods over urban catchments with sufficient lead time in order to provide timely warnings. In this paper new improvements to the traditional Support Vector Machine (SVM) prediction technique for rainfall prediction are presented. The results obtained using the new improvements, such as enhancement of SVM prediction using Singular Spectrum Analysis (SSA) for pre-processing the data and combined SSA and SVM with a statistical analysis that give stochastic results to AI-based prediction, are compared with the results obtained using the SVM technique only. When applying the SVM technique to the rainfall data used in this study, the results showed an underestimation of the rainfall peaks. When using SSA for pre-processing the rainfall data the results are significantly better. The new stochastic approach proved to be useful for estimating the level of confidence of the forecast.

KEYWORDS
Pluvial flooding, support vector machine, singular spectrum analysis, rainfall forecasting

INTRODUCTION
Short-term surface flood modelling requires reliable estimation of the distribution of floods over urban catchments with sufficient lead time in order to provide timely warnings. The predictability of these events is limited mainly due to the restricted capabilities of the existing short-term rainfall forecast models, which generally rely on extrapolating the rainfall measurements obtained from networks of raingauges and meteorological radars. The lead time is usually short when compared to the reasonable small response time to extreme floods.
The use of radar data is becoming common; however most of the catchment system managers are still unable to access and process this type of data. Thus, only raingauges are used and consequently the extrapolation of a rainfall time series is crucial. Several authors have obtained interesting results in rainfall forecast using the technique of SVM (Support Vector Machine). Dibike *et al.* (2001) demonstrated the capability of SVM in hydrological prediction for modelling rainfall-runoff processes and found that the SVM provided better prediction of runoff on testing data when compared to the ANN (Artificial Neural Network) model. Gupta

et al. (2009) applied the SVM to forecast rainfall with a lead time from 15-min to 30-min by integrating and analysing the raingauge data of three consecutive years in Mumbay.
Short duration and extreme intensity rainfall events that cause urban floods (usually 30 to 60 min events) are extremely difficult to predict due to the high nonlinearity of sub-hourly data. This paper presents new developments to perform reliable forecast of rainfall time series. In this paper new improvements to the traditional SVM prediction technique are presented and their results are evaluated and compared with traditional SVM only.

METHODOLOGY

In this paper the following three techniques are used to forecast rainfall time series. Their results are then evaluated and compared.

1) SVM only;
2) Enhancement of SVM with SSA: The SSA is used for pre-processing the rainfall data. It separates the rainfall data into two series (smoothed series and residuals). In this technique the SVM is applied to each series separately;
3) Combined SSA and SVM using a stochastic approach: After pre-processing the rainfall data with SSA techniques, SVM is applied to the smoothed data series and a statistical analysis is applied to the residuals.

Support Vector Machine

Support Vector Machines were developed with an objective to solve pattern recognition and classification problems. This technique has been further extended to solve nonlinear regression estimation problems and have been successfully applied to solve forecasting problems in many fields (Hong, 2008). The SVM technique leads to a unique and global solution, which employs a structural risk minimization (SRM) principal as opposed to an empirical risk minimization (ERM) principal, employed by conventional ANN (Debike et al. 2001). The SRM places an upper bound on the expected risk, as opposed to the ERM, which minimizes the error on the training data only. It is this difference that provides SVM with a great ability to generalize when compared to ANN (Gupta et al, 2009).

The process to implement time series prediction using SVM includes two steps: training and forecasting. In this study, part of the historical rainfall data was used to construct (train) the prediction model. Let be the length N rainfall sequence used to train prediction model. Based on the analyses in Alonso et al. (2005) and some preliminary tests, this work employs a length 4 training window, composed of 3 predictors that will recognize the forth element (Figure 1). This indicates that the forecasted precipitation will be derived from the previous three observations. This window then slides one time step forward (the dotted-line window in Figure 1) and repeats the procedure to establish the prediction model.

Based on the works developed by Vapnik (1995) and Joachims (1999, 2002), SVM^{light} is a free software developed by Thorsten Joachims from Department of Computer Science at Cornell University. The regression module of the SVM^{light} is used in this work to carry out the time series prediction.

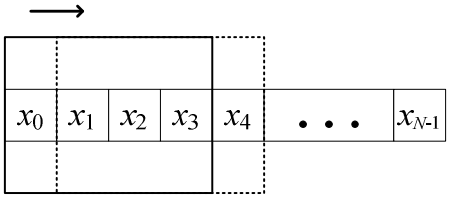

Figure 1: Schematic of the sliding-window for time series prediction using SVM

Singular Spectrum Analysis (SSA)

Singular Spectrum Analysis can be used for pre-processing data (Sivapragasam et al, 2001). SSA is a non-parametric technique used in the analysis of time series. Its usefulness has been proven in the analysis of climatic, meteorological and geophysical time series (Alonso et al., 2005). Basic SSA technique performs four steps (Golyandina et al., 2001). At the first step, called the embedding step, the one-dimensional series is represented as a multidimensional series whose dimension is called the window length. The multidimensional time series (which is a sequence of vectors) forms the trajectory matrix. The second step is the singular value decomposition (SVD) of the trajectory matrix into a sum of rank-one bi-orthogonal matrices. The third step is the grouping step and corresponds to splitting the matrices, computed at the SVD step, into several groups and summing the matrices within each group. The result of the step is a representation of the trajectory matrix as a sum of several resultant matrices. The last step transfers each resultant matrix into a time series, which is an additive component of the initial series. The corresponding operation is called diagonal averaging. It is a linear operation and maps the trajectory matrix of the initial series into the initial series itself. In this way we obtain a decomposition of the initial series into several additive components.

The general purpose of the SSA analysis is the decomposition with additive components that are 'independent' and 'identifiable' time series (Golyandina et al., 2001). Sometimes, one can also be interested in particular tasks, such as extraction of signal from noise, extraction of oscillatory components and smoothing. In this work SSA is used for smoothing, which means representing the series as a sum of two series where the first one is a 'smooth approximation' of it (original series = smoothed series + residuals). In figure 2 it is shown a rainfall data series, the smoothed data series and its residuals.

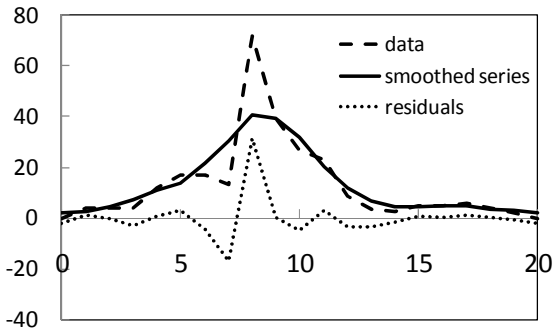

Figure 2 – Rainfall data series, smoothed series and residuals.

Enhancement of SVM with SSA

In this improved technique, the SSA is used for pre-processing the rainfall data. Support vector machine is applied to each series separately. Consequently SSA is applied to all data series and two data sets are generated: the smoothed series and the residual. Each one of the data series will generate two SVM models and the prediction is done separately for each series. At the end, the smoothed and residuals predictions are added and the rainfall predictions are obtained.

Combined SSA and SVM using a stochastic approach

The SVM technique is not very efficient to predict very irregular patterns as the residuals. Consequently a new approach was developed for the nowcasting of a rainfall time series. Firstly the decomposition of the time series is carried out using SSA, producing a smoothed series and its associated residuals. Then the forecast of the smoothed series is done with SVM.

With this result it is then possible to identify the magnitude and peak time of the rainfall event.

Using the SVM technique involves training the model with historical data. In this new approach historical data is also needed. Firstly the instant of the peak in all residuals should coincide. Secondly is the definition of the characteristic curves. These curves are defined as the average of the residual in each instant (c1), average plus the standard deviation of the residual (c2), and average plus two times the standard deviation of the residual (c3). The main reason to add standard deviation is because the SVM forecast under-estimate the peaks. The next step is to sum the smoothed series forecast with the characteristic curves. Finally, a stochastic approach is obtained and an "envelope" for the forecast results is created. The observed data should be between the smoothed series plus the c1 and smoothed series plus c3. This procedure is explained in Figure 3. Figure 4 shows and example of the historical residual data and its characteristic curves.

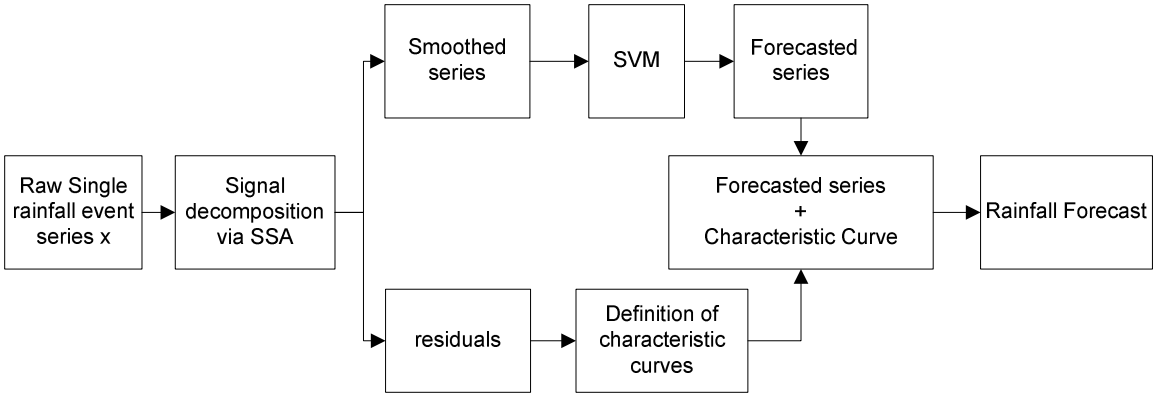

Figure 3 – Stochastic SSA-SVM techniques

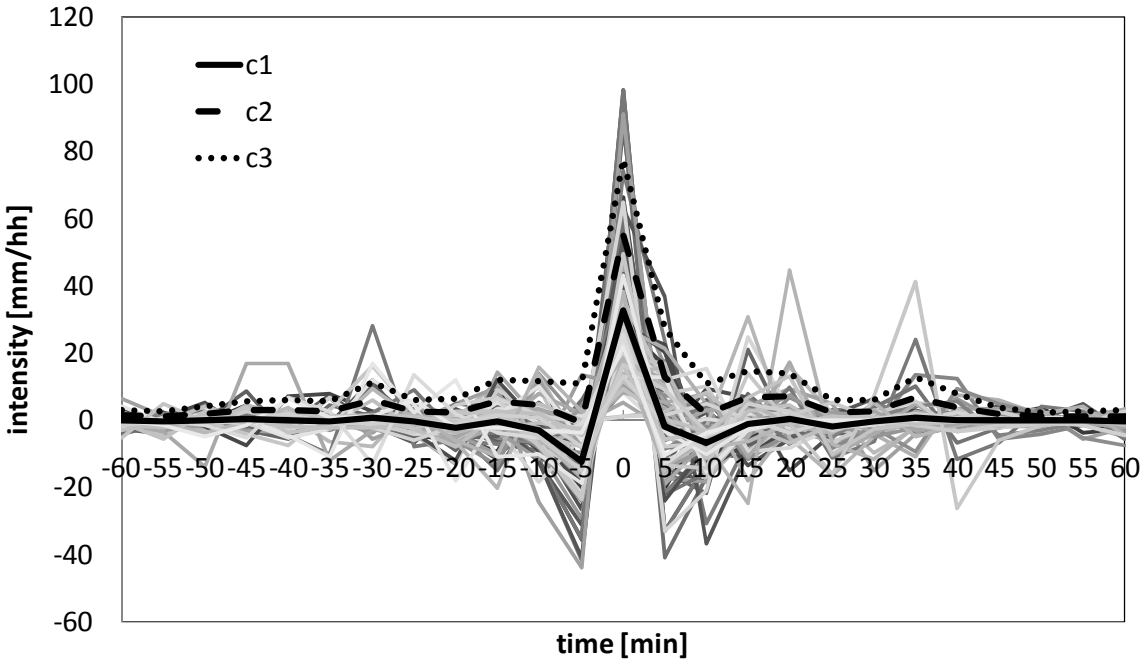

Figure 4 –Residuals series and characteristic curves.

RESULTS

Rainfall data set

Coimbra meteorological station is located at the Geophysical Institute of University of Coimbra and was installed in 1864. The siphon udograph daily charts from 1935 to 2005 were recently digitalized. The dataset was digitalized by INAG, the Portuguese Water Institute, using the SIFDIA program that allows one minute discretization (Carvalho et al, 2008). In this work, approximately 70 years of continuous data was used. The inter-event interval used was 6 hours and the minimum event depth was 0.2mm. The events chosen have a return period higher than 2 years for average intensity during 30, 45 or 60 minutes. In conclusion, 84 events with 5 minutes (with the real 5min peak) time step were selected.

Rainfall events

In this paper three rainfall events were chosen to test the methodologies. The 09-04-1946 event has a high return period. The 23-05-2004 has 2 year return period and the 25-10-2006 has an intermediate return period with several rainfall peaks.
Figure 5 presents the rainfall events and the return period for different durations.

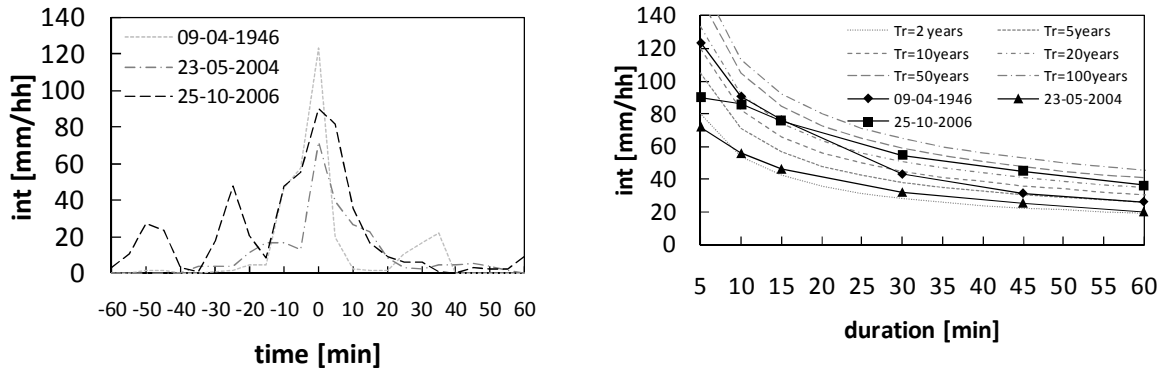

Figure 5: rainfall events (left) and return period of the selected rainfall events (right)

From Figure 6 to 11 the hyetographs for the different techniques are presented for each event. In each graph a hyetograph starting with a forecast starting time (fst) of 5 minutes prior to the peak to 25 minutes after the peak is presented. The developed schemes where applied

23/05/2004 event

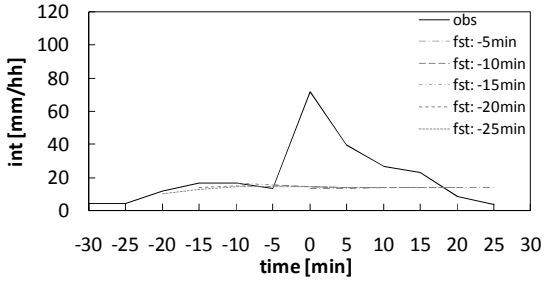

Figure 6: Forecast using SVM

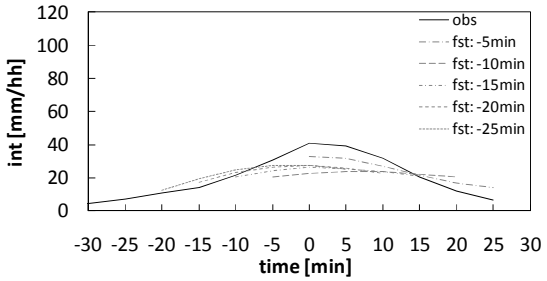

Figure 7: Forecast of smoothed series using SVM

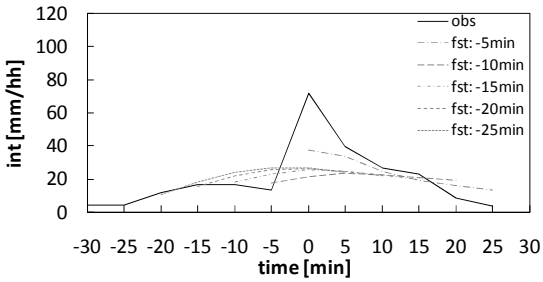

Figure 8: Forecast using SSA+SVM (smoothed +residuals)

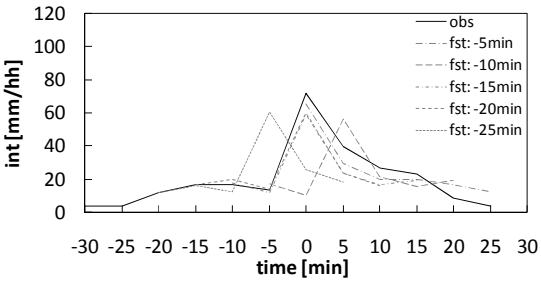

Figure 9: Forecast using SSA+SVM (smoothed +c1)

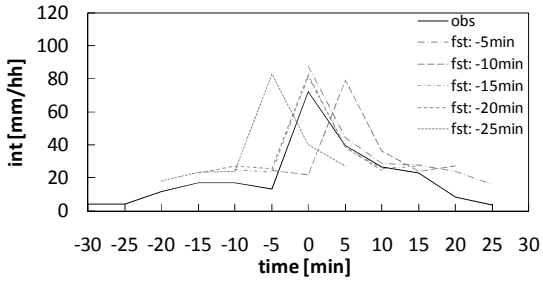

Figure 10: Forecast using SSA+SVM (smoothed +c2)

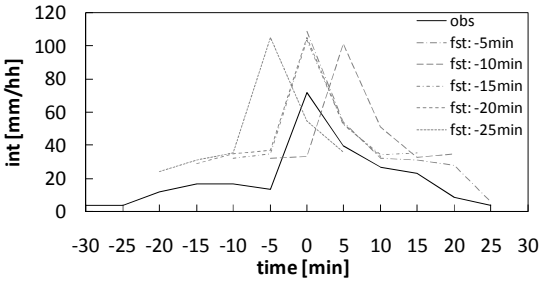

Figure 11: Forecast using SSA+SVM (smoothed +c3)

Figure 6 shows that the SVM technique highly underestimates the rainfall event. In Figure 7 the hyetograph of the smoothed rainfall series and several forecasts is presented; it can be seen that the forecast of the smooth series is good. Figure 8 presents the rainfall forecast doing the smoothed series and residuals separately. The results are not satisfactory but better than the traditional SVM approach. Figures 9 to 11 show the rainfall forecast using the stochastic approach described previously. The results show a very good agreement between observed data and the forecasted data. It can be seen that some peaks are predicted 5 minutes before or after the time they really happen.

Figures 12 to 16 show the quality of the peak prediction using the different techniques, for different forecast start times. It can be seen that the results obtained using the SVM only technique underestimates the peak and that the SSA-SVM techniques results are always better than using the SVM technique isolated. The stochastic approach predicts the peak reasonably accurately. It can be seen in Figures 14 to 16 that, when the rainfall event has 2 peaks and the forecast starts very early, the method does not work so well, but still is the one that performs better among the methods evaluated in this study.

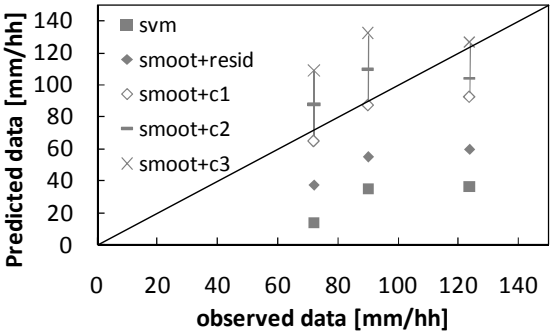

Figure 12: Peak values for fst:-5min

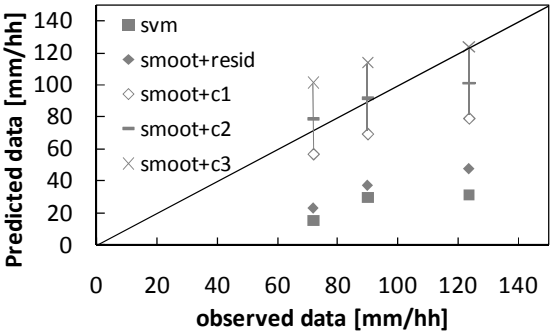

Figure 13: Peak values for fst:-10min

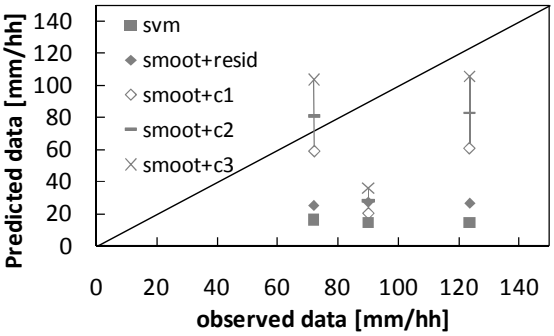

Figure 14: Peak values for fst:-15min

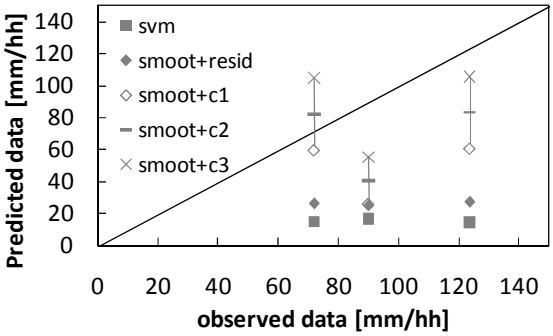

Figure 15: Peak values for fst:-20min

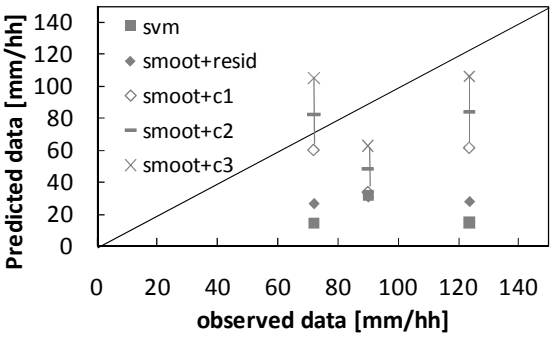

Figure 16: Peak values for fst:-25min

DISCUSSION AND CONCLUSIONS

New improvements to the traditional SVM prediction technique were presented in this paper and their results were evaluated and compared to the results obtained using only the SVM technique. When applying the SVM technique to the rainfall data, the results underestimate the rainfall actual peaks. In contrast, when using SSA for pre-processing the rainfall data and, in turn, applying the SVM to the smooth series and its residual separately the results are significantly better, especially in the case of the smoothed series where the predicted rainfall peaks show a good agreement with the actual rainfall peaks. Due to the irregular behaviour of the residual, the SVM technique is not very efficient in this case. The new stochastic approach proved to be useful to overlap this drawback and for estimating the level of confidence of the forecast.

Future works include tests with data using different time steps. Ten or twenty minute data is smoother and consequently easier to predict using Artificial Inteligence algorithms. However in urban floods the time discretization is very important and works that combine SVM with Cascade Methods can be studied. The combination of these techniques with interpolation methods allow having a spatial distribution of rainfall, which also need to be tested.

ACKNOWLEDGEMENT

Nuno Simões would like to acknowledge the financial support from the Fundação para a Ciência e Tecnologia - Ministério para a Ciência, Tecnologia e Ensino Superior, Portugal [SFRH/BD/37797/2007]. Li-pen Wang would like to acknowledge the full financial support of the Ministry of Education of Taiwan for his postgraduate research programme.

REFERENCES

Alonso, F.J., Del Castillo, J.M, Pintado, P., (2005), Application of singular spectrum analysis to the smoothing of raw kinematic signals. J. Biomech. 38, 1085-1092

Carvalho, R.F.; David, L.M.; Martins, C.; Temido, G.; de Lima, JLMP, (2008). Statistical characterization of extreme rainfall climate along the future high-speed rail track in Portugal. European Geosciences Union (EGU) General Assembly, Vienna, Austria.

Dibike, Y. B., Velickov,S., Solomatine, D. and Abbott M. B. (2001): Model induction with support vector machines: introduction and applications. Journal of Computing in Civil Eng, 15(3), 208–216.

Golyandina, N., Nekrutkin, V. and Zhigljavsky, A. (2001). Analysis of Time Series Structure: SSA and related techniques. Chapman & Hall/CRC.

Gupta, K. and Nikam V., Rainfall forecast for extreme monsoon rainfall conditions for urban area, Proceedings of ‘‘8th International Workshop On Precipitation In Urban Areas.’’ St. Moritz, Switzerland, 2009.

Hong W.C. (2008). Rainfall forecasting by technological machine learning models. Applied Mathematics and Computation, 200, 41-57

Joachims T. (1999). Making large-Scale SVM Learning Practical. Advances in Kernel Methods. In: B. Schölkopf, C. Burges and A. Smola (eds.), Support Vector Learning, MIT-Press, Chapter 11, pp. 41-56.

Joachims T. (2002). Learning to Classify Text Using Support Vector Machines. Kluwer Academic Publishers, NL.

Sivapragasam, C., Liong, S. and Pasha, M., Rainfall and runoff forecasting with SSA–SVM approach, Journal of Hydroinformatics, 03.3, 2001

Vapnik V. N. (1995). The nature of statistical learning theory. Springer, New York.