An enhanced blend of SVM and Cascade methods for short-term rainfall forecasting

L. Wang[1,*], N. Simões[1,2], 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]*Imperial College London, Department of Civil and Environmental Engineering, South Kensington Campus, London, SW7 2AZ, United Kingdom*
[2]*Departamento de Engenharia Civil, Universidade de Coimbra, Rua Luís Reis Santos, Pólo II da Universidade, 3030-788 Coimbra, Portugal*
[3] *Laboratório Nacional de Engenharia Civil (LNEC), 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 li-pen.wang08@imperial.ac.uk*

ABSTRACT
A more reliable flood forecasting could benefit from higher-resolution rainfall forecasts as inputs. However, the prediction lead time of the operational rainfall forecasting models will substantially diminish while sub-hourly (e.g., 5-min) rainfall forecasting is required. A method that integrates the SVM (Support Vector Machine) and Cascade-based downscaling techniques is therefore developed in this work to carry out high-resolution (5-min) precipitation forecasting with longer lead time (45-60 minutes). The 5-min raingauge observations from Coimbra (Portugal) are employed to assess the proposed methodology. A comparison with the conventional SVM is also conducted to study the possible benefit of using the proposed methodology to carry out short-term rainfall forecasting.

KEYWORDS
Support vector machine, cascade, log-Poisson, rainfall forecasting, downscaling

INTRODUCTION

Integrated rainfall modelling
Accurate and timely rainfall forecast is crucial for the corresponding flood simulation and prevention, especially in the case of urban surface/pluvial floods. Prediction of the rainfall volumes caused by a shower an hour or more before it arrives at a particular location is an important and useful topic (Tsonis and Austin, 1981). A more reliable estimation of flood distribution could benefit from higher-resolution rainfall forecasts as inputs. However, the predictability of the operational rainfall forecasting models will substantially diminish while sub-hourly (e.g., 5 min) rainfall forecasting is required. To circumvent this difficulty, the integration of rainfall models over multiple scale ranges has been widely used to carry out high-resolution rainfall forecasting with longer lead time (Bowler *et al.*, 2006; Sokol, 2006; Golding, 1998).

Based upon this integration scheme, this work combines the Support Vector Machine (SVM) and the cascade-based downscaling techniques, in order to carry out 5-minute rainfall forecasting with up to 45-60 min lead time.

Support Vector Machines (SVMs)

The SVM is an AI (Artificial Intelligent)-based learning method that was developed with an objective to solve pattern recognition and classification problems which have 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 leads to a unique and global solution because of its formulation, which employs a structural risk minimization (SRM) principal as opposed to an empirical risk minimization (ERM) principal, employed by conventional neural networks (Debike *et al.*, 2001). The SRM places an upper bound on the expected risk, as opposed to an ERM, which minimizes the error on the training data only. It is this difference that equips SVM with a great ability to generalize compared to ANN (Gupta *et al.*, 2009). The SVM has been applied to quantitative short-term rainfall and hydrological forecasting. For example, 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 as compared to the ANN 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. These results demonstrated the SVM's potential to synthesise the complex patterns of nonlinear geophysical processes; however, they also indicate that the SVM is incapable of well producing the patterns of time series at finer resolutions, such as 5 min.

Cascade-based downscaling

The cascade-downscaling methods are developed based upon the investigation of the scale-invariant behaviour of complex nonlinear processes. The cascade is a single process to generate fine-scale data by subdividing a unit set into smaller and smaller subsets according to a fixed set of contracting (fragmentation) ratios (**S** in Figure 2) and at the same time subdividing the associated unit measure by another set of contracting ratios (**W** in Figure 2). Many efforts have been made to characterise these ratios and to apply them to spatially- or temporally-distributed rainfall downscaling in the literature (Over and Gupta, 1996; Deidda *et al.*, 1999; Onof *et al.*, 2005; Pathirana *et al.*, 2003; Onof and Arnbjerg-Nielsen, 2009; Wang *et al.*, 2010, 2011), among which random cascade methods are the mainstream and have been widely developed. The general idea is to construct rainfall generators based upon analysing statistical features (or probability distributions) of **W**. The associated parameters of generators can be empirically estimated from historical rainfall observations. For example, Onof *et al.* (2005), for example, referring to the derivation in Deidda *et al.* (1999), used a log-Poisson cascade to disaggregate hourly rainfall sequences to 5-min. The results showed that specific statistics were satisfactorily reproduced, which is crucial for the uses of the corresponding hydrological modelling (e.g., ground runoff and sewer network simulation).

RAINFALL DATA

Coimbra meteorological station of Geophysical Institute of University of Coimbra was installed in 1864 at 141 m altitude in Coimbra. 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 5 minutes peak) time step were selected.

METHODOLOGY

Based upon the concept of integrated rainfall modelling, this work blends the SVM and cascade-based downscaling techniques to carry out high-resolution rainfall forecasting with longer lead time. Instead of being used to directly generate 5-min forecasts, the SVM technique is used herein to produce 15-30 min rainfall forecasts. It is because the pattern of 15-30 min rainfall sequences is somewhat smoother and is expected to be better predicted; moreover, a longer lead time forecasting (approximately 3 time steps ahead) can be conducted due to larger time intervals. The coarser forecasts are then disaggregated to finer ones through the cascade-based downscaling process. Finally, the 5-min rainfall forecasts with approximately 45-90 min lead time are expected to be produced in this study. The details of the implementation of each technique are explained as follows.

Time Series Prediction: SVM regression

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 $\boldsymbol{X} = \{x_0, x_1, \dots, x_{N-1}\}$ be the length N rainfall sequence used to train prediction model. Based upon the analyses in Alonso *et al.* (2005) and some preliminary tests, this work employs a length 4 training window, composed of 3 predictors and 1 predictand (see the solid-line window in Figure 2). This indicates that, in this research, the precipitation of interest (e.g., x_3) is assumed to be able to be derived from the previous three observations (i.e., x_0, x_1 and x_2). This window then slides one time step forward (the dotted-line window in Figure 2) and repeats to establish the prediction model. The rest of historical rainfall data was then assumed to be unknowns in this work to validate forecasting results and to assess the predictability of the proposed method.

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

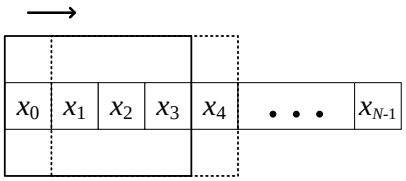


Figure 1: Schematic of the sliding-window for time series prediction using SVM

Log-Poisson Cascade Methods

Log-Poisson cascade, which has been widely used to disaggregate hourly precipitation to sub-hourly (Deidda et al., 1999; Onof et al., 2005; Onof and Arnbjerg-Nielsen, 2009), is employed in this work to downscale the coarser forecasts by SVM to 5-min. The associated generator is formed as,

$$W = A\beta^N, \quad (1)$$

where N is a log-Poisson distributed random variable, and A and β are two parameters that can be identified by fitting the observed structure function ($K(q)$), which is empirically plotted based upon scale invariance of historical raingauge observations. An scaling investigation of the observed rainfall data was carried out over the time scales ranging from 5 to 120 min (Figure 3 (Left)), which shows that, within the investigating temporal range, scale invariance is well observed. Based upon this, the observed $K(q)$ is further derived and plotted as the grey dashed line in Figure 3 (Right). The observed $K(q)$ curve then can be substituted into the relation of the theoretical structure function ($K(q)$) and the Log-Poisson generator (W), expressed as:

$$K(q) = -\log_2 E[W^q] = -C\frac{q(1-\beta)+\beta^q-1}{\ln 2}, \quad (2)$$

the associated parameters (β and C) for the generator then are derived, where q is a real value and the parameter A in Eq. (1) can be further obtained using $A = e^{C(1-\beta)}$.

In this study, the *Cascade* computer programme for disaggregation is used (Onof, 2009), in which the key parameters (β and C) is optimally solved by minimising the difference of the observed and the theoretical structure functions for a certain range of q. The optimally-derived theoretical (log-Poisson) $K(q)$ curve is plotted as the dark solid line in Figure 3 (Right); the associated parameters β and C are respectively 0.263 and 0.519 (A thus equals 1.466). A very good fit can be seen using log-Poisson distribution, particularly for $q \geq 1.0$. These parameters are then substituted into Eq. (1) to construct the rainfall generator to produce 5-min precipitation for this work.

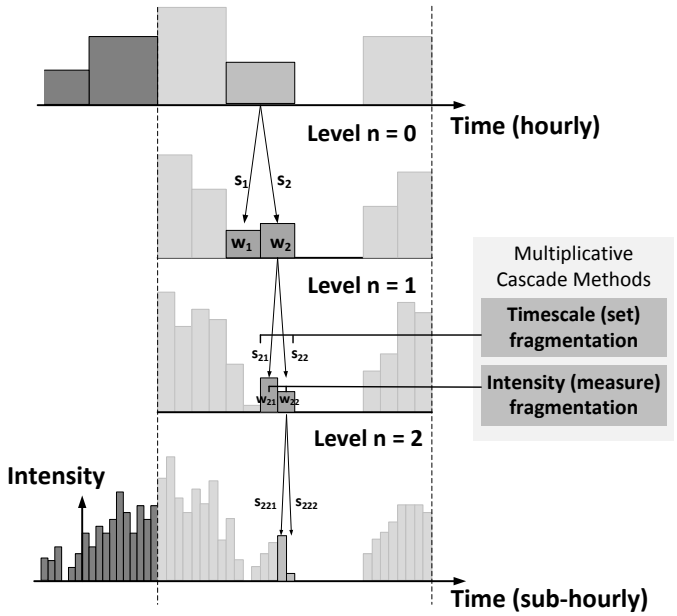


Figure 2: Conceptual schematic of a cascade process, where a coarser volume in a specific scale is repeatedly subdivided into numbers of sub-volumes according to certain timescale (set) and intensity (measure) fragmentation ratios (**S** and **W**).

RESULTS

Two events (denoted as event A and B in the following context) were randomly selected from the historical rainfall data sets and assumed to be unknown to evaluate the proposed methodology. Each component of the proposed methodology, i.e., SVM and Log-Poisson cascade methods, is separately tested first to evaluate their capacities. An integrated testing is then applied to assess the possible improvements that can be obtained through this blend.

SVM Time Series Prediction

The results of time series prediction merely using SVMs over 5 – 30 min time scales are shown in Figure 4 and 5. Substantial underestimates of the peak values can be generally observed in the 5-min cases for both events (the grey lines in Figure 4) and in the 15- and 30-min cases for the event A (Figure 5 (Left)); nonetheless it is somewhat improved for the event B when the 15- and 30-min rainfall forecasting is carried out (Figure 5 (Right)). These indicate that the proper time scale that SVMs can be employed to predict rainfall time series shall be at least larger than 15 – 30 minutes.

Log-Poisson Cascade Downscaling

The log-Poisson distributed rainfall generator used herein is constructed based upon the parameters obtained from the scaling analyses process. An evaluation of using this generator to synthesise 5-min precipitation respectively from 15- and 30-min rainfall observations is carried out. The downscaled profiles (from 15-min) are plotted as the dark dotted lines in Figure 4. Good syntheses can be observed in both events, particularly the ability to reproduce extreme values, which is very crucial in short-term rainfall forecasting. These results depict that the log-Poisson distributed generator is an appropriate tool with promising abilities to reproduce the complex patterns of 5-min rainfall in these two events.

Integrated Rainfall Forecasting

Based upon the proposed methodology, an integrated forecasting is carried out. The 15- and 30-min rainfall forecasts that were firstly produced using the SVM time prediction technique are shown in Figure 4. The log-Poisson distributed rainfall generator is then employed to downscale these forecasts into 5-min ones. For the event A (Figure 6 (Left)), except the case of dsft: -5min(-15min), no obvious improvements can be seen compared with the SVM-based prediction. It is due to the poor forecasting at 15- and 30-min time scales; so in spite of the higher variability that can be seen among downscaled forecasts, substantial underestimation is still observed. For the event B (Figure 6 (Right)), thank to the better prediction of 15- and 30-min rainfall time series, obvious improvements can be seen, particularly the reproduction of peak values.

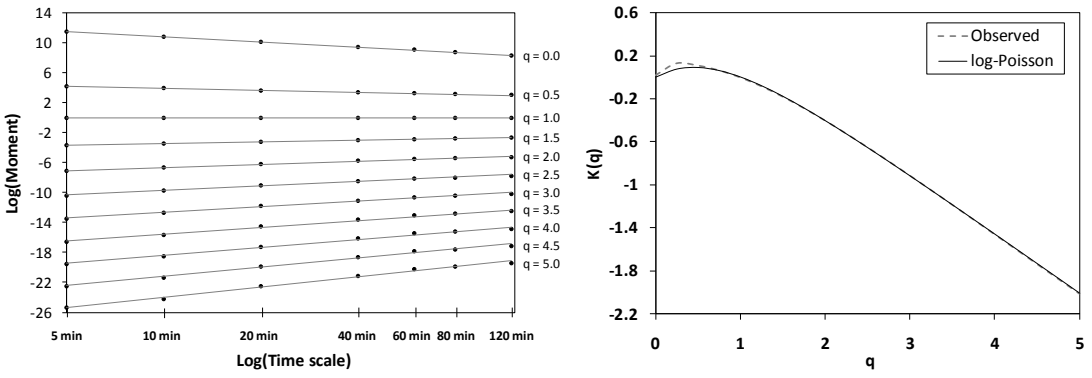


Figure 3: (Left) Log-log plot of moment as a function of timescale between 120 min and 5 min for $0.0 \leq q \leq 5.0$; (Right) The observed and theoretical (log-Poisson) $K(q)$ curves, respectively drawn by the grey dashed and the dark solid lines.

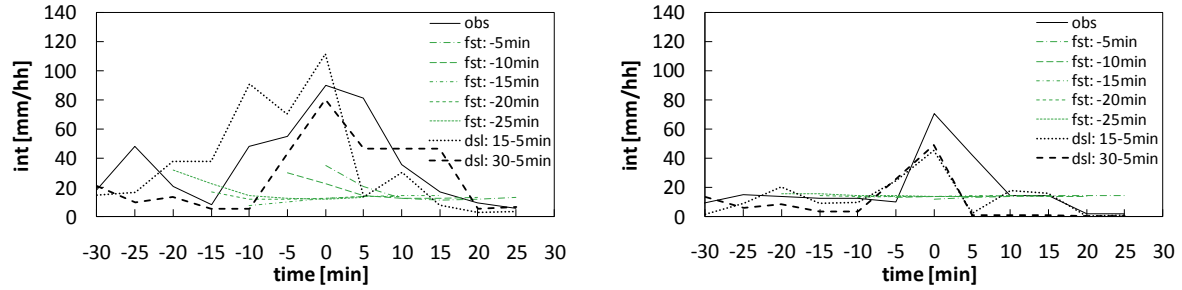


Figure 4: Profiles of event A (Left) and B (Right) in the 5-min time interval, where the peak is defined to occur at time = 0. The dark solid line represents the observed data; the dark dotted (from 15 min) and dashed (from 30 min) lines represent the downscaled rainfall profiles; the other lines are forecasting results with different starting time points.

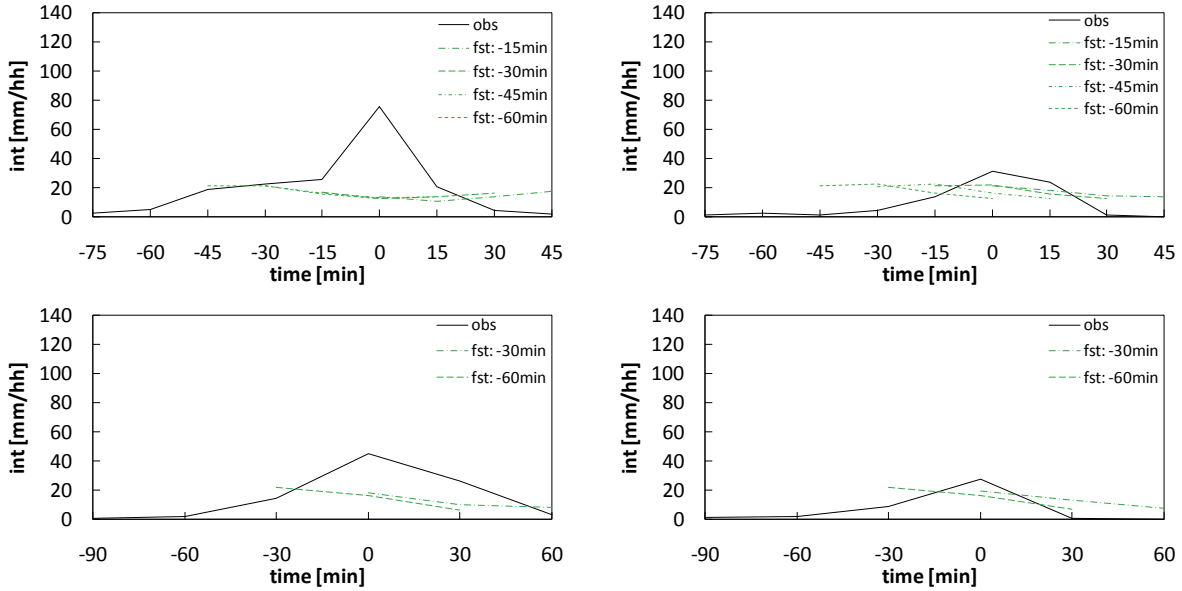
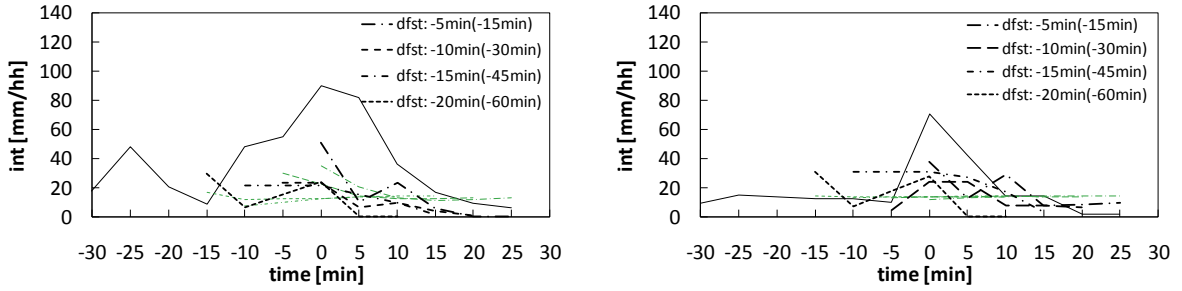


Figure 5: Profiles of event A (Left) and B (Right) and the associated forecasts respectively in the 15- (Upper) and 30-min (Lower) time intervals.



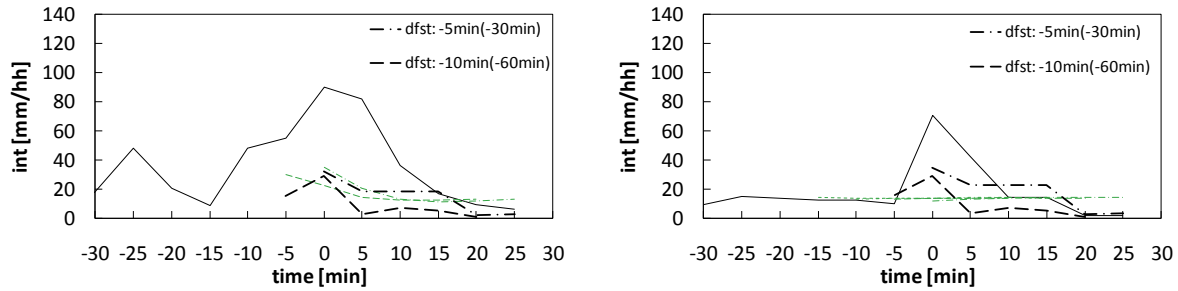


Figure 6: Forecasts of event A (Left) and B (Right) in the 5-min interval respectively downscaled from 15- (Upper) and 30-min (Lower) forecasts, where the dark lines represent downscaled forecasts and the grey lines are SVM-based forecasts. The dfst: -5min (-30min) means the downscaled forecasts (originally starting at the -30 min time point before being downscaled0 that start being compared at -5 min time point.

DISCUSSION AND CONCLUSIONS

In this study, a methodology that blends the SVMs and log-Poisson cascade techniques is proposed and implemented. Based upon the concept of integrated rainfall modelling, the proposed methodology aims to carry out high-resolution (5-min) rainfall time series forecasting with the longer lead time (45 – 60 min). Rainfall observations from Coimbra (Portugal) were used in this work to assess the capacity of this blend. Preliminary results suggest that the appropriate time scale of SVM-only time series prediction is larger than 15-30 min. This implies the necessity of SVM-based forecasting methods to be combined with downscaling techniques and the potential of the proposed methodology to carry out finer-resolution rainfall prediction. However, the results also show that, although the log-Poisson cascade methods enable well reproducing the pattern of 5-min rainfall time series, the predictability of this blend mainly relies on the accuracy of 15- and 30-min rainfall prediction.

The perspective work of this blend will therefore be focused on improving the state-of the-art SVM-based time series prediction techniques; for some preliminary achievements, the readers may refer to Simões et al. (2011).

ACKNOWLEDGEMENTS

The research was conducted as part of the Flood Risk Management Research Consortium (FRMRC2, SWP3). The first author acknowledges the full financial support of the Ministry of Education of Taiwan for his postgraduate research programme. The second author acknowledges the financial support from the Fundação para a Ciência e Tecnologia - Ministério para a Ciência, Tecnologia e Ensino Superior, Portugal [SFRH/BD/37797/2007].

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**(5), 1085-1092

Bowler N., Pierce C. and Seed A. (2006). STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. Q. J. Roy. Meteor. Soc., **132**(620), 2127-2155.

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.

Deidda R., Benzi R. and Siccardi F. (1999). Multifractal modeling of anomalous scaling laws in rainfall. Wat. Resour. Res., **35**(6), 1853-1867.

Dibike Y. B., Velickov S., Solomatine D. and Abbott M. B. (2001). Model induction with support vector machines: introductionand applications. J. Comput. Civil Eng., **15**(3), 208–216.

Golding B. (1998). Nimrod: a system for generating automated very short range forecasts. Meteorol. Appl.. **5**(1), 1-16.

Gupta K. and Nikam V. (2009): Rainfall forecast for extreme monsoon rainfall conditions for urban area. In: P. Molnar, P. Burlando and T. Einfalt (eds), Rainfall in the Urban context: forecasting, Risk and Climate Change. Proc. 8th Int. Workshop on Precipitation In Urban Areas, St. Moritz, Switzerland, 10-13 December 2009.

Hong W.C. (2008). Rainfall forecasting by technological machine learning models. Appl. Math. Comput., **200** (1), 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.

Onof C. (2009). Cascade computer programme for disaggregation. Imperial College London.

Onof C. and Arnbjerg-Nielsen K. (2009). Quantification of anticipated future changes in high resolution design rainfall for urban areas. Atmos. Res., **92**(3), 350-363.

Onof C., Townend J. and Kee R. (2005). Comparison of two hourly to 5-min rainfall disaggregators. Atmos. Res., **77**(1-4), 176-187.

Over T. M. and Gupta V. K. (1996). A space-time theory of mesoscale rainfall using random cascades. J. Geophys. Res., **101**(D21), 26319-26331.

Pathirana A., Herath S. and Yamada T. (2003). Estimating rainfall distributions at high temporal resolutions using a multifractal model. Hydrol. Earth Syst. Sc,, **7**(5), 668-679.

Simões N., Wang L., Ochoa S., Leitão J. P., Pina, R., Onof C., Sá Marques A., Maksimović Č., Carvalho R. and David L. (2011). A coupled SSA-SVM technique for stochastic short-term rainfall forecasting. 12nd International Conference on Urban Drainage, Porto Alegre, Brazil, 10-16 September 2011.

Sokol Z. (2006). Nowcasting of 1-h precipitation using radar and NWP data. J. Hydrol., **328**(1-2), 200-211.

Tsonis A. and Austin G. L. (1981). An evaluation of extrapolation techniques for the short-term prediction of rain amounts. Atmos. Ocean, **19** (1),

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

Wang L., Onof C. and Maksimović Č. (2010). Reconstruction of sub-daily rainfall sequences using multinomial multiplicative cascades. Hydrol. Earth Syst. Sc. Discuss., **7**(4), 5267-5297.

Wang L., Onof C. and Maksimović Č. (2011). An improved discrete cascade method for sub-daily rainfall modelling. European Geosciences Union General Assembly 2011, Vienna, Austria, 3-8 April 2011.