On Prediction of Subang, Selangor Daily Rainfall Data: An Application of Local Approximation Method

Peramalan Data Hujan Harian di Subang, Selangor: Suatu Aplikasi Kaedah Penghampiran Setempat

Nor Zila Abd Hamid[1] and Mohd Salmi Md Noorani[2]
[1]Department of Mathematics, Faculty of Science and Mathematics,
Universiti Pendidikan Sultan Idris, 35900, Tanjung Malim, Perak, Malaysia
[2]School of Mathematical Sciences, Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia
e-mail: [1]nor.zila@fsmt.upsi.edu.my and [2]msn@ukm.my

Abstract

Prediction has been done on the daily rainfall data taken from Subang Meteorology Station, Selangor, located in Malaysia. Firstly, the original daily rainfall data need to be systemized through the reconstruction of phase space approach (RPS). Then, the system was utilized to predict rainfall via local approximation method (LAM). Out of seven days predictions, the results are found to be in good agreement with the observed ones for five days. Even though the results seem well, the values of overall prediction errors are quite high. So, some analysis have been done, a few remarks are highlighted and a number of suggestions are proposed in order to produce better prediction.

Keywords rainfall, chaos, local approximation method, reconstruction of phase space

Abstrak

Peramalan ke atas data hujan harian dari Stesen Kajicuaca Subang, Selangor, Malaysia telah dilaksanakan. Peramalan bermula dengan membina sistem untuk data hujan asal melalui kaedah pembinaan semula ruang fasa. Kemudian, dengan mengaplikasikan kaedah Penghampiran Setempat, sistem yang dibina digunakan untuk meramal data hujan. Tujuh hari peramalan dilakukan dan lima hari daripadanya menunjukkan penghampiran yang baik dengan data hujan yang sebenar. Walau peramalan kelihatan baik, namun, nilai ralat peramalan secara keseluruhan agak tinggi. Maka, analisis terhadap hasil dapatan telah dijalankan, beberapa isu dibangkitkan dan terdapat cadangan diberikan agar peramalan yang lebih baik dapat dihasilkan di masa hadapan.

Kata kunci hujan, kalut, kaedah penghampiran setempat, pembinaan semula ruang fasa

Introduction

Prediction of a system is important because from prediction, people can be well prepared to face what is going to happen in the future. Prediction is important to various fields such as finance, hydrology, traffic and many more. However, this presented study only focuses on prediction of one of hydrology system, rainfall. The prediction of rainfall is important because if we know early that the flood is going to occur, a well preparation can be made to face it.

Flood is one of the worst disasters to human life. Flood occurs because of heavy rain for such a long period of time. Flood cause lot of lost. For example, the flood occurred in North Malaysia on November 2010. This flood caused the transportation problem including the land road, train, highway and airlines because the roads and airport are under water. Flood also caused the water contamination that affect human's health. The production of paddy also decreased because about 45,000 hectares of the paddy field reported destroyed.

Moreover, some peoples lost their love one and some more lost their place to stay at. Another flood disaster also have been reported in Sabah and Sarawak (January 2010), Kelantan (November 2010), Johor, Malacca, Pahang and Negeri Sembilan (January 2011) and many more (Utusan Malaysia Online, 2011). These motivate us that the prediction of Malaysia's rainfall needs to be done properly. The prediction of rainfall might save human life and reduce lost.

Before predicting the rainfall, we have to systemize the original rainfall data first. Systemizing the data is important because the behavior of the data can be observed through the system. In this study, we use the reconstruction of phase space approach (RPS) to systemize the original data. By then, we can track where the system is going to in future. In other word, we can predict the system. In order to predict our observed rainfall data, we applied the local approximation method (LAM).

The RPS has been used by numbers of researcher in hydrologic area. As to date, there are two main reasons of using RPS in any system; 1) to detect the chaotic behavior of the system and 2) to predict the system. Sivakumar *et al*. (2001), Islam and Sivakumar (2002) and Sivakumar (2002) detect the chaotic behavior of Brazil's monthly runoff, Denmark's daily runoff and Mississippi River's sediment concentration in United State, respectively. In Singapore, Sivakumar *et al*. (1999) successfully detect the presence of chaos in daily rainfall data via RPS.

While in Malaysia, Radhakrishnan and Dinesh (2006) and Betty *et al*. (2010) proved that rainfall system in Malaysia is also chaos via the same manner. In theory, if any system was proven as chaos, this means that the system is lying between noisy (unpredictable) and well studied (predictable) system (Abarbanel, 1996). This means that chaos system having noise but slightly predictable. This chaotic system is really sensitive on initial conditions. Hence, with those characteristic, only short term prediction can be done.

The second reason of using RPS is to predict the system. The real data were observed in scalar vector x_1, x_2, x_3, x_4,... x_N (Equation 1), with total data N. Then, via RPS, the data was embedded into phase space $Y_t = \left\{ x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(d-1)\tau} \right\}$ (Equation 2) and we call it reconstructed phase space $\boldsymbol{Y}$ at time $\boldsymbol{t}$ with embedding dimension $\boldsymbol{d}$ and delay time, $\boldsymbol{\tau}$. The method on how to find the value of d and τ which are false nearest neighbor (FNN) and average mutual information (AMI) shall be elaborated later, in methodology section.

The phase space $\boldsymbol{Y_t}$ is then been utilized to predict rainfall via the approximation method. As to date, there are two main methods; namely global and local approximation method. Via global method, it is an important issue to choose appropriate functional form that represents the whole phase space. On the other hand, the basic idea of local method is to use only the nearby state to the present state in phase space in order to make predictions. In general, global method provides good approximations if the observed data are well behaved and not very complicated (Velickov, 2004).

Since we cannot see the behavior of the observed rainfall data, we decided to use the LAM as the predictor. LAM employed that the behavior of current state were related to the previous state. This means that, what happen today is relating on what had happened in past. Moreover, what will happen tomorrow is relating on what happen today, which relate on what had happened in past.

LAM was introduced by Farmer and Sidorowich (1987) who applied it to the short term predictions. LAM is then been widely used. See for example: Jayawardena (1997), Sivakumar *et al*. (2002) and Mekonnen and Jayawardena (2004). In United State, Sivakumar (2002) and Khan *et al*. (2005) utilized LAM in order to predict sediment concentration in Mississippi River and stream flow in Colorado and Arkansas River, respectively. Moreover, Sivakumar *et al*. (2001), Islam and Sivakumar (2002) and Pao-Shan *et al*. (2004) utilized LAM to predict monthly runoff in Brazil, daily runoff in Denmark and hourly rainfall in Taiwan, respectively.

With a slightly modification, Wu and Chau and She and Yang (2010) utilized LAM to predict stream flow in Hong Kong and daily discharge in Russia. In early 2011, Siek and Solomantine modified LAM in order to predict storm in North Sea. All of the prediction results via LAM seem to agree with the observed ones except research by Singh and McAtackney (1998) and Andreou *et al*. (2009) who suggested that the data have to be cleaned first because the noise influence the prediction results much.

In Malaysia, as far as we concerned, no such research that use LAM to predict any system. So, that is why this present study has been carried out. We desire to see weather LAM can be utilized to predict our rainfall data. But, we must bear in mind that since the data have been proved by Radhakrishnan and Dinesh (2006) and Betty *et al*. (2010) are chaotic in behavior, so we cannot really predict the rainfall in long-term. We are trying our best to predict a week rainfall data (seven days).

In this study, we are using the original (pure) data. Simply said, our data are unfiltered data which contain noise (resulted from precipitation equipment or the rainfall contains other impurities). Since this is the first research in Malaysia that utilizes LAM, so, we just use all of the data that we have. But, if the prediction error is high, we might conduct another experiment with the reduce-noise data.

The next section will elaborate about the data, methodology, result and discussion. Lastly, there are some remarks of the whole study and few suggestions in order to obtain better prediction result.

Data

The daily rainfall data in *mm*^{3} unit which have been precipitated in Subang Meteorology Station in Selangor were analyzed in this study. The data from 1st January 1990 to 31st December 2005 were used to train the method while data from 1st January 2006 to 7th January 2006 were used to test the method's performance. This interpret that we used 5844 data to predict seven data ahead. Figure 1 shows the observed daily rainfall data.

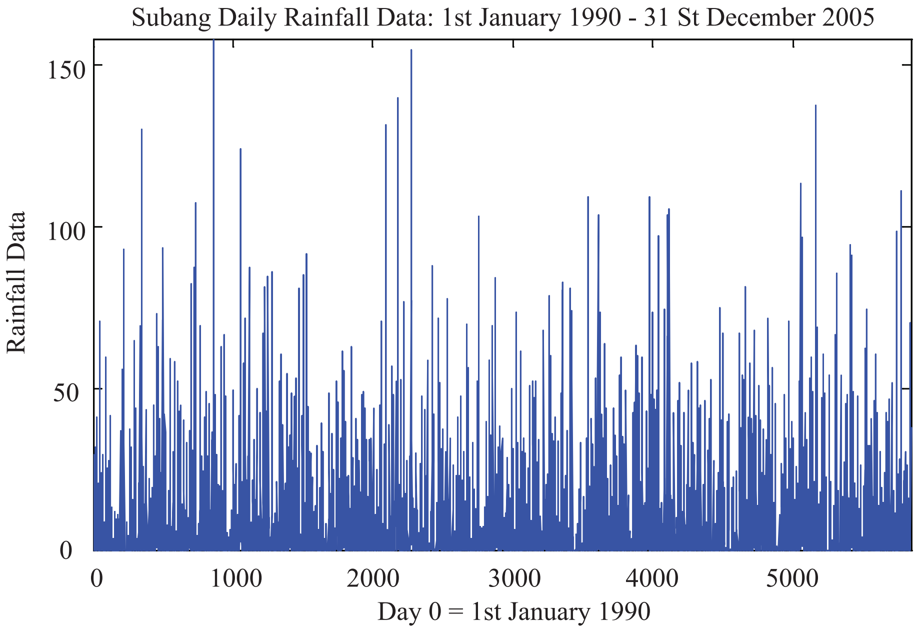

Figure 1 Subang Daily Rainfall Data

Methodology

Reconstruction of Phase Space (RPS)

The dynamics of time series $x_1, x_2, x_3, x_4, ..., x_N$ (Equation 1) are fully embedded in ***d***-dimensional phase space with delay time, $\boldsymbol{\tau}$ defined in $Y_t = \left\{ x_t, x_{t+\tau}, x_{t+2\tau}, ..., x_{t+(d-1)\tau} \right\}$ (Equation 2). The embedding dimension ***d*** can be thought as the minimum number of state variables required to describe the system and the delay time $\boldsymbol{\tau}$ is the average length of memory of the system (Regonda *et al*., 2005).

The dynamics of the system can be seen by the phase space whose trajectories describe its evolution from some initial state which is assumed to be known (Jayawardena and Lai, 1993). If the trajectories converge to some sub-space, then it is call an attractor.

Average Mutual Information (AMI)

The delay time, $\boldsymbol{\tau}$ needs to be chosen appropriately in order to fully capture the structure of attractor (Velickov, 2004). If $\boldsymbol{\tau}$ is too small, then the vectors of the space are not independent, and resulting the loss of attractors characteristics. But, if $\boldsymbol{\tau}$ is too large, the different coordinates may be almost dynamically uncorrelated and cause a loss of information of the original system (Regonda *et al*., 2005).

There are several ways to compute delay time, but according to Velickov (2004) the method of Average Mutual Information, AMI demonstrates a good performance. The AMI of lag time *T* is *I(T)*;

$$I(T) = \frac{1}{N} \sum_{k=1}^{N} p(u_k, u_{k+T}) \log_2 \left[\frac{p(u_k, u_{k+T})}{p(u_k) p(u_{k+T})} \right] \qquad \text{(Equation 3)}$$

where $p(u_k)$ and $p(u_{k+T})$ are probability of observing u_k and u_{k+T} in observed time series and $p(u_k, u_{k+T})$ is the joint probability of observing u_k and u_{k+T}. The delay time τ is the first minimum value of T from the graph T versus *I(T)*.

False Nearest Neighbor (FNN)

Three famous methods for estimating embedding dimension are Grassberger-Procacia method, FNN and Cao method. This present study chooses FNN method because Abarbanel (1996) stated that the method is much more robust.

The FNN algorithms are as follows. Say reconstructed phase space at time *i* with dimension *p* and delay time, τ is $Y_i = \left\{ x_i, x_{i+\tau}, x_{i+2\tau}, \ldots, x_{i+(p-1)\tau} \right\}$. *p* was started at 3 because Takens theorem (in Velickov, 2004) state that in order to preserve the topological of the original data attractor, it was suggested that $d \geq 2m+1$, where *m* is the dimension of original data. Since our observed rainfall data are in scalar vector, this interpret that $m = 1$. Hence, $d \geq 3$.

Say Y_j is the nearest neighbor to Y_i with minimum Euclidean distance, where $Y_j = \left\{ x_j, x_{i+j}, x_{i+2j}, \ldots, x_{i+(p-1)j} \right\}$. But, if

$$\frac{\left| x_{i+p\tau} - x_{j+p\tau} \right|}{\left\| Y_i - Y_j \right\|} \geq R_t \qquad \text{(Equation 4)}$$

then $Y_j$is not the nearest neighbor, or simply said, it is a false nearest neighbor (FNN) of Y_i. Here, $|\bullet|$ is a symbol for absolute value and $\|\bullet\|$ is a symbol for Euclidean distance. R_t is some threshold common range from 10 to 30. For all points *i* in dimension *p*, (Equation 4) was performed. The percentage of FNN is then calculated.

The algorithm was repeated for increasing *p* until the percentage of FNN drops to zero or some acceptable small number as 1%. If when $p = d$, the percentage of FNN drops to zero, then *d* is the embedding dimension. The dimension *d* can be thought as the minimum number of state variables required to describe the system. Besides, *d* also represents the number of column in the phase space of (Equation 2). The proper work need to be done in order to choose *d* because it is wasteful in term of time and computerization to construct more column than its necessary, but if the value of *d* is not sufficient, the state space will not accurately reflect the true topology of the system attractor (Velickov, 2004).

Local Approximation Method (LAM)

The method of local approximation is illustrated as follow. Say we have time series $x_1, x_2, x_3, x_4, \ldots., x_N$ where *N* is total days. Recall the reconstructed phase space at time *t* with embedding dimension *d* and delay time τ is $Y_t = \left\{ x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(d-1)\tau} \right\}$ for $t = 1, 2, 3, \ldots$ For example, say we have $N = 30$, $\tau = 10$, $d = 3$ and we want to predict the value of x_{31}. The reconstructed phase spaces are $Y_1 = \left\{ x_1, x_{1+10}, x_{1+20} \right\} = \left\{ x_1, x_{11}, x_{21} \right\}$, $Y_2 = \left\{ x_2, x_{12}, x_{22} \right\}$, and so on. It was found that the phase spaces end at $Y_{10} = \left\{ x_{10}, x_{20}, x_{30} \right\}$ while the value of x_{31}contains in $Y_{11} = \left\{ x_{11}, x_{21}, x_{31} \right\}$.

By LAM, if we want to predict the value of x_{31} in Y_{11}, we have to find the nearest phase neighbor (or simply said, the nearest neighbor) to the last known phase space, Y_{10}. Say Y_2 is the nearest neighbor with minimum Euclidean distance to Y_{10}. Then, Y_3 is an approximation toY_{11}. As $Y_3 = \{x_3, x_{13}, x_{23}\}$, then the value of $x_{31}$was approximated to x_{23}. The algorithm was repeated with $N = 31$, to find the approximate value of x_{32} and so on.

Performance Measure

In order to measure the performance of the approach, nowadays, there are many types of performance measures which are using to compare the predicted and observed value, such as absolute error, average absolute error, mean square error, root mean square error and correlation coefficient.

In this study, average absolute error was used to observe the average different value between the predicted and the observed ones. The formulas are:

$$e = \frac{\sum_{i=1}^{k} \left| y_{observed(i)} - y_{predicted(i)} \right|}{k} \qquad \text{(Equation 5)}$$

with k number of predicted days.

Results and Discussion

The result of AMI is as Figure 2. According to AMI method, the suitable delay time τ is the first minimum lag time T . Referring to Figure 2, τ for Subang data is 6. The result of FNN is as Figure 3. As mentioned in methodology section, the value of d was chosen upon its corresponding FNN percentage. If the percentage drops to 1%, than the match d is the embedding dimension. However, after long calculation, we found that the value of FNN percentage is not going to drop to 1%. So, we decided to choose the dimension with minimum value of FNN percentage as d. Referring to Figure 3, d for Subang data is 4.

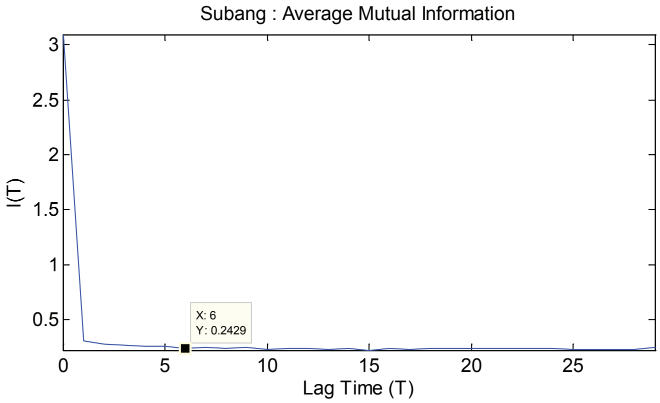

Figure 2 Result of AMI

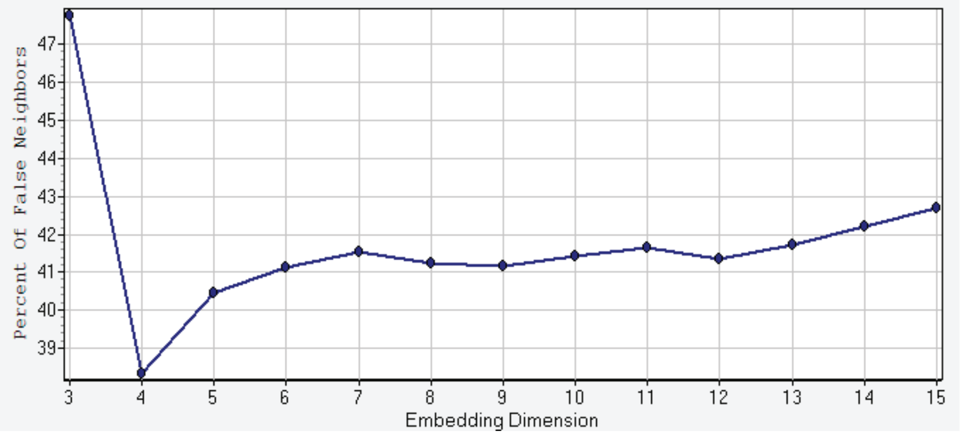

Figure 3 Result of FNN

The value of τ and d give the information that the reconstructed phase spaces are: $Y_1 = \{x_1, x_7, x_{13}, x_{19}\}$, $Y_2 = \{x_2, x_8, x_{14}, x_{20}\}$ and so on. These RPS end at $Y_{5826} = \{x_{5826}, x_{5832}, x_{5838}, x_{5844}\}$. Next, by utilizing LAM, the rainfall data value of day 5845 was predicted via algorithms that have been described in methodology section. The algorithms were repeated to predict the rainfall data value of day 5846 and ahead.

The results of prediction are as Fig. 4. The value of e is 27.33. From Figure 4, we can observe that five of seven days predictions are good. Our findings are quite similar to Jayawardena (1997), where except at a few high values, all of the predictions generally seem to agree with the observed value. Since the value of e is quite high, overall, we conclude that the prediction was not so agreed with the observed ones. These resulted that it is either the approach needs to be modified and improved or the data are not suitable to be used along with the approach. Besides, these also might influenced by the noise that contains in our observed data.

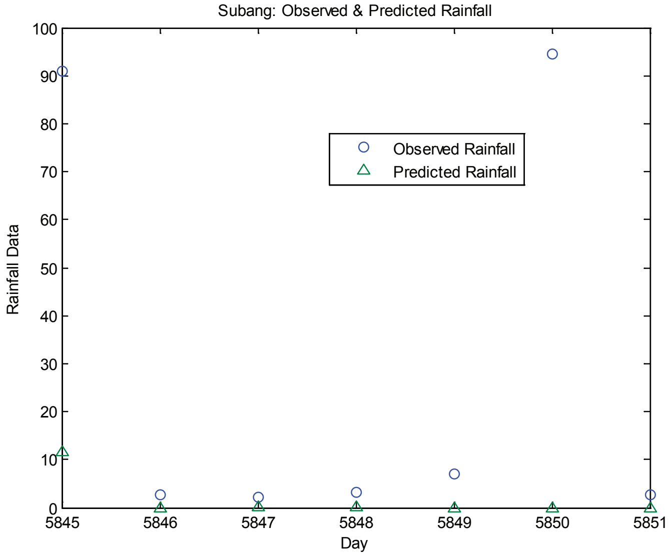

Figure 4 Results of Prediction

Conclusion and Further Studies

Prediction has been done on the daily rainfall data from Subang Meteorology Station, Selangor, located in Malaysia. Out of seven days predictions, the results are found to be in good agreement with the observed ones for five days. Even though the results seem well, the values of overall prediction errors are quite high.

As a conclusion, from the value of errors obtained, we admit that the predictions via LAM on Subang rainfalls data are at weak level. But since this approach has been tested by other researchers, and their results are good, we try to find out why this approach results high error to our experiment. After some discussions, we realized that the purely data that we used may contain noise that give big impact to the prediction results; and the approach needs to be modified in order to suit our observed data.

As mentioned in the introduction part, LAM is widely used in hydrology area. A lot of researches have been done and almost all predictions seem promising good approximation with the observed ones. So, it is suggested that the observed data need to be cleaned first and the prediction method needs to be modified in order to suit our data. As suggested by Singh and McAtackney (1998), we will choose Fourier analysis as a noise filter in our next experiment.

As mentioned before, our findings are quite similar to Jayawardena (1997), which except at a few high values, all of the predicted data generally seem to agree with the observed ones. So, we are planning to study the suitable method in order to fix the problem of those isolated (high) data values.

As explain in methodology part, FNN was chosen as the method to find the value of embedding dimension, *d*. The value of *d* was found once the percentage of FNN drop to 1%. However, after evaluating the value of *d* until 15, we cannot find any value of *d* that its corresponding FNN percentage near 1%. So, we just chose the value of *d* with the minimum FNN percentage. We are suspecting that the value of *d* which is 4 is not our desired *d*. So, in our next experiment, we will consider the method of Grassberger-Procacia or Cao in order to find the value of *d*.

In LAM, we employed the concept that in order to predict weather tomorrow, one looks in the past (the nearest neighbor) to find closest weather pattern. But in Velickov (2004), it was argued whether one neighbor is enough for such prediction. So, we plan to make a modification of the method by considering several neighbors of the phase space as used by Jayawardena and Lai (1993), Jayawardena (1997) and Mekonnen and Jayawardena (2004) in their studies.

Sugihara and May (1990) in Pao-Shan *et al*. (2004) suggested that the number of neighbors are *d* +1, with *d* is the embedding dimension. So, we might experiment this theory in our future study. We also looking forward to implement the approximation method which fits several nearest neighbors to a functional form as been done by Sivakumar *et al*. (1999), Sivakumar *et al*.(2000), Sivakumar *et al*. (2001), Islam and Sivakumar (2002), Sivakumar (2002) and Sivakumar *et al*. (2002) in their researches.

References

Abarbanel, H. D. I. (1996). *Analysis of Observed Chaotic Data*. New York: Springer-Verlag, Inc.

Andreou, Andreas, S., Zombanakis, George, A., Georgopoulos, E. F. & Likothanassis, S. D. (2009).

Forecasting Exchange-Rates via Local Approximation Methods and Neural Networks. Mpra Paper No. 17764.

Betty, V. W. N., Mohd. S. M. N. & Fredolin, T. (2010). Deterministic Behavior in Malaysian Rainfall. *Proceedings of Applied Mathematics International Conference 2010*.

Farmer, J. D. & Sidorowich, J. J. (1987). Predicting Chaotic Time Series. *Physical Review Letters.* 59(8): 845-848.

Islam, M. N. & Sivakumar, B. (2002). Characterization and Prediction of Runoff Dynamics: A Nonlinear Dynamical View. *Advances in Water Resources.* 25: 179–190.

Jayawardena, A. W. & Lai, F. (1993). Chaos in Hydrological Time Series. Extreme Hydrological Events: Precipitation, Floods and Droughts. *Proceedings of the Yokohama Symposium*, 213: 59-66.

Jayawardena, A. W. (1997). *Runoff Forecasting Using a Local Approximation Method*. IAHS, pp. 167-171.

Khan, S., Ganguly A. R. & Saigal, S. (2005). Detection and predictive modeling of chaos in finite hydrological time series. *Nonlinear Processes in Geophysics.* 12: 41–53.

Mekonnen, Z. Z. & Jayawardena, A. W. (2004). Development of Flood Forecasting Model in Middle Awash River Basin of Ethiopia. http://www.icharm.pwri.go.jp/master/publication/pdf/2010/zinas/pdf.

Pao-Shan, Y., Shien-Tsung, C., Che-Chuan, W. & Shu-Chen, L. (2004). Comparison of Grey and Phase-space Rainfall Forecasting Models using a Fuzzy Decision Method. *Hydrological Sciences* 49(4): 655–671.

Radhakrishnan, P. & Dinesh, S. (2006). An Alternative Approach to Characterize Time Series Data: Case Study on Malaysian Rainfall Data. *Chaos, Solitons and Fractals.* 27: 511–518.

Regonda, S. K., Rajagopalan, B., Lall, U., Clark, M. & Moon, Y. I. (2005). Local Polynomial Method for Ensemble Forecast of Time Series. *Nonlinear Processes in Geophysics*. 12: 397-406.

She, D. & Yang, X. (2010). A New Adaptive Local Linear Prediction Method and Its Application in Hydrological Time Series. *Mathematical Problems in Engineering Volume* 2010, Artikel ID 205438. http:/www.hindawi.com/journal/mpe/2010/205438/.

Siek, M. & Solomatine, D. P. (2011). Real-time Data Assimilation for Chaotic Storm Surge Model Using NARX Neural Network. *Journal of Coastal Research, Proceedings of the 11th International Coastal Symposium*, pp. 1189 – 1194.

Singh, S. & McAtackney, P. (1998). Dynamic Time-Series Forecasting using Local Approximation. *Proceedings of 10th IEEE International Conference on Tools with AI*, Taiwan, pp. 392-399.

Sivakumar, B. (2002). A Phase Space Reconstruction Approach to Prediction of Suspended Sediment Concentration in Rivers. *Journal of Hydrology* 258: 149-162.

Sivakumar, B., Berndtsson, R. & Persson, M. (2001). Monthly Runoff Prediction Using Phase Space Reconstruction. *Hydrological Sciences Journal* 46(3): 377-387.

Sivakumar, B., Berndtsson, R., Olsson, J., Jinno, K. & Kawamura, A. (2000). Dynamics of Monthly Rainfall-Runoff Process at the Gota Basin: A Search for Chaos. *Hydrology and Earth System Sciences*. 4(3): 407-417.

Sivakumar, B., Jayawardena A. W. & Fernando, T. M. K. G. (2002). River Flow Forecasting: Use of Phase-Space Reconstruction and Artificial Neural Networks approaches. *Journal of Hydrology.* 265: 225–245.

Sivakumar, B., Liong, S., Liaw, C. & Phoon, K. (1999). Singapore Rainfall Behavior: Chaotic?. *Journal of Hydrologic Engineering*, pp. 38-48.

Utusan Malaysia Online. (2011). http://utusan-malaysia-online.com/tiga-maut-banjir-semakin-buruk-di-empat-negeri.

Velickov, S. (2004). *Nonlinear Dynamics and Chaos*. London. Taylor & Francis Group plc.

Wu, C. L. & Chau, K. W. (2010). Data-driven Models for Monthly Streamflow Time Series Prediction. *Engineering Applications of Artificial Intelligence.* 23(8): 1350-1367.