19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Exploring detrending techniques in detecting Long-Memory of ozone time series in Malaysia by simulation

Muzirah Musa[a] **and Abdul Aziz Jemain**[b]

[a]*Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak.*
[b] *School of Mathematical Sciences, Faculty of Science and Technology,Universiti Kebangsaan Malaysia 43600, Bangi, Selangor, Malaysia.*
Email: muzirah@fst.upsi.edu.my

Abtract:

Air pollutants, and specifically ozone in the atmosphere, have received extensive attention since last few decades, mainly because of their adverse effect on people's health. Generally, the collected ozone data are often recorded as time series and long-memory behaviour in ozone levels usually exist. Long-memory or persistency is one of the statistical properties in time series which can be estimated by the *Hurst* coefficient, *H* determination. Currently, many methods to estimate *H* are available. Most of them, even if very effective, need prior information to be applied (in particular about the stationary nature of the series). In order to assess the long-term ozone behaviour in Malaysia, this study aimed to explore the role of detrending techniques of three existing methods used in detecting long-memory. Simulation series in the range of $0.1 \le H \le 0.9$ without any assumption on the stationary nature of the time series were used to detect long-memory. The quality of estimation was evaluated in terms of biases and variability. These methods are then applied to the daily mean hourly ozone concentration at 6 monitoring stations in Malaysia over 9 years. Our aim is to plan an optimal procedure to estimate the value of the *Hurst* coefficient and in addition to explain the degree of persistency in long-term ozone concentration in data Malaysia.

Keywords: *Long-memory, Persistency, Hurst coefficient, Stationarity, Ozone.*

1. INTRODUCTION

Ozone is a secondary pollutant and is formed by photochemical reactions involving oxides of nitrogen and volatile organic compounds in sunlight. Human activities are causing changes in ozone levels in much of the atmosphere. There have been numerous studies on ozone pollution research since this variable is one of particular importance due to its adverse affect on human and ecosystem health (Atkinson-Palombo et al. 2006; Chelani 2009; Pudasainee et al. 2010; Varotsos et al. 2005). Generally, the collected ozone data are often recorded as time series and are characterized by many large fluctuations with autocorrelation that is difficult to interpret (Lee et al. 2006). Understanding the statistical properties of ozone behaviour is considered as the most highly significant scientific topic and is critically important for pollution control measures.

The first step in environmental management is to find out exactly what is going wrong and why, and whether it is getting worse. The second is to make the information available to those who need to assess or manage the environment; governmental agencies, scientists, industrialist, concerned organizations and the public. This study seeks to evaluate several methods to detect the presence of long memory in time series and investigate the possible relationship between the ozone concentration behaviour and the time scales.

Long memory is a feature of many geophysical time series and commonly used to describe persistent dependence within the time series observations as lag increases. In recent studies, it has been shown that many time series in the nature are characterized by self-similarity and scale invariance. These characteristics exhibit what is known as long memory or high persistence that implies a strong correlation between the successive data points. (Chelani 2009; Kai et al. 2008; Lee et al. 2006; Varotsos; Davidoff 2006; Varotsos et al. 2005; Weng et al. 2008). Statistically, long-memory sometimes described by the autocorrelation function decays slower than exponential decay with increasing lag, possibly by a power-law decay. According to Beran (1992), the slowly decaying correlation is called *Hurst Effect* or *Hurst Phenomenon* which closely can be related to climate changes.

In the present of long-memory, the widely applied Box and Jenkins models for analyzing short-memory series are no longer appropriate. Consequently, long-memory analysis is pertinent for obtaining accurate information regarding the variability of high persistent in the data. In detecting a long-memory, most of studies suggested that *Hurst coefficient* (*H*) is suitable as a long-memory indicator. There are several methods that can be used to determine the *H* values. The most commonly used method are rescaled range analysis and detrended fluctuation analysis.

However, Wang et. al, (2007) in their study reported that the detection of long memory is affected by some uncertainty and it would be better to use several estimators so as to increase reliability of the estimation. This uncertainty is related to issues of non-stationary time series and the methods used. Since, data points are often non-stationary or have means, variances and covariance that change over time, and each tools or method has different performance, prerequisite conditions and limitations, therefore each needs thorough evaluation in order to avoid bias or misinterpretation of the derived *Hurst* coefficient, *H*.

The aim of this study is to plan an optimal procedure to estimate the value of Hurst's coefficient, using existing methods without any prior knowledge about stationary nature of the time series. This study investigates three existing methods used in detecting long memory by exploring detrending technique through simulation. Then, these methods are applied to daily mean hourly ozone concentration data in Malaysia, recorded at six monitoring stations over 9 years. In section 2, the methods used to detect long-memory will be described briefly. The implementation of the Davis & Harte Simulation (Davies; Harte 1987) is presented in section 3. The methods then applied to daily mean hourly ozone concentration data series to detect the existence of long-memory are discussed in section 4, and finally, some conclusions are drawn in section 5.

2. METHODS FOR DETECTING THE EXISTENCE OF LONG-MEMORY

Methods for detecting long-memory share a common structure. It involves the determination of the fluctuation of the detrended series. Three different trend removal techniques were explored. For linear trends, the simple linear and the line connecting the first and last point in an interval (bridge) detrended were used. Next, the polynomial trend with degree 2 and 3 were examined. These two techniques were applied into the *Windowed Local Detrended method* (*WLD*) and *Detrended Fluctuation method* (*DF*). Finally the simple and exponential weighted moving average trend removal techniques were applied. This method is called

Detrended Moving Average (DMA). The difference between these methods are that, the original series $\{x_j\}$ was used for the *WLD* method and the integrated series $\{y_k\}$ was used for the *DF* and *DMA* methods. Next, in detecting the presence of long-memory in the series, the fluctuation is determined by calculating the root mean square of the detrended series. Mathematically given as

$$f(r)=\sqrt{\frac{1}{T}\sum\left(r-\hat{r}_{(\bullet,\bullet)}\right)^2} \tag{1}$$

where r is the series, T is the total number of observation and $\hat{r}_{(\bullet,\bullet)}$ is the theoretical value of the trend considered. Mathematically given that

$$r\in\begin{cases}\{x_j\}=\{x_{n(i-1)+j}\} & for \quad j=1,2,...,N \ and \ i=1,2,...,M\\ \{y_k\}=\{y_{n(i-1)+k}\} & for \quad y_k=\sum_{j=1}^{k\in[1,n]}\left(x_j-\bar{x}\right)\end{cases} \tag{2}$$

where *j* and *k* be the number of observations and i be the interval of size *n* with $M=N/n$. The detrending techniques used

$$\hat{r}_{(\bullet,\bullet)}\in\begin{cases}\hat{r}_{(l,b)}=(1-w_j)r_{i1}+w_jr_{in}\ ;\ w_j=(j-1)/n & for \text{ linear-bridge}\\ \hat{r}_{(p,d)}\ =\ \hat{a}+\hat{b}j+\hat{c}j^2+...+\hat{z}j^p & for \text{ polynomial with choosen } d\in\{1,2,3\}\\ \hat{r}_{(m,s)}\ =\ \dfrac{1}{n}\sum_{j=0}^{n-1}r_{k-j} & for \text{ simple moving average}\\ \hat{r}_{(m,w)}\ =\ (1-\lambda)r_k+\lambda\tilde{r}_{k-1}\quad ;\ \lambda=n/(n+1) & for \text{ exponential weighted moving average}\end{cases} \tag{3}$$

The algorithms for the three methods are as follows:

2.1. Windowed Local Detrended method (WLD)

This method was adopted from the scaled windowed variance (SWV) method developed by Cannon et al. (1997). The series $\{x_j\}$ is partitioned into *M* non-overlapping intervals of equal length *n*. Then, for each interval *i*, the standard deviation, S_i is calculated using the formula

$$S_i=\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(x_j-\hat{x}\right)^2} \tag{4}$$

where $\hat{x}=\hat{x}_i$ is a local trend in each interval *i* which is determined by assumption of linear and polynomial trend. Finally the average standard deviation, $\bar{S}$ of all intervals of length *n* is computed. The computation is repeated for different length size *n* and the relationship between $\bar{S}$ and *n* is given by

$$\bar{S}\propto n^H \tag{5}$$

where *H* is the degree of persistency of the series.

2.2. Detrended Fluctuation method (DF)

A Detrended Fluctuation (DF) method begins with constructing the integrated series y_k. The y_k is a cumulative sum of deviation of the series from its mean, $\bar{x}$ as given in (2). Then, the y_k series is partitioned into M non-overlapping intervals of equal size length *n*. In each interval, the assumptions of linear and polynomial trend were then locally detrended. A Fluctuation for this integrated and detrended series for a given length *n* is calculated by

$$\sigma_n=\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(y_k-\hat{y}\ \right)^2} \tag{6}$$

which $\hat{y}$ represent the assumption of linear and polynomial trend. By repeating this calculation for different length *n*, the relationship between σ_n and *n* is given by

$$\sigma_n \propto n^{\alpha} \tag{7}$$

where $\alpha = H + 1$.

2.3. Detrended Moving Average method (DMA)

A Detrended Moving average (DMA) method looks very similar to DF. The main difference one meets here is that instead of linear and polynomial detrendisation procedure in equally length *n*, one uses moving average of a given length *n*. The determination of σ_n given in (6) was done by substitute $\hat{y}$ with the theoretical values of simple and exponential weighted MA, $\tilde{y}$.

$$\sigma_n = \sqrt{\frac{1}{N-n+1}\sum_{k=1}^{N}\left(y_k - \tilde{y}\ \right)^2} \tag{8}$$

By repeating this calculation for different length *n*, the relationship between σ_n and *n* is given by the power law relationship as in (7). The quantity of *H* was obtained from the log-log plot of σ_n as a function of *n*. The existence of long-memory in data series are determined by the values of *H*. *Hurst* coefficient with $0.5 < H < 1$, characterizes persistent or positively correlated time series while $0 < H < 0.5$ indicates anti-persistent or negatively correlated time series . For $H = 0.5$ the time series are said to be independent or random series (Beran 1994).

3. DAVIS & HARTE SIMULATION

An extensive Davis and Harte (1987) investigation was performed in order to find out how good the previously described methods for detecting the presence of long-memory. The main advantage of this method is the speed. The source code, the *Splus* programs is available in Beran (1994). The simulation begins with generating 40 fractional Gaussian noise (fGn) series of 2048 data points for each of 9 values of *H* ranging from 0.1 to 0.9 by steps of 0.1. These series were then cumulatively summed to obtain the corresponding fractional Brownian motions (fBm). fBm is non-stationary with time dependent variance. Each method was applied on the entire series (2048 points) and then on the first 1024, 512, 256, 128 and 64 points. In order to facilitate comparisons, the same series lengths were adopted for all methods used to estimate *H* ($\hat{H}$). For each length of series, the mean, standard deviation, bias and mean squared error, MSE of $\hat{H}$ are calculated. The performance of all methods is measured in terms of bias and variability.

4. RESULTS AND DISCUSSION

4.1. Result of simulations

Results concerning the effect of series length on variability in *H* estimation with eight detrending techniques are displayed in figure 1. For WLD (K1) method, the detrending techniques used were simple linear (M1), linear bridge (M2), quadratic (M3) and cubic (M4). DF (K2) method considered simple linear (M5) and quadratic (M6) detrending techniques. Finally, simple moving average (M7) and exponential weighted moving average (M8) techniques were applied for DMA (K3) method. The biasness of estimation is determined by the different between theoretical values of *H* (true *H*) and estimated *H* ($\hat{H}$).Values of MSE (variance + bias^2) of $\hat{H}$ was used as performance indicator for comparative purpose. Figure 1 show that, the four techniques tested (M1-M4), gave essentially similar results as MSE values approximates to zero which approaches over the entire value of true *H*. These techniques seemed to provide very accurate mean estimates of *H* even for very short series where a smallest MSE is generally interpreted as best explaining the variability in the observation.

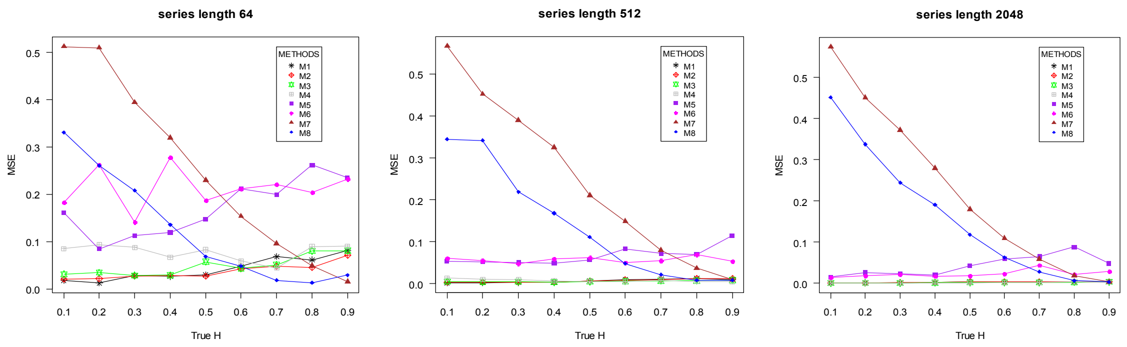


Figure 1. Plots of MSE of $\hat{H}$ versus true H . Each box compares detrending methods M1-M8. The result are shown for the series length n = 64, 512 and 2048.

The techniques M6 and M7 shows the variations of estimates tended to decrease as series length increased. However, the variation was slightly larger for $H > 0.5$. While for M7 and M8, the variations of estimates tended to decreased as series length and H increased. Clearly, the least biased method for this series is WLD (M1-M4). Overall of this finding may suggest that WLD with four detrending techniques give the best result in detecting long memory process.

4.2. Result with ozone concentrations data

In Malaysia, a network of air quality monitoring stations is run on behalf of the Air Quality Division of the Department of the Environment, Malaysia (DOE) through long-term monitoring by a private company, Alam Sekitar Sdn Bhd (ASMA). The ozone concentrations data was recorded as part of a Malaysian Continuous Air Quality Monitoring (CAQM) program based on UV absorption. The pollutant measurements are performed on an hourly basis and reported in part per million (*ppm*).

For the purpose of the study, data were taken from six urban background stations (S1-S6) which well represent an urban area in Peninsular Malaysia. These sampling stations are surrounded by industrial, residential and commercial areas and consequently, congested roads. Due to rapid urbanization, industrialization and transportation, these areas were expected to be affected by the pollutants which discharged continuously from the human activities'. Figure 2 shows geographical location of the study area. Most of the stations have a data recorded spanning the period from January 1998 to December 2006 except for station S2 which data taken from December 1998 to December 2006.

The daily mean hourly ozone concentration for a 9-year time span were used and the time series plots at two out of six stations are shown in Figure 3. The time series, however, reveal implicit seasonal trend in the data. To overcome the natural non-stationarity in the data, an anomaly method proposed by Windsor and Toumi (2001) was applied to remove the seasonal trend.

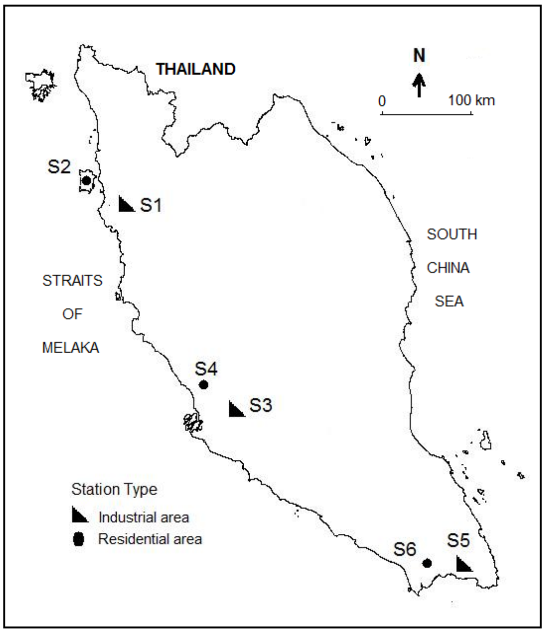
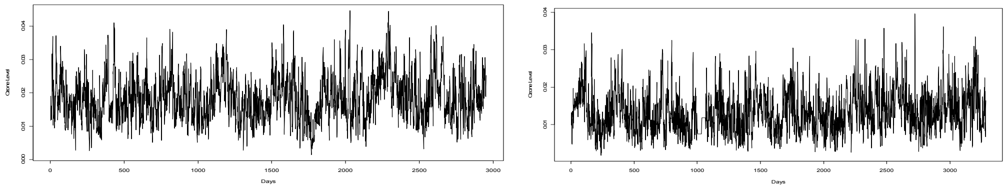


Figure 2: Map indicating location of study area in Peninsular Malaysia

The anomaly was calculated by subtracting the hourly average for that day of the year from the observed values. Then, the daily mean hourly ozone concentrations are obtained and used for further analysis.



(a) (b)

Figure 3. Series of daily mean hourly ozone concentration at (a) S2 and (b) S3 monitoring stations.

The results of detecting long-memory in daily mean hourly ozone concentrations using eight detrending techniques are reported in Table 2. All methods give similar results in detecting the presence of long memory for all data series. The estimated H values ($\hat{H}$) varied between 0.47 and 0.96 (± 0.01 and ± 0.05) shows that long-memory exist in ozone data series for each station. The effect of different detrending techniques used is clearly observed for methods K1 and K2 when the linear detrending (M1,M2,M5) and polynomial detrending (M3,M4,M6) techniques are considered in removing the local trends. For K1 method, the $\hat{H}$ values for M3 and M4 lie in the range of 0.78 to 0.96, while $\hat{H}$ values for M1 and M2 are in between 0.63 and 0.88. For K2 method, the $\hat{H}$ values for M6 are varied in between 0.82 and 0.97. For M5, the $\hat{H}$ values lie in between 0.68 and 0.93. As can be seen, the estimated H values ($\hat{H}$) are about 10% increased when the polynomial detrending techniques are used in detecting a long-memory in ozone data series. However, method K3 doesn't give much different values of values $\hat{H}$ for both M7 and M8 detrending technique used. Their values vary between the ranges of 0.47 to 0.77. The results show that, the evidence of the existence of a long-memory is found in daily mean hourly ozone concentration in urban area of Peninsular Malaysia.

Table 2: Estimated H ($\hat{H}$) values for six monitoring stations.

		METHODS							
		K1				K2		K3	
		M1	M2	M3	M4	M5	M6	M7	M8
		$\hat{H}$ (± *sd*)	$\hat{H}$ (± *sd*)	$\hat{H}$ (± *sd*)	$\hat{H}$ (± *sd*)	$\hat{H}$ (± *sd*)	$\hat{H}$ (± *sd*)	$\hat{H}$ (± *sd*)	$\hat{H}$ (± *sd*)
STATIONS	S1	0.78 (*0.02*)	0.78 (*0.01*)	0.87 (*0.02*)	0.88 (*0.02*)	0.68 (*0.05*)	0.86 (*0.05*)	0.68 (*0.01*)	0.68 (*0.01*)
	S2	0.75 (*0.02*)	0.71 (*0.02*)	0.83 (*0.02*)	0.89 (*0.01*)	0.93 (*0.04*)	0.97 (*0.03*)	0.57 (*0.01*)	0.56 (*0.01*)
	S3	0.85 (*0.02*)	0.86 (*0.01*)	0.91 (*0.01*)	0.91 (*0.01*)	0.77 (*0.03*)	0.86 (*0.03*)	0.77 (*0.01*)	0.76 (*0.01*)
	S4	0.85 (*0.02*)	0.88 (*0.01*)	0.96 (*0.01*)	0.95 (*0.01*)	0.92 (*0.04*)	0.96 (*0.03*)	0.54 (*0.01*)	0.53 (*0.01*)
	S5	0.75 (*0.01*)	0.73 (*0.01*)	0.80 (*0.01*)	0.83 (*0.01*)	0.74 (*0.03*)	0.84 (*0.03*)	0.67 (*0.01*)	0.66 (*0.01*)
	S6	0.68 (*0.01*)	0.63 (*0.01*)	0.78 (*0.01*)	0.81 (*0.01*)	0.68 (*0.03*)	0.82 (*0.02*)	0.48 (*0.01*)	0.47 (*0.01*)

Regarding to the previous studies done in investigating a long-memory or persistency of air pollutant, our results reached similar conclusion with many researchers. Studies done by Windsor and Toumi (2001) applied R/S, Sigma-T and Kurtosis analysis to average hourly ozone, PM_{10} and $PM_{2.5}$ series in UK with mean *Hurst coefficient* estimator of 0.77,0.80, 0.77, respectively. Varotsos et al.(2005) used Detrended fluctuation analysis (*DFA*) to identify a long-memory of hourly observations of ozone, NO_x, PM_{10} and $PM_{2.5}$ in Greece and Maryland. Their result showed the *Hurst* value for the daytime ozone was found to range from 0.87 to 0.89 and 0.79 to 0.81 for the nightime. While those for NO_x ranged from 0.71 to 0.73. The *Hurst* value for PM_{10} and $PM_{2.5}$ were found to vary from 0.88 to 0.93 and 1.19 to 1.21, respectively. Weng et al.(2008) in their study, concluded that daily maximum hourly ozone time series in Taiwan was identified as persistent and long-memory with Hurst value 0.75. While Kai et al. (2008) analyzed SO_2, NO_2, PM_{10} pollution indexes and daily air pollution indexes (API) of Shanghai using R/S, DFA and spectrum. The result shows that the Hurst values obtained is 0.81 for SO_2, 0.84 to 0.92 for NO_2 and 0.78 to 0.81 for PM_{10}. Whereas, for API was found from 0.78 to 0.83. R/S analysis on hourly ozone data series in Delhi, done by Chelani (2009) found that *Hurst coefficient* estimator are with 0.77 for whole data series over year 2006 and above 0.88 for monthly analysis.

5. CONCLUSION

Air pollution is one of the most important environmental problems attracting concerns of environmentalists, policy makers and the public. It is well known that ozone is chiefly related to pollutants emission but the mechanics that derives its temporal evolutions are not understood very clearly. In the literature, many studies have been carried out on different aspect of ozone concentration and one of the aspect is detecting long memory. The statistical approach for the detection of long memory is affected by some uncertainty. It would be better to use several estimators so as to increase the reliability of the estimation. Exploring detrending techniques used in this analysis in an attempt to reduce bias in the *H* values estimated and in providing the best procedure for detecting long-memory in the time series. By simulation, three existing methods with two different approaches of detrending techniques; local detrended (WLD and DF) and smoothed dynamic detrended (DMA), were explored with several trend removal techniques. WLD methods with four detrending techniques gave best result in detecting long-memory for all short and long simulated series. However, the results also show that DF and DMA methods can be used for the long series. The three methods with several trend removal techniques were applied to daily mean hourly ozone concentrations from six monitoring stations in Malaysia and show that the existence of long-memory or persistency in the ozone data series through the determination of *Hurst* coefficient, *H*. The knowledge of the persistence in ozone offers a better understanding of the dynamics of the system governing the ozone formation. Consequently, this information would help the decision maker in sustaining the environmental monitoring system.

ACKNOWLEDGEMENT

The authors are grateful to the Air Quality Division of the Department of the Environment, Malaysia, (DOE); Alam Sekitar Sdn Bhd (ASMA) for providing the air quality data for this study. Special thanks to Universiti Pendidikan Sultan Idris, Universiti Kebangsaan Malaysia and Ministry of Higher Education, Malaysia for their sponsorship. Without whom this research would not have been possible. Funded Grant No [UKM-AP-2011-19].

REFERENCES

Atkinson-Palombo, C. M., J. A. Miller, and J. R. C. Balling (2006). Quantifying the ozone "weekend effect" at various locations in Phoenix, Arizona. *Atmospheric Environment*, 40 (39), 7644-7658.

Beran, J. (1992). Statistical Methods for Data with Long-Range Dependence. *Statistical Science*, 7 (4), 404-416.

Beran, J.(1994). *Statistics for long-memory processes,* Chapman & Hall, New York.

Cannon, M. J., D. B. Percival, D. C. Caccia, G. M. Raymond, and J. B. Bassingthwaighte (1997). Evaluating scaled windowed variance methods for estimating the Hurst coefficient of time series. *Physica A: Statistical and Theoretical Physics*, 241 (3-4), 606-626.

Chelani, A. B. (2009). Statistical persistence analysis of hourly ground level ozone concentrations in Delhi. *Atmospheric Research*, 92 (2), 244-250.

Davies, R. B., and D. S. Harte (1987). Tests for Hurst Effect. *Biometrika*, 74 (1), 95-101.

Kai, S., L. Chun-qiong, A. Nan-shan, and Z. Xiao-hong (2008). Using three methods to investigate time-scaling properties in air pollution indexes time series. *Nonlinear Analysis: Real World Applications*, 9 (2), 693-707.

Lee, C.-K., L.-C. Juang, C.-C. Wang, Y.-Y. Liao, C.-C. Yu, Y.-C. Liu, and D.-S. Ho (2006). Scaling characteristics in ozone concentration time series (OCTS). *Chemosphere*, 62 (6), 934-946.

Pudasainee, D., B. Sapkota, A. Bhatnagar, S.-H. Kim, and Y.-C. Seo (2010). Influence of weekdays, weekends and bandhas on surface ozone in Kathmandu valley. *Atmospheric Research*, 95 (2-3), 150-156.

Varotsos, C., and D. K. Davidoff (2006). Long-memory processes in ozone and temperature variations at the region 60 S–60 N. *Atmospheric Chemistry and Physics*, 6, 4093-4100.

Varotsos, C., J. Ondov, and M. Efstathiou (2005). Scaling properties of air pollution in Athens, Greece and Baltimore, Maryland. *Atmospheric Environment*, 39 (22), 4041-4047.

Wang, W., P. H. A. J. M. Van Gelder, J. K. Vrijling, and X. Chen (2007). Detecting long-memory: Monte Carlo simulations and application to daily streamflow processes. *Hydrology and Earth System Sciences*, 11 (2), 851-862.

Weng, Y.-C., N.-B. Chang, and T. Y. Lee (2008). Nonlinear time series analysis of ground-level ozone dynamics in Southern Taiwan. *Journal of Environmental Management*, 87 (3), 405-414.

Windsor, H. L., and R. Toumi (2001). Scaling and persistence of UK pollution. *Atmospheric Environment*, 35 (27), 4545-4556.