

# Assessment of Risk of Rainfall Events with a Hybrid of ARFIMA-GARCH

Ibrahim Lawal Kane<sup>1</sup> & Fadhilah Yusof<sup>1</sup>

<sup>1</sup> Department of Mathematical Sciences, Faculty of Science, University of Technology Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

Correspondence: Ibrahim Lawal Kane, Department of Mathematical Sciences, Faculty of Science, University of Technology Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia. E-mail: [ibrahimlk131@gmail.com](mailto:ibrahimlk131@gmail.com)

Received: October 18, 2013      Accepted: November 13, 2013      Online Published: November 28, 2013

doi:10.5539/mas.v7n12p78

URL: <http://dx.doi.org/10.5539/mas.v7n12p78>

## Abstract

Hazardous situations related to rainfall events can be due to very intense rainfall, or to the persistence of rainfall over a long period of time. Such events may result to an exceedence of the capacity of drainage systems resulting in the heap of basements which may lead to landslides and flooding. This study assesses the persistence dependence of rainfall time series of Chui Chak, a station in Peninsular Malaysia that observed the highest rainfall event for the period 01/01/1975-31/12/2008. The persistence dependence of the rainfall time series was modelled via fractional ARIMA model augmented with the GARCH model. The Ljung-Box test for testing autocorrelation proves that the combined ARFIMA-GARCH model captures the temporal persistence behaviour in the Chui Chak rainfall time series data with persistence measure 0.839. This measure represents a relatively lasting persistence, that is, the process variability should return to the historical average after a relatively long period of time which may have a risk of extreme event.

**Keywords:** kriging, ARFIMA, GARCH, McLeod-Li test, persistence

## 1. Introduction

Conventionally, rainfall has been considered as a random process with several peculiarities, mostly associated to intermittency and non-Gaussian behaviour (Papalexioiu et al., 2011). Hazardous situations related to rainfall events can be due to very intense rainfall, or to the persistence of rainfall over a long period of time. Such events may give rise to an exceedence of the capacity of sewer systems resulting in the inundation of roads and basements which may lead to landslides or flooding. To prevent such hazards, an analysis of hydrological and climatological records may aid in modelling by giving enhanced understanding of the processes that might derived the extremes.

Persistence dependence is generally recognized in hydrological, climatological and financial time series over a varied range of time scales. Rainfall analysis have a substantial role in the successful planning, development and implementation of water resource management to evaluate engineering works and environmental problems such as hydropower generation, reservoir operation, flood control and water quality control. Henceforth, an efficient study of rainfall temporal behaviour is the critical mission in hydrology (Modarres & Ouarda, 2013). Most of the traditional methods for measuring the risk related with behaviour of the data set are done through study of the variance or volatility. Nevertheless, the study of the variance or volatility does not give explanation that there might be a predictable component in the data.

Stochastic models are designed to replicate the imperative behaviour or patterns evident in the data set grounded on the current knowledge of the process. The application of stochastic models to hydrology and water resources management is mainly for two important explanations. Firstly, stochastic models provide a basis for decisions making. Secondly, stochastic models also increase the understanding of physical phenomena. This is the crucial goal of the physical sciences. A worthy modelling framework will extract previously unknown insight from the data so as to plan for the future.

The dependence structure of rainfall data is typically very complex both in space and time both in time. Daily rainfall data sets from dense networks of rain-gauge stations in peninsular Malaysia are commonly analysed in this paper using ordinary Kriging technique to identify the location that realized highest rainfall for the period

under study. Very high rainfall events have a substantial effect on society and may lead to loss of life and property. To conduct an in-depth study on the temporal dependence the daily rainfall series of the identified location with highest rainfall event would be very vital.

Simple linear or nonlinear stochastic models may not always be appropriate in modelling the temporal dependence in rainfall time series. Certainly, rainfall time series data usually have both linear and nonlinear patterns. To adequately fit these chaotic time series, other more complicated models that have the ability to capture the dynamics of the series more precisely have to be taken into account (Yusof & Kane, 2013). There are a number of studies on time series modelling and forecasting that argue that performance improves in hybrid models (Zhang, 2003). Numerous authors published a lot of papers combining different models e.g., ARIMA with GARCH, ARFIMA with GARCH, HMM with GARCH, ANN with SVM etc. (Chau & Wu, 2010) investigates the joint effect of the hybrid of ANN-SVR model and the singular spectrum analysis on improving the accuracy of daily rainfall forecasting. Their results indicated that the hybrid of support vector regression model performed the best. Sallehuddin et al. (2007) proposed a model that integrates nonlinear Grey Relational Artificial Neural Network (GRANN) and linear ARIMA model. The forecasting ability of the model was compared with several other models, which include: ARIMA, Multiple Regression, GRANN, etc. The result have shown that the proposed hybrid of GRANN\_ARIMA model has outperformed other models with 99.5% forecasting accuracy for small-scale data and 99.84% on large-scale data. Yin (2008) proposed a variant of an ARFIMA process by adding an explanatory variable into the ARFIMA long memory model. Comparing the forecast ability of the proposed model with other models such as; ARFIMA, the Random Walk and the ARMA process of the order one, the result show that the proposed model outperform other models in terms of out sample forecasting.

Suhartono and Muhammad (2011) proposed a hybrid model to study trends and seasonal fuzzy time series. The proposed hybrid model is based on a Winter's model and weighted fuzzy time series. The result revealed that more accurate forecast is obtained from the proposed hybrid method. It can be observed that most hybrid models performed better than the individual models in terms of forecast ability and accuracy.

Rainfall time series is often characterized by long memory behavior (Yusof et al., 2013) and are highly skewed with volatility clustering (Villarini et al., 2010). Long memory is an asymptotic property of a stochastic process, characterizing the decay of the correlation between observations separated by increasingly large time lags. The fractionally integrated process is used in explicitly modeling of the long term memory structure. The possible way of representing the correlation structures is by the use of fractional differencing order  $d$ . It has been demonstrated that a stationary process will possess long memory behavior if  $0 < d < 0.5$ . For  $-0.5 < d < 0$ , the process is considered by a slow decay of autocorrelation but does not retain the long memory property. Therefore, in this case, the process is said to be anti-persistent. This behavior is often related to an over differentiation problem by the first difference filter. For  $0.5 < d < 1$ , the case is called non-stationary mean-reverting (Claude & Vivien, 2002). The testing of hydrological data with Long Memory models was done by number of authors (Koutsoyiannis, 2006; Koscielny-Bunde et al., 2006; Mudelsee, 2007; Wang et al., 2007; Gil-Alana, 2012; Yusof et al., 2013) and found it to be a common characteristic in hydrological data.

The volatility clustering indicates that the residual of the series exhibits time-varying heteroskedastic effect, that is, unconditional standard deviations are not constant through time. These characteristics can well be captured using the GARCH family models. The motivation for using both models comes from the notion that rainfall time series is characterised by both linear and non-linear pattern, therefore neither one of the models can identify the true data generating process DGP (Terui & Dijk, 2002). Modarres and Ouarda (2013) investigated the advantages of non-linear GARCH model over linear ARIMA model using stream flow data of the Matapedia River, Quebec, Canada. The result shows that the linear ARIMA model became inadequate for modelling stream flow time series due to the existence of Heteroskedasticity in the residuals of the ARIMA model as shown by Engle test. Therefore, combining an ARIMA (13, 1, 4) and GARCH (3, 1) models fitted the data very well. Therefore, recommended that "the application of a GARCH model is strongly suggested for hydrological time series modelling as the conditional variance of the residuals of the linear models can be removed and the efficiency of the model will be improved".

Wang et al. (2005) verified the presence of conditional heteroskedasticity in the residuals from linear models fitted to the daily and monthly stream flow processes of the upper Yellow River using the McLeod-Li and Engle's Lagrange Multiplier tests. In a related study, HongRui et al. (2012) established a GARCH model to the daily runoff data for the period 1949–2001 of Yichang hydrological station. Szolgayova (2011) investigates whether heteroskedastic effect can be detected in selected time series of daily discharges in Slovakia. The results show that heteroskedasticity was present in all of the tested time series. Other application of

heteroskedastic models to hydrological data can be seen in (Chen et al., 2008; Laux et al., 2011; Yusof & Kane, 2013).

In this study, we first analyse the spatial variability of network of rainfall stations to identify the location with the highest rainfall event in peninsular Malaysia. The temporal pattern of the rainfall data of the identified location is then modelled via long memory model augmented with the Generalised Autoregressive Conditional Heteroskedastic (GARCH) model taking in mind the observed features of the hydrological time series. The combined model would be good in capturing the persistence dependence in the rainfall data. The remaining of the paper is organized as follows: The methodology employed in the spatial analysis and the temporal modelling that captures both the linear and nonlinear behaviour of the daily rainfall series is presented in section 2. Section 3 presents and discusses the results of the findings. Section 4 gives the conclusions and recommendations.

## 2. Methods

### 2.1 Rainfall Spatial Variability

Kriging uses semivariance to measure the spatially correlated component, a component that is also called spatial dependence or spatial autocorrelation which expresses the spatial dependence between neighboring observations. The semivariogram quantifies the relationship between the semivariance and the distance between sampling pairs. A semivariogram plots the average semivariance against the average distance, this function may be used alone as a measure of spatial autocorrelation (Kang, 2012).

The **semivariance** is computed by:

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} (z(s_i) - z(s_i + h))^2$$

where  $\gamma(h)$  is the average semivariance between known points  $s_i$  separated by distance  $h$ , the number of pairs of sample points arranged by direction in the bin is denoted by  $n$  and  $z$  is the attribute value. Semivariogram is modeled by fitting a theoretical function such as: Spherical, Exponential, Linear and Gaussian, to ensure that the solution is unbiased and has minimum variance (Kang, 2012). After having a suitable semivariogram model, the kriging technique can then be applied in estimating the value of  $z(s_0)$  in each grid point  $s_0$  where no value is available. The Kriging estimate is of the linear form:

$$\hat{z}(s_0) = \sum_{i=1}^N \lambda_i z(s_i) \quad (1)$$

Where  $z(s_i)$  is the measured value in the  $i$ th location,  $\lambda_i$  is an unknown weight for the measured value at the  $i$ th location,  $s_0$  is the prediction location and  $N$  is the number of measured values (Journel & Huijbregts, 1978).

The issue is to find a set of weighted coefficients in such a way that the kriging estimator satisfied the conditions of unbiasedness and the minimum estimation variance respectively, that is:

$$\sum_{i=1}^N \lambda_i = 1$$

$$\sigma^2 = E\{[\hat{z}(s_0) - z(s_0)]^2\}$$

These two constraints make the kriging technique a good interpolator against other interpolation methods.

### 2.2 Testing and Estimation of Fractional Degree of Integration ( $d$ )

In testing and estimation of the fractional order of integration, Geweke and Porter-Hudak (1983) proposed the log-periodogram regression estimate and has been the most widely used (Shimotsu, 2002). Consider a fractional integrated process  $\{Y_t\}$ , its spectral density is given by:

$$f(\omega) = [2 \sin(\omega / 2)]^{-2d} f_u(\omega) \quad (2)$$

where  $\omega$  is the Fourier frequency,  $f_u(\omega)$  is the spectral density corresponding to  $u_t$  and  $u_t$  is a stationary short memory noise with 0 mean. Consider the set of harmonic frequencies  $\omega_j = 2\pi j / n, j = 0, 1, \dots, n/2$ , where  $n$  is the sample size. Taking the logarithm of Equation (2) we have:

$$\ln f(\omega_j) = \ln f_u(\omega_j) - d \ln[4 \sin^2(\omega_j / 2)], j = 0, 1, \dots, n/2 \quad (3)$$

Equation (3) can be re-written in an alternative form following (Wang et al., 2007) as:

$$\ln f(\omega_j) = \ln f_u(0) - d \ln[4 \sin^2(\omega_j/2)] + \ln \left[ \frac{f_u(\omega_j)}{f_u(0)} \right], j = 0, 1, \dots, n/2 \quad (4)$$

The fractional order of differencing  $d$  can be estimated using the regression equation in Equation (4). Using the periodogram estimate of  $f(\omega_j)$ , if the number of frequencies  $m$  used in Equation (4) is a function  $g(n)$  (a positive integer) of the sample size  $n$ , where  $m = g(n) = n^\alpha$  with  $0 < \alpha < 1$ , it can be demonstrated that the least squares estimate  $\hat{d}$  using the above regression is asymptotically normally distributed in large samples

$$\hat{d} \sim N \left( d, \frac{\pi^2}{6 \sum_{j=1}^{g(n)} (U_j - \bar{U})^2} \right)$$

where  $U_j = \ln[4 \sin^2(\omega_j/2)]$  and  $\bar{U}$  is the sample mean of  $U_j, j = 1 \dots g(n)$ .

Under the null hypothesis of no long memory ( $d = 0$ ), the t-statistic

$$t_{d=0} = \hat{d} \left( \frac{\pi^2}{6 \sum_{j=1}^{g(n)} (U_j - \bar{U})^2} \right)^{-1/2}$$

has limiting standard normal distribution.

The value of the power factor  $\alpha$  is the main determinant of the ordinates included in the regression, for more details see (Yusof et al., 2013).

### 2.3 ARFIMA ( $p, d, q$ ) Model

Fractionally integrated processes are an extension of the class of Box and Jenkins methodology known as ARIMA processes. Autoregressive fractionally integrated moving average models (ARFIMA) grant more freedom and flexibility in modelling processes with long memory behaviour. Modelling data with fractionally integrated processes allows capturing and modelling slower rates of decay of autocorrelation than with other common time series methods such as AR, MA or ARIMA models (Baillie, 1996). The general ARFIMA ( $p, d, q$ ) model was first introduced by Granger and Joyeux (1980) and Hosking (1981) and is formulated as:

$$\Phi(B)(1-B)^d X(t) = \Theta(B)\varepsilon_t \quad (5)$$

where  $\varepsilon_t$  is a white noise process with mean zero and variance  $\sigma^2$ ,  $B$  is the lag operator,  $\Phi(B)$  and  $\Theta(B)$  are polynomials of orders  $p$  and  $q$  denoted  $(1 - \phi_1 B - \dots - \phi_p B^p)$  and  $(1 - \theta_1 B - \dots - \theta_q B^q)$  respectively, with roots outside of the unit circle. The standard differencing operator  $(1 - B)$  of an ARIMA processes is replaced with a fractional operator  $(1 - B)^d$  where  $d$  is the fractional differencing parameter. The aim is to find the order of the AR and MA processes. The order of differencing can be decided using Equation (4). There exist several techniques for selecting the order ARMA model. One of them is based on the information criteria. The idea is to minimize the risk of under fitting that is, selecting an order smaller than the true order, and over fitting that is, selecting an order larger than the true order. One of the mathematical formulation of the parsimony criterion of model building is proposed by Akaike (1974) named Akaike Information Criterion (AIC). This method was formulated for the purpose of selecting an optimal model fits to a given time series data. Mathematically the AIC is formulated as:

$$\text{AIC}(M) = \ln \hat{\sigma}^2 + \frac{2(p+q)}{n}$$

where  $p$  and  $q$  are the order of autoregressive and moving average parameters to estimate. The model with minimum AIC can be selected as the parsimonious model (Yurekli & Kurunc, 2007).

#### 2.3.1 Method of Estimating ARFIMA ( $p, d, q$ ) Model

For a tentative formulation of a model, identification of the model parameters, then it is needed to obtain the co-efficient estimates of the parameters. After the parameters have been estimated, the fitted model is then

subjected to diagnostic checks. Since the order of differencing has been estimated already, now we consider an ARMA model of the form:

$$Y = c + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \quad (6)$$

To estimate the value of  $(c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2)$  on the basis of observations on  $Y$ , The principle will be based on maximum likelihood estimation. This method makes it possible to estimate the parameters of the process. Let  $\theta = (c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2)'$  denote the vector of population parameters. Suppose we have observed a sample of size  $T (y_1, y_2, \dots, y_T)$ , the approach is to calculate the joint probability density:

$$f_{y_T, y_{T-1}, \dots, y_1}(y_T, y_{T-1}, \dots, y_1; \theta) \quad (7)$$

which implies the probability of having observed this particular sample. The maximum likelihood estimate (MLE) of  $\mu$  is the value for which this sample is most likely to have been observed; that is, it is the value of  $\mu$  that maximizes (7).

### 2.3.2 Test for Heteroskedasticity

The identification of conditional heteroskedasticity is often based on testing whether squared residuals from the ARFIMA model are autocorrelated (Bollerslev & Mikkelsen, 1996; Andersen & Bollerslev, 1998) among many others. Testing for the lack of correlation of particular residuals, can be carried out using different methods, among the methods the McLeod-Li Test would be implemented in this paper.

This test for heteroskedasticity was proposed by McLeod and Li (1983). It looks at the autocorrelation function of the squares of the residuals from the fitted model and tests whether the correlation between  $(x_t^2, x_{t-j}^2)$  is not 0 for some  $j$ . The autocorrelation at the  $j$  lag for the squared residuals  $\{x_t^2\}$  is estimated using:

$$\hat{\varepsilon}(j) = \frac{\sum_{t=1}^N (x_t^2 - \sigma^2)(x_{t-j}^2 - \sigma^2)}{\sum_{t=1}^N (x_t^2 - \hat{\sigma}^2)} \quad (8)$$

For the null hypothesis that  $x_t$  is an independently identically distributed (i.i.d) process, McLeod and Li (1983) demonstrated that, for fixed  $L$ ,  $\sqrt{N}\varepsilon = (\varepsilon(1), \dots, \varepsilon(L))$  is asymptotically a multivariate unit normal. Consequently, if  $L$  is sufficiently large, the usual Ljung-Box (LB) statistic:

$$Q = N(N+2) \sum_{j=1}^L \frac{\hat{\varepsilon}_j^2}{(N-j)}$$

is asymptotically  $\chi^2(L)$ . If the probability of  $Q$  is higher than  $\alpha = 0.05$ , then is a strong evidence that the residuals are uncorrelated and the model is suitable for the data set. If the probability is less than  $\alpha = 0.05$ , then the residuals are correlated and the model is inadequate, therefore need to repeat the model building process to achieve an adequate model (Modarres & Ouarda, 2012).

### 2.4 GARCH Modeling

The ARFIMA model is adequate for modelling linear dependence of rainfall data; however, the model cannot capture the heteroskedasticity in the time series typically observed in the form of existence of serial dependence in squares of the model residuals. Therefore, the residual  $\varepsilon_t$  from the fitted model is modelled using GARCH family models to reduce or remove the heteroskedasticity.

Introduced by Engle (1982), the Autoregressive Conditional Heteroskedastic (ARCH) model, and generalized by Bollerslev (1986), named Generalized Autoregressive Conditional Heteroskedastic model (GARCH). The term "conditional" implies the level of association on the past sequence of observations and the "autoregressive" describes the feedback mechanism that incorporates past observations into the present (Laux et al., 2011).

The variance equation of the GARCH ( $p, q$ ) model can be expressed as:

$$\varepsilon_t = z_t \sigma_t, \quad z_t \sim \psi_\tau(0, 1) \quad (9)$$

$$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 = \omega + \alpha(B) \varepsilon_t^2 + \beta(B) \sigma_t^2 \quad (10)$$

where  $\varepsilon_t$  is the residuals from the fitted ARFIMA model and  $\psi_\tau(0, 1)$  denote the pdf (probability density function) of the residuals with 0 mean and variance 1.  $\tau$  is the distributional parameter that define the the shape of the distribution. The GARCH ( $p, q$ ) process is called a stationary process if the following conditions hold:

$$E(\varepsilon_t) = 0$$

$$\text{var}(\varepsilon_t) = \frac{\omega}{(1 - \alpha(1) - \beta(1))}$$

$$\text{cov}(\varepsilon_t, \varepsilon_s) = 0, \quad t \neq s, \text{ if and only if } \alpha(1) + \beta(1) < 1$$

The stationarity assumption in statistical analysis of a time series data is necessary. In GARCH modelling this considers the constraints on the estimated parameters in the maximum likelihood-estimation. It follows the theorem that states the restrictions on the estimated parameters in the GARCH ( $p, q$ ) model for stationarity in the GARCH process. The required and sufficient condition for the existence of a second order stationary solution is that:  $\alpha + \beta < 1$  holds. The simplest GARCH model is GARCH (1, 1) model (Hu et al., 2010).

### 2.5 Choice and Validation of Time Series Model

When a good model for the data generating process of a time series has been built, it is common practice to test for the model acceptability. Test for residuals autocorrelation is the prominent tool for this task. To judge whether ARCH effects and autocorrelation have been completely removed or not, a well-known example is a Portmanteau test for residual autocorrelation and ARCH-LM test. The former is implemented in this paper.

#### 2.5.1 Portmanteau Test for Residual Autocorrelation

A portmanteau test is a test used for investigating the presence of autocorrelation in time series. The portmanteau test checks the pair of hypotheses:

$H_0 : \rho_{u,i} = \dots = \rho_{u,h} = 0$  i.e. all lags correlations are zero.

$H_1 : \rho_{u,i} \neq 0$  for at least one  $i = 1, \dots, h$ , is tested, at least one lag with non-zero correlations.

where  $\rho_{u,i} = \text{corr}(u_t, u_{t-i})$  is the autocorrelation coefficients of the residual series. If  $\hat{u}_t$  are residuals from an estimated ARFIMA ( $p, d, q$ ) model, the portmanteau test statistic is

$$Q_h = T \sum_{j=1}^h \hat{\rho}_{u,j}^2,$$

where  $\hat{\rho}_{u,j} = T^{-1} \sum_{t=j+1}^T \hat{u}_t^s \hat{u}_{t-j}^s$  and  $\hat{u}_t^s = \hat{u}_t / \hat{\sigma}_u$  are the standardized estimation residuals. The test statistic has an approximate  $\chi^2(h - p - q)$ -distribution if the null hypothesis holds. An advance version with potentially better small sample properties was proposed by Ljung and Box (1978) given as:

$$Q_h^* = T^2 \sum_{j=1}^h \frac{1}{T-j} \hat{\rho}_{u,j}^2 \approx \chi^2(h - p - q) \quad (11)$$

The null hypothesis is rejected if probability value is less than the significance level of  $\alpha = 0.05$ .

## 3. Data Analysis and Discussion

### 3.1 Data Used

The daily average rainfall data of 75 rain-gauge stations across peninsular Malaysia for the period January 1975 to November 2008 obtained from the Malaysian meteorological department were used in this study.

#### 3.1.1 Results and Discussion

The rainfall data sets from 75 gauged stations were first analyzed using OK technique. The reason behind this is to identify the location that realized highest rainfall for the period under study. Figure 1 gives the result of the OK displaying the spatial variability of rainfall amount in peninsular Malaysia with the location with the highest average rainfall amount for the period under study.

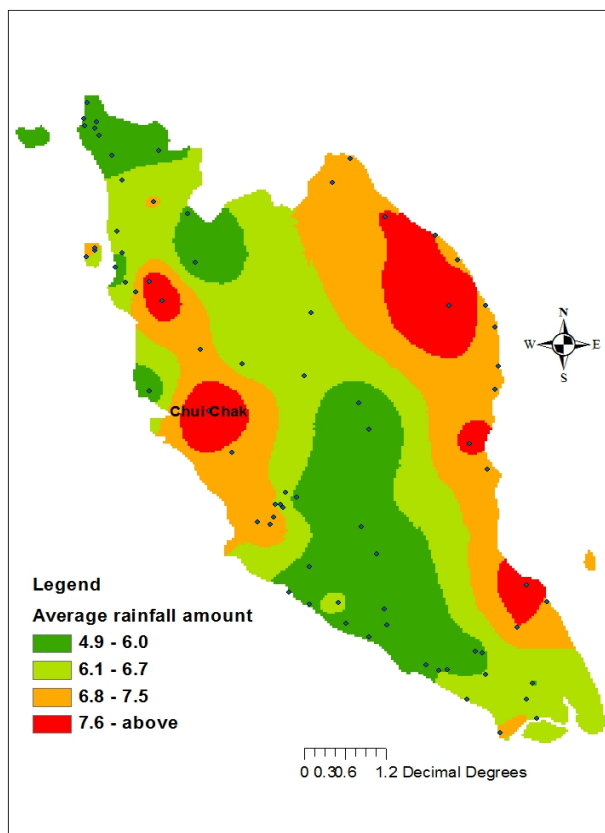


Figure 1. Spatial distribution of Peninsular Malaysia daily precipitations in millimetre averaged over the time period January 1975 to November 2008

It can be observed in (Figure 1) that some other stations falls within the range of 7.6 and above signifying areas high average rainfall amount, but the interest in this study is particularly on single station with the highest amount (the station was selected using a defined threshold in the arcGIS software) so as to modelled the temporal behaviour of the rainfall in the identified station. In modelling the temporal behaviour of the rainfall time series of the identified station, the daily rainfall time series for the period January 1975 to November 2008 of the identified station (Chui Chak) is used.

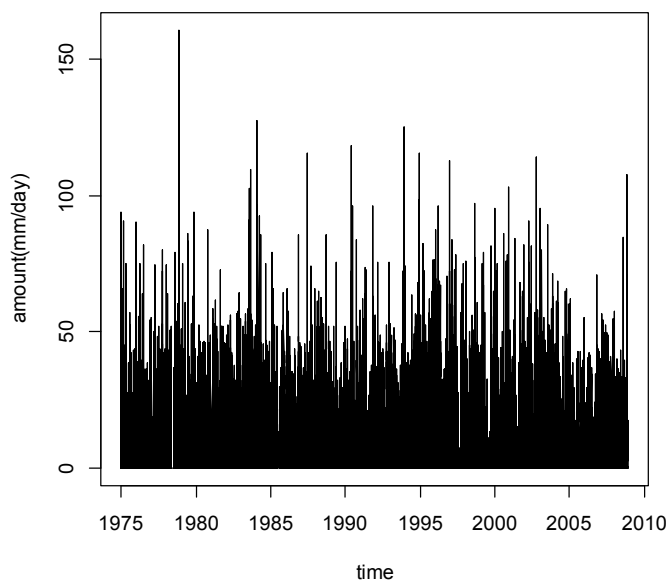


Figure 2. Time series plot for the daily rainfall of Chui Chak station

Figure 2 depicts time series plot of the daily rainfall of Chui Chak station, from the figure, the time series show a very persistent behaviour. The autocorrelation function (ACF) that provides a measure of temporal correlation between rainfall data points with different time lags is given in Figure 3. For a purely random event, all autocorrelation coefficients are zero, apart from lag 0, which is equal to 1 (Yusof et al., 2013). The autocorrelation function (Figure 3) decays slowly which indicate that the time series are strongly correlated which is a main feature of long memory appearance. Fractionally integrated processes possess genuine long memory in the sense that the present state of a system is temporally dependent on all past states (Rea et al., 2007).

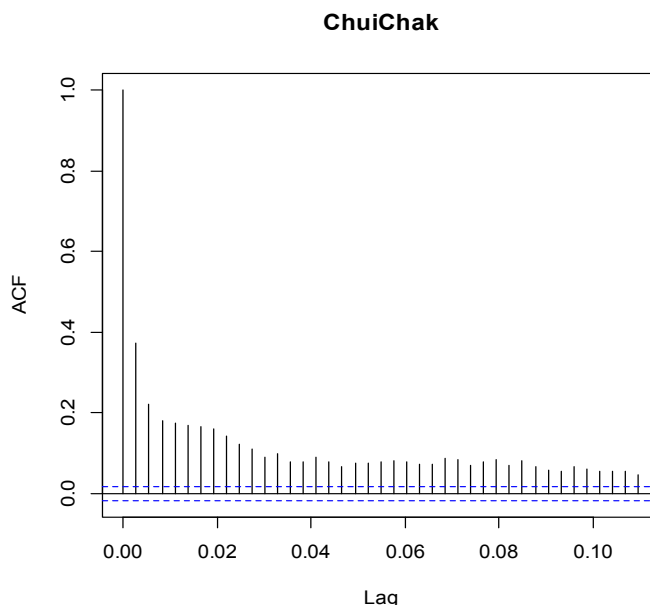


Figure 3. ACF plot for the daily rainfall of Chui Chak station

The approach here follows fitting a suitable ARFIMA model to the daily rainfall series of the station identified by first transforming the data via fractional integration. The Geweke-Porter-Hudak method (GPH) was employed to estimate the fractional order of differencing  $d$  using bandwidth of 0.8. The value of  $d$  was found to be 0.2070681 which falls within  $0 < d < 0.5$  indicating a long memory process. A stationary and invertible process can be represented either in a moving average form or in an autoregressive form. The problem of either of these representations, though, is that it may contain too many parameters even for a finite-order moving average and a finite order autoregressive model. In general, a large number of parameters reduce efficiency of the estimates. Hence, in building a good model, it may perhaps be essential to include both AR and MA terms in the model; this will leads to the useful mixture of ARMA process (Wei, 2006). The AR and MA orders  $p$  and  $q$  respectively of the ARFIMA model were identified and estimated using Box-Jenkins methodology. The results are given in Table 1.

Table 1. Summary ARFIMA-GARCH estimation results

Model	ARFIMA parameters							ARFIMA-GARCH parameters		
	$d$	$\psi_1$	$\psi_2$	$\varphi_1$	$\varphi_2$	$\varphi_3$	$\varphi_4$	$\omega$	$\alpha_1$	$\beta_1$
Estimate	0.207	1.45	-0.609	1.354	0.463	0.058	0.026	36.77	0.149	0.690
Std.err.	0.015	0.11	0.101	0.114	0.093	0.019	0.011	3.696	0.026	0.012



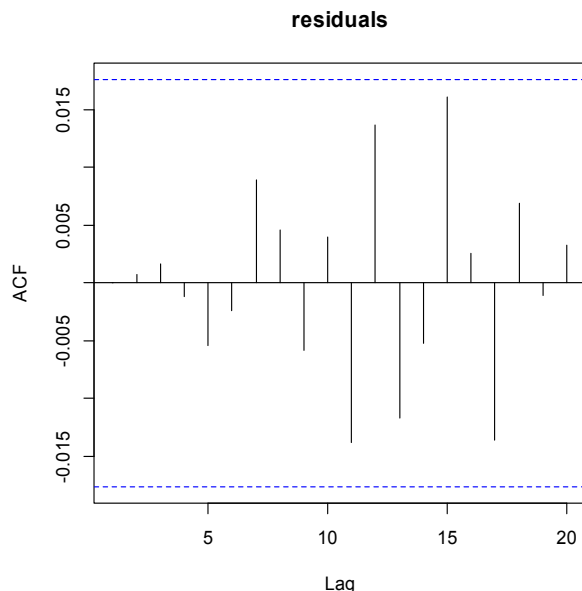


Figure 4. Autocorrelation function of residuals of the fitted ARFIMA model

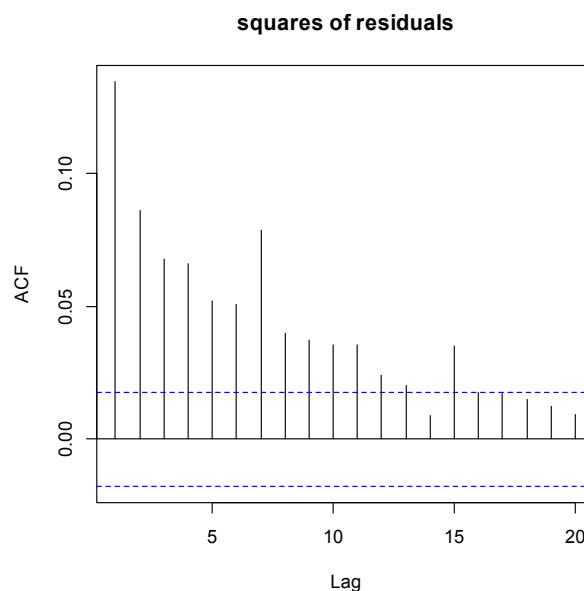


Figure 5. Autocorrelation functions of squared residuals of the fitted ARFIMA model

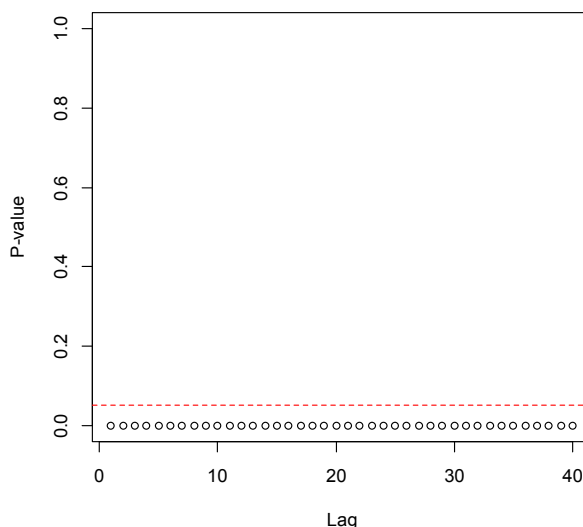


Figure 6. McLeod-Li test for the heteroskedastic effect in the residuals of the fitted ARFIMA model

The autocorrelation function (Figure 4) of the residual from the fitted ARFIMA model shows that the residuals are uncorrelated; therefore, the ARFIMA model seems to be adequate. However, the autocorrelation function (Figure 5) of squares of residual of the fitted ARFIMA model shows autocorrelation, this indicates the presence of heteroskedastic effect and confirmed by Mcleod-Li test given in (Figure 6) which can be modelled using GARCH model. As mentioned earlier, the linear models e.g. (ARMA, ARFIMA etc) can only be sufficient in modelling the information confined in the autocorrelation of the differentiated time series, which is a low requirement and is not capable of modelling Heteroskedasticity. Due to this problem, the maximum likelihood method to estimate the ARFIMA parameters used to be biased when the error variance is not constant (De Montera et al., 2008). The remedy to this consists in modelling Heteroskedasticity as a nonlinear relationship. Therefore the residuals from ARFIMA model is augmented by a GARCH model, the results are given in (Table 1). Table 2 gives the Ljung-Box test results for the ARFIMA-GARCH error model, it can be seen that both residual and squares of residual are uncorrelated meaning that no other heteroskedastic effect left and this is the necessary condition to conclude that the model is appropriate. Therefore the addition of the GARCH (1,1) specification for the ARFIMA error helps to capture the serial correlations in the squared residuals.

Table 2. Ljung-Box test results for the ARFIMA-GARCH model

Lag	Residuals		Sqr. Residuals	
	Statistic	P-value	Statistic	P-value
Q(10)	5.090112	0.885077	8.693195	0.561446
Q(15)	13.87522	0.535013	18.17034	0.253816
Q(20)	18.29635	0.567893	19.18531	0.509810

In the ARFIMA-GARCH (1, 1) model, the sum of  $\alpha$  and  $\beta$  measures the extent to which the variance of current volatility remains significant for long periods into the future (i.e. volatility persistence). As this sum approaches unity, the persistence of variability to the volatility becomes greater. Moreover, when the sum of  $\alpha$  and  $\beta$  equals to 1, then any variance to volatility is permanent and the unconditional variance is infinite. In this case, the process is denoted an IGARCH and implies that volatility persistence is permanent; hence, past volatility is significant in predicting future volatility over all finite horizons (McMillan & Thupayagale, 2011). The volatility is said to be explosive if the sum of  $\alpha$  and  $\beta$  is greater than 1, then, the variance to the volatility in one period will result in even greater volatility in the next period (Chou, 1988), as such the higher the volatility the riskier the security.

It could be noticed that the sum parameters value for the ARFIMA-GARCH model is relatively high i.e.  $\alpha + \beta = 0.839$ , therefore represent a relatively lasting persistence, that is the process variability should converge to the historical average after a relatively long period of time.

#### 4. Conclusions

The spatial variability of rainfall time series is modelled using Ordinary Kriging technique in identifying the location with highest rainfall event in peninsular Malaysia. Chui Chak station realised to be the identified station. Moreover, the temporal dependence of the rainfall time series of the identified station is modelled via fractional ARIMA model augmented with the GARCH model. The Ljung-Box test for testing autocorrelation proves that the combined ARFIMA-GARCH model captures the temporal behaviours in the Chui Chak rainfall time series data. Thus, modelling the dependence in level and variability enhances our understanding for the properties of the rainfall series. Therefore the model could be a very good model for modelling rainfall for water resource management.

#### References

- Akaike, H. (1974). A look at the statistical model identification. *IEEE transactions on automatic control, AC-19*, 716-723. <http://dx.doi.org/10.1109/TAC.1974.1100705>
- Andersen, T.G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review, 39*, 885-905. <http://dx.doi.org/10.2307/2527343>
- Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics, 73*, 5-9. [http://dx.doi.org/10.1016/0304-4076\(95\)01732-1](http://dx.doi.org/10.1016/0304-4076(95)01732-1)
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J Econometrics, 31*(3), 307-327. [http://dx.doi.org/10.1016/0304-4076\(86\)90063-1](http://dx.doi.org/10.1016/0304-4076(86)90063-1)
- Bollerslev, T., & Mikkelsen, H. O. (1996). Modeling and pricing long memory in stock market volatility. *Journal of Econometrics, 73*, 151-184. [http://dx.doi.org/10.1016/0304-4076\(95\)01736-4](http://dx.doi.org/10.1016/0304-4076(95)01736-4)
- Claude, D., & Vivien, G. (2002). Long Memory Time Series and Fractional Integration. *A Cliometric contribution to France and German Economic and Social History, 25*(314), 4-22.
- Chau, K. W., & Wu, C. L. (2010). A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *Journal of Hydroinformatics, 12*(4), 458-473. <http://dx.doi.org/10.2166/hydro.2010.032>
- Chen, S., Jeong, K., & Härdle, W. (2008). Support Vector Regression Based GARCH Model with Application to Forecasting Volatility of Financial Returns. *SFB 649 Discussion Paper*. Retrieved from <http://edoc.hu-berlin.de/series/sfb-649-papers/2008-14/PDF/14.pdf>
- Chou, R.Y. (1988). Persistent volatility and stock returns – some empirical evidence using ARCH. *Journal of Applied Econometrics, 3*, 279-94. <http://dx.doi.org/10.1002/jae.3950030404>

- De Montera, L., Mallet, C., Barthes, L., & Gole, P. (2008). Short-term prediction of rain attenuation level and volatility in Earth-to-Satellite links at EHF band. *Nonlinear Processes in Geophysics*, 15(4), 631-643. <http://dx.doi.org/10.5194/npg-15-631-2008>
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom inflation. *Econometrica*, 50, 987-1008. <http://dx.doi.org/10.2307/1912773>
- Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *J. Time Ser. Anal.*, 4, 221-237. <http://dx.doi.org/10.1111/j.1467-9892.1983.tb00371.x>
- Gil-Alana, L. A. (2012). U.K. Rainfall Data: A Long-Term Persistence Approach. *J. Appl. Meteorol. Clim.*, 51, 1904-1913. <http://dx.doi.org/10.1175/JAMC-D-11-0201.1>
- Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-range time series models and fractional differencing. *J. Time Ser. Anal.*, 1, 15-30. <http://dx.doi.org/10.1111/j.1467-9892.1980.tb00297.x>
- HongRui, W., Xiong, G., LongXia, Q., & Song, Y. (2012). Uncertainty analysis of hydrological processes based on ARMA-GARCH model. *Science China Technological Sciences*, 8(55), 2321-2331.
- Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, 68, 165-176. <http://dx.doi.org/10.1093/biomet/68.1.165>
- Hu, B., Song, X., Ming, Z., Wang, S., & Yang, J. (2010). Empirical Analysis on the Fluctuation of Increasing Rate of China's Energy Consumption. *Modern Applied Science*, 4(6), 60-66.
- Hurvich, C. M., Deo, R. S., & Brodsky, J. (1998). The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series, *Journal of Time Series Anal.*, 19, 19-46. <http://dx.doi.org/10.1111/1467-9892.00075>
- Journel, A. G., & Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic Press London.
- Kang, T. C. (2012). *Introduction to Geographic Information Systems* (6th ed., pp. 314-338). McGRAW HILL INTERNATIONAL EDITION, Singapore.
- Koscielny-Bunde, E., Kantelhardt, J. W., Braun, P., Bunde, A., & Havlin, S. (2006). Long-term persistence and multifractality of river runoff records: Detrended fluctuation studies. *J. Hydrol.*, 322, 120-137. <http://dx.doi.org/10.1016/j.jhydrol.2005.03.004>
- Koutsoyiannis, D. (2006). Nonstationarity versus scaling in hydrology. *J. Hydrol.*, 324, 239-254. <http://dx.doi.org/10.1016/j.jhydrol.2005.09.022>
- Laux, P., Vogl S., Qiu, W., Knoche, H. R., & Kunstmann, H. (2011). Copulabased statistical refinement of precipitation in RCM simulations over complex terrain. *Hydrol Earth Syst Sci Discuss*, 8, 3001-3045. <http://dx.doi.org/10.5194/hessd-8-3001-2011>
- Ljung, G. M., & Box, G. E. P. (1978). On a measure of a lack of fit in time series models. *Biometrika*, 65(2), 297-303. <http://dx.doi.org/10.1093/biomet/65.2.297>
- McLeod, A. I., & Li, W. K. (1983). Diagnostic checking ARMA time series models using squared residual autocorrelations. *Journal of Time Series Analysis*, 4, 269-273. <http://dx.doi.org/10.1111/j.1467-9892.1983.tb00373.x>
- McMillan, D., & Thupayagale, P. (2011). Measuring volatility persistence and long memory in the presence of structural breaks Evidence from African stock markets. *Managerial Finance*, 37(3), 219-241. <http://dx.doi.org/10.1108/03074351111113298>
- Modarres, R., & Ouarda, T. B. M. J. (2012). Generalized autoregressive conditional heteroscedasticity modelling of hydrologic time series. *Hydrol. Process*. <http://dx.doi.org/10.1080/02626667.2012.743662>
- Modarres, R., & Ouarda, T. B. M. J. (2013) Modelling heteroscedasticity of streamflow times series. *Hydrological Sciences Journal*, 58(1), 54-64. <http://dx.doi.org/10.1080/02626667.2012.743662>
- Mudelsee, M. (2007). Long memory of rivers from spatial aggregation. *Water Resour. Res.*, 43, W01202. <http://dx.doi.org/10.1029/2006WR005721>
- Papalexiou, S. M., Koutsoyiannis, D., & Montanari, A. (2011). Can a simple stochastic model generate rich patterns of rainfall events? *Journal of Hydrology*, 411, 279-289. <http://dx.doi.org/10.1016/j.jhydrol.2011.10.008>

- Rea, W., Reale, M., & Brown, J. A. (2007). Long Memory or Structural Breaks in Temperature and Proxy Time Series. Christchurch, New Zealand: *MODSIM 2007 International Congress on Modeling and Simulation Proceedings* (pp. 3010-3016).
- Sallehuddin, R., Mariyam, S., Shamsuddin, Hj., Zaiton, M. H. S., & Ajith, A. (2007). Forecasting Time Series Data using Hybrid Grey Relational Artificial Neural Network and Auto Regressive Integrated Moving Average Model. *ICS AS CR 2007. Neural Network World* 6/07, 573-605.
- Shimotsu, K. (2002). *Exact local whittle estimation of fractional integration with unknown mean and time trend*. Department of Economics, University of Essex (pp. 1-2).
- Suhartono, & Muhammad, H. L. (2011). A Hybrid Approach based on Winter's Model and Weighted Fuzzy. Time Series for Forecasting Trend and Seasonal Data. *Journal of Mathematics and Statistics*, 7(3), 177-183.
- Szolgayova, E. (2011). Modelling and forecasting daily river discharge considering autoregressive heteroskedasticity. *Geophys Res Abstr*, 13, EGU2011-10681-1.
- Terui, N., & Van Dijk, H. K. (2002). Predictability in the Shape of the Term Structure of Interest Rates. *International Journal of Forecasting*, 18, 421-438. [http://dx.doi.org/10.1016/S0169-2070\(01\)00120-0](http://dx.doi.org/10.1016/S0169-2070(01)00120-0)
- Villarini, G., Smith, J. A., & Napolitano, F. (2010). Nonstationary modeling of a long record of rainfall and temperature over Rome. *Advances in Water Resources*, 33(10), 1256-1267. <http://dx.doi.org/10.1016/j.advwatres.2010.03.013>
- Wang, W., Van Gelder, P. H. A. J. M., Vrijling, J. K., & Ma, J. (2005). Testing and modelling autoregressive conditional heteroskedasticity of streamflow processes. *Nonlin. Processes Geophys.*, 12, 55-66. <http://dx.doi.org/10.5194/npg-12-55-2005>
- Wang, W., Van Gelder, P. H. A. J. M., Vrijling, J. K., & Chen, X. (2007). Detecting long-memory: Monte Carlo simulations and application to daily streamflow processes, *Hydrol. Earth Syst. Sci.*, 11, 851-862. <http://dx.doi.org/10.5194/hess-11-851-2007>
- Wei, W. W. S. (2006). *Time series analysis: univariate and multivariate methods* (2nd ed.).
- Yin, Y. C., & Hoe, Q. S. (2008). A New Variant of ARFIMA Process and Its Predictive Ability. *Modern Applied Sciences*, 2(2), 142-159.
- Yurekli, K., & Kurunc, A. (2007). Performances of Stochastic Approaches in Generating Low Streamflow Data for Drought Analysis. *Journal of special hydrology*, 5(1), 22-24.
- Yusof, F., & Kane, I. L. (2012). Volatility modeling of rainfall time series. *Theoretical and Applied Climatology*, 113(1-2), 247-258. <http://dx.doi.org/10.1007/s00704-012-0778-8>
- Yusof, F., Kane, I. L., & Yusop, Z. (2013). Structural break or long memory: an empirical survey on daily rainfall data sets across Malaysia, *Hydrol. Earth Syst. Sci.*, 17, 1311-1318. <http://dx.doi.org/10.5194/hess-17-1311-2013>
- Zhang, G. P. (2003). Time series forecasting using a hybrid of ARIMA and neural network model. *Neurocomputing.*, 50, 159-175. [http://dx.doi.org/10.1016/S0925-2312\(01\)00702-0](http://dx.doi.org/10.1016/S0925-2312(01)00702-0)

## Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).