A simulation method for flood risk variables

Yuzhi Cai[1], Ben Gouldby[2], Paul Dunning[2] and Peter Hawkes[2]

1 University of Plymouth
2 HR Wallingford

Abstract

Across the UK, storms often occur in sequences as a series of low pressure systems passing at intervals of one to four days. These systems can significantly influence the characteristics of subsequent flooding. Gaining an understanding of the sequencing of storm events and representing their characteristics within flood risk analysis is therefore of importance. In this paper we present a general method for simulating a time series with given marginal distributions and required autocorrelation structures. A simulation study was also carried out which shows that the generalized method works very well.

1. Introduction

Within flood risk analysis, serial dependence within a flood risk variable can arise frequently and can be important. Low pressure weather systems that track eastwards across the North Atlantic can result in high rainfall and/or high wind speeds and hence high wave conditions and surges. Often these low pressure weather systems occur in a series (typically numbering four or five), sometimes known as a family, whereby the centres of the successive cyclones maybe separated by a day's travel time. Damage to flood defences arising as a result of the initial depression is then exacerbated by subsequent storms, prior to any temporary repair work occurring. This scenario arose during the Towyn storm in North Wales in February 1990 (HR Wallingford, 1990) where defence breaches occurred and flooding arose on three successive high tides.

Whilst capturing the serial dependence is important, simulating a long time series of data, with a realistic extreme value distribution, is also of importance. For example, the long term rate of deterioration of flood defences can be particularly sensitive to extremes and their serial dependence. Consider scour at the toe of a sea defence; beach lowering as a result of extreme wave action can expose the structure footings (the magnitude of the extreme is of importance), subsequent storms occurring in quick succession (the degree of serial dependence), can cause rapid deterioration. Continuous simulation modelling (Calver *et al*, 2001) is increasingly being used within hydrological models to overcome complexities relating to the spatial and temporal dependencies.

In this paper we present a general method for simulating a time series with the required marginal distributions and required autocorrelation structures. This work formed part of Task 2 (*Estimation of extremes*) of the EU-funded FLOOD*site* programme. Related work on longer term temporal dependence (seasonality and trend) including illustrative case studies related to flood risk, is given in Hawkes (2007).

Cario and Nelson (1998) presented a numerical method for fitting and simulating autoregressive-to-anything processes, later Deler and Nelson (2001) and Biller and Nelson (2003) extended the work to multivariate cases, where the marginal distributions are in the Johnson family of distributions. One of the limitations of their approach is that the estimated base process, AutoRegressive process of order p, $AR(p)$ may not be stationary, and hence a very long simulated time series may not be available. One way to overcome such problems is to define a different base process. In this paper we generalize their method by defining an AutoRegressive Moving Average process of order p and q, $ARMA(p,q)$ as the base process. Such a generalization can avoid non-stationarity problems, for example, by using an $MA(q)$ as the base process, and can capture more complicated autocorrelation structures in observed time series. In this paper we focus on one dimensional cases; the methods are, however, extendable to higher dimensions.

The arrangement of the paper is as follows: the generalized method is presented in Section 2; we discuss the method for modelling the marginal distributions in Section 3; and Section 4 details a simulation study undertaken to test the method. Further comments and conclusions are provided in Section 5.

2. The simulation method

In this section, we provide the details about the generalized method.

Let Z_t be the one dimensional base process defined by

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + u_t + \theta_1 u_{t-1} + \ldots + \theta_q u_{t-q}$$

Where $\phi\, and\, \theta$ are the *AR* and *MA* coerfficents respectively and the u_t are independently and identically distributed (iid) normal random variables with mean zero and variance σ^2. The autocovariance function $\{r_h\}$ of the base process Z_t is given by

$$r_h = \phi_1 r_{h-1} + \phi_2 r_{h-2} + \ldots + \phi_p r_{h-p} + \sigma^2 \left(\sum_{i=h+1}^{\min\{p,q\}} \theta_i \phi_{i-h} + \sum_{i=h}^{q} \theta_i \theta_{i-h} \right), \tag{1}$$

where $r_{-h} = r_h, h = 0,1,\ldots,$ and $\theta_0 = 1$.

It is noticed that the marginal distribution of Z_t is the standard normal distribution, if we start the base process from a joint normal distribution with zero mean and covariance matrix

$$\Sigma_z = \begin{pmatrix} 1 & r_1 & \cdots & r_{p-1} \\ r_1 & 1 & \cdots & r_{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p-1} & r_{p-2} & \cdots & 1 \end{pmatrix}$$

and if we let

$$\sigma^2 = \frac{1 - \phi_1 r_1 - \ldots - \phi_p r_p}{\sum_{i=1}^{\min\{p,q\}} \theta_i \phi_i + \sum_{i=0}^{q} \theta_i^2}. \tag{2}$$

Furthermore, since r_0=1 we see that r_h is also the autocorrelation of the Z_t process at lag h.

It is also seen that Cario and Nelson's base process can be obtained by letting $q=0$ and hence is a special case of ours.

Let $F(y)$ be the marginal distribution function of the Y_t process, which is the time series we want to simulate. Define

$$Y_t = F^{-1}(\Phi(Z_t)),\ t = 1,2,\ldots,$$

where $\Phi(\cdot)$ is the standard normal distribution function. Then it is easy to show that Y_t has the required marginal distribution function F.

Therefore, we only need to guarantee that the simulated Y_t process has the required autocorrelation structure.

Let ρ_h be the autocorrelation of Y_t at lag h, that is,

$$\rho_h = Corr(Y_t, Y_{t+h}),\ h = 0,1,...$$

Then

$$\rho_h = Corr\left(F^{-1}(\Phi(Z_t)), F^{-1}(\Phi(Z_{t+h}))\right).$$

It is seen that the autocorrelation structure of the Y_t process is related to the autocorrelation structure of the Z_t process. For the simulated Y_t process to have the required autocorrelation structure ρ_h, we need to find the autocorrelation structure r_h of the base process Z_t.

It follows from the definition that

$$\rho_h = \frac{E(Y_t Y_{t+h}) - E(Y_t)^2}{Var(Y_t)},$$

where $E(Y_t)$ and $Var(Y_t)$ are the mean and variance of Y_t respectively and are determined by the marginal distribution function F, and

$$E(Y_t Y_{t+h}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F^{-1}(\Phi(Z_t))F^{-1}(\Phi(Z_{t+h}))\phi_{r_h}(z_t, z_{t+h})dz_t dz_{t+h},$$

where ϕ_{r_h} is the standard bivariate normal probability density function with correlation r_h and is given by

$$\phi_{r_h}(z_t, z_{t+h}) = \frac{1}{2\pi\sqrt{1-r_h^2}} e^{-\frac{z_t^2 - 2r_h z_t z_{t+h} + z_{t+h}^2}{2(1-r_h^2)}}$$

Therefore, the autocorrelation function $\{r_h\}$ of the Z_t process can be determined by solving the following equations

$$A + B\rho_h = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F^{-1}(\Phi(Z_t))F^{-1}(\Phi(Z_{t+h}))\phi_{r_h}(z_t, z_{t+h})dz_t dz_{t+h}, \qquad (3)$$

Where $A=(E(Y_t))^2$, $B=Var(Y_t)$ and $h=1,2\ldots$

If $p=0$, then the Z_t process is an $MA(q)$ process. In this case, $r_h=0$ for $h>q$. So only $r_1,\ldots,r_q$ need to be found. If $q=0$, the Z_t process is an $AR(p)$ process. Cario and Nelson (1998) considered this case and proposed a numerical search procedure for solving the equations (3).

In general cases, these are nonlinear equations, so a numerical method needs to be used. In this paper, we use the Newton-Raphson's method to find r_h with details given below.

Let

$$f(r_h) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F^{-1}(\Phi(Z_t))F^{-1}(\Phi(Z_{t+h}))\phi_{r_h}(z_t, z_{t+h})dz_t dz_{t+h} - B\rho_h - A.$$

It can be shown that if $f(r_h)$ converges, then $f'(r_h)$ also converges and is given by

$$f'(r_h) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F^{-1}(\Phi(Z_t))F^{-1}(\Phi(Z_{t+h}))\phi'_{r_h}(z_t, z_{t+h})dz_t dz_{t+h},$$

where

$$\phi'_{r_h}(z_t, z_{t+h}) = \phi_{r_h}(z_t, z_{t+h}) \cdot \xi_{r_h}(z_t, z_{t+h}),$$

and

$$\xi_{r_h}(z_t, z_{t+h}) = \frac{r_h}{1-r_h^{\,2}} + \frac{z_t z_{t+h} + r_h^{\,2} z_t z_{t+h} - r_h z_t^{\,2} - r_h z_{t+h}^{\,2}}{\left(1-r_h^{\,2}\right)^2}$$

Therefore, by the Newton-Raphson's method, an approximated value of r_h can be obtained by using the following recursive formula to a required accuracy:

$$r_h^{(n+1)} = r_h^{(n)} - \frac{f\left(r_h^{(n)}\right)}{f'\left(r_h^{(n)}\right)},\ n = 0,1,2,... \tag{4}$$

where $f(r_h^{(n)})$ and $f'(r_h^{(n)})$ are estimated by the following simulation method.

For $i=1....M$, where M is a given number, we simulate $\left(z_t^{\,i}, z_{t+k}^{\,i}\right) \sim N(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & r_k^{(n)} \\ r_k^{(n)} & 1 \end{pmatrix}.$$

Then

$$f\left(r_h^{(n)}\right) \approx \frac{1}{M} \sum_{i=1}^{M} F^{-1}\left(\Phi\left(Z_t^i\right)\right) F^{-1}\left(\Phi\left(Z_{t+h}^i\right)\right) - B\rho_h - A \tag{5}$$

and

$$f'\left(r_h^{(n)}\right) \approx \frac{1}{M} \sum_{i=1}^{M} F^{-1}\left(\Phi\left(Z_t^i\right)\right) F^{-1}\left(\Phi\left(Z_{t+h}^i\right)\right) \xi_{r_h^{(n)}}\left(Z_t^i, Z_{t+h}^i\right) \tag{6}$$

where $\Phi(z)$ is the distribution function of a standard normal random variable.

Once the autocorrelation structure r_h of the base process is available, the coefficients θ_i ($i=1,\ldots,p$) and ϕ_j ($j=1,\ldots q$) of the base process can be found by solving the system of equations (1).

If $p=0$, then Z_t follows an $MA(q)$ process. Hence, the system of equations (1) becomes

$$r_h = \begin{cases} \theta_h\theta_0 + \theta_{h+1}\theta_1 + \ldots + \theta_q\theta_{q-h} & h = 0,1,\ldots,q \\ 0 & h > q \end{cases}$$

and the parameters $\theta_1,\ldots,\theta_q$ can be obtained and $\sigma^2 = 1/\left(\theta_0^{\,2} + \ldots + \theta_q^{\,2}\right)$.

Alternatively, if $q=0$, then Z_t is an $AR(p)$ process and the system of equations (1) becomes

$$\begin{cases} r_1 = \phi_1 r_0 + \phi_2 r_1 + \ldots + \phi_p r_{p-1} \\ r_2 = \phi_1 r_1 + \phi_2 r_0 + \ldots + \phi_p r_{p-2} \\ \vdots \\ r_p = \phi_1 r_{p-1} + \phi_2 r_{p-2} + \ldots + \phi_p r_0 \end{cases}$$

Therefore, in this case, the parameters $\phi_1,\ldots,\phi_p$ can be obtained and $\sigma^2 = 1 - \phi_1 r_1 - \ldots - \phi_p r_p$.

Generally, since the Z_t process contains $p+q+1$ parameters, only the first $p+q+1$ equations in the system (1) are used to find those parameters. Again, in general cases no exact solutions are available, hence a numerical procedure is required. In this case the multivariate Newton-Raphson's method is used.

Finally, a time series z_t can be simulated from the base process, and then be transformed into the required Y_t process by $y_t = F^{-1}(\Phi(z_t))$ for t=1,2….

It is worth mentioning that $p+q$ is the maximum lag value that we are interested in for the Y_t process. In other words, $\rho_0, \rho_1, \ldots, \rho_{p+q}$ are the required correlation structure of the Y_t process, and any ρ_h for $h>p+q$ can be ignored.

3. Marginal distributions

The above method can be applied to any time series with any marginal distributions. As we will apply our method to flood variable data (e.g. sea condition data such as waves and water levels) we discuss a special marginal distribution function.

Sea condition variables such as wave height and water level are key variables within flood risk analysis. Because extreme events are rare, it is difficult to collect sufficient data for risk analysis based upon them. To study extreme events, Generalized Pareto Distributions (GPD) are often used and the marginal distribution of a sea condition variable can be taken as a mixture of a Weibull distribution and a GPD.

Let $Y = (Y_1,\ldots,Y_k)'$ be a vector of sea condition variables. The following theorems provide an explicit formula for the marginal distribution function and its inverse function.

Theorem 1 *The marginal distribution of Y_i is given by*

$$F_{Y_i}(y) = ((1-\beta_i)G_{1i}(y) + \beta_i)I_{[y>u_i]} + \frac{\beta_i}{G_{2i}(u_i)} G_{2i}(y) I_{[0<y\le u_i]},$$

where β_i satisfies $P(Y_i \le u_i) = \beta_i$,

$$G_{1i}(y) = 1 - \left(1 + \frac{\xi_i (y-u_i)}{s_i}\right)^{-1/\xi_i}, \quad 1 + \frac{\xi_i (y-u_i)}{s_i} \ge 0, \qquad y > u_i$$

is a GPD function,

$$G_{2i}(y) = 1 - e^{-\left(\frac{y}{b_i}\right)^{c_i}}, \qquad y>0$$

is a Weibull distribution function, and u_i is a proper threshold value.

Theorem 2 *The inverse function of* $F_{Y_i}(y)$ *is given by*

$$F_{Y_i}^{-1}(p)=\begin{cases} u_i+\left[\left(\dfrac{1-p}{1-\beta_i}\right)^{-\xi_i}-1\right]\dfrac{s_i}{\xi_i}, & p>\beta_i \\ \\ b_i\left[-\ln\left(1-\left(1-e^{-(u_i/b_i)^{c_i}}\right)p/\beta_i\right)\right]^{1/c_i}, & p\le\beta_i \end{cases}$$

where $0<p<1$.

4. Simulation study

In this section, we carry out a simulation study to investigate the performance of our method. Firstly, a time series sample of pseudo flood variable data was simulated. We henceforth refer to this as our "observed" data.

Now let *Y* be our random pseudo flood variable and, for the purposes of generating some "observed" data, assume it follows an *MA(2)* process:

$$y_t=\varepsilon_t-0.5\varepsilon_{t-1}+0.1\varepsilon_{t-2}$$

where $\varepsilon_t \sim N\left(0,1.5^2\right)$.

For this time series, the autocorrelation function is given by

$$\rho_h=\begin{cases} 1 & h=0 \\ \dfrac{\theta_1+\theta_1\theta_2}{1+\theta_1^{\ 2}+\theta_2^{\ 2}}=-0.437 & h=1 \\ \dfrac{\theta_2}{1+\theta_1^{\ 2}+\theta_2^{\ 2}}=0.079 & h=2 \\ 0 & h>2 \end{cases}$$

and the marginal distribution function of y_t is *N(0,2.835)*.

An "observed" series y_t', t=1,…,5000 was generated with the resulting sample parameters being $N(-0.010,\ 2.848)$ and $\rho_1'=-0.430$, $\rho_2'=0.068$.

The full y_t' series is shown in Figure 1 and a reduced series representing the first 250 timesteps is shown in Figure 2.

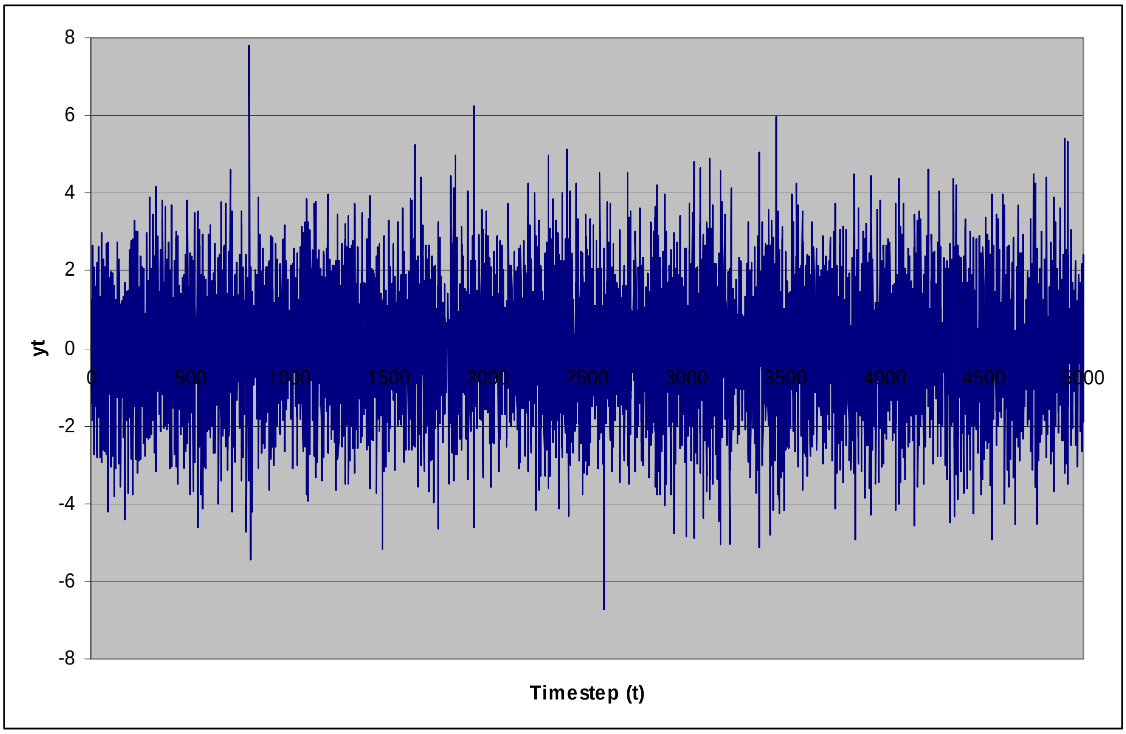

Figure 1 Realisations of the pseudo flood variable, generated from the *MA(2)* process, for 5,000 timesteps.

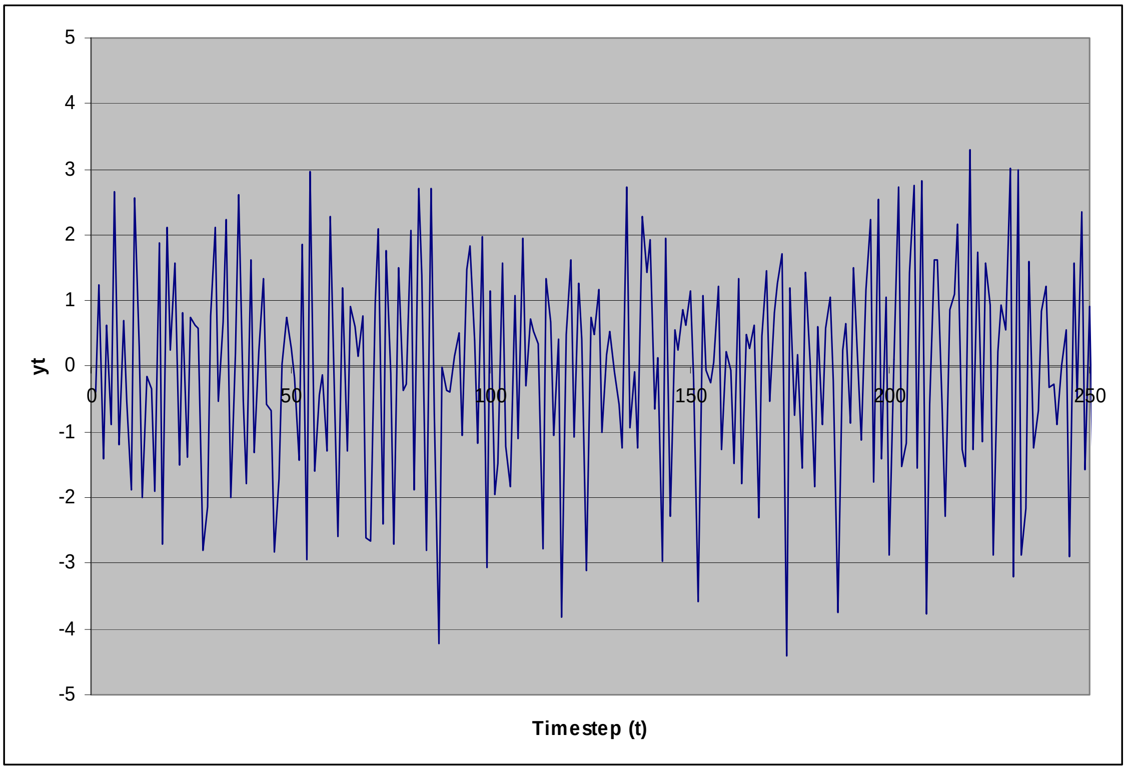

Figure 2 The first 250 realisations of the pseudo flood variable from the 5,000 timestep series, generated from the *MA(2)* process.

The objective of the simulation study was to use the method detailed in Section 2 to simulate a long series, $\hat{y}_t$, covering many more timesteps, whilst preserving the marginal distribution and autocorrelation parameters of our "observed" data.

We let our base process Z_t be

$$Z_t \sim ARMA(1,1) \quad \Rightarrow \quad Z_t = \phi_1 Z_{t-1} + \mu_t + \theta_1 \mu_{t-1} \qquad (7)$$

To obtain converged estimates of r_h, the autocorrelation function of the Z_t process at lag h, we generated a bivariate series, $\left(Z_i^t,\, Z_i^{t-h}\right),\ i = 1,\dots,M$, with our initial estimate of r_h. We then used this series, together with the parameters A and B estimated from our "observed" data to evaluate equations (5) and (6), in order to obtain a revised estimate of r_h from equation (4). This procedure was repeated until a converged estimate was obtained. In the context of its application here, the Newton Raphson algorithm was efficient, with convergence occurring in fewer than 10 iterations. The estimates of r_1 and r_2 were -0.523 and 0.072, respectively.

Having estimated the autocorrelation function of the Z_t series, the remaining *ARMA* paramaters $(\sigma^2, \theta_1, \phi_1)$ were found.

Equation 1 can be expanded to give:

$$\begin{cases} r_0 = \phi_1 r_1 + \sigma^2\left[\theta_0{}^2 + \theta_1(\theta_1 + \phi_1)\right] & h = 0 \\ r_1 = \phi_1 r_0 + \sigma^2(\theta_1\theta_0) & h = 1 \\ r_2 = \phi_1 r_1 & h = 2 \end{cases}$$

A quadratic in θ_1 was derived from these equations and solved to give the parameter estimates provided in Table 1.

Parameter	Value
ϕ_1	-0.14
θ_1	-0.60
σ^2	0.64

Table 1: Estimated parameters of the *Zt* (ARMA(1,1)) model

A large Z_t series was then simulated using these parameters (Eqn. 7), and transformed to obtain a $\hat{y}_t$ series of size 50,000. The parameters of the simulated $\hat{y}_t$ series compare well to those of our "observed" data given in Table 2.

Parameter	"Observed" data	Simulated series
Mean	-0.01	0.0
Variance	2.85	2.86
ρ_1	-0.43	-0.52
ρ_2	0.07	0.06

Table 2: Comparison of parameters from the "observed" data and the simulated series

The first 30,000 timesteps of the $\hat{y}_t$ series are shown in Figure 3 and the first 250 timesteps in Figure 4.

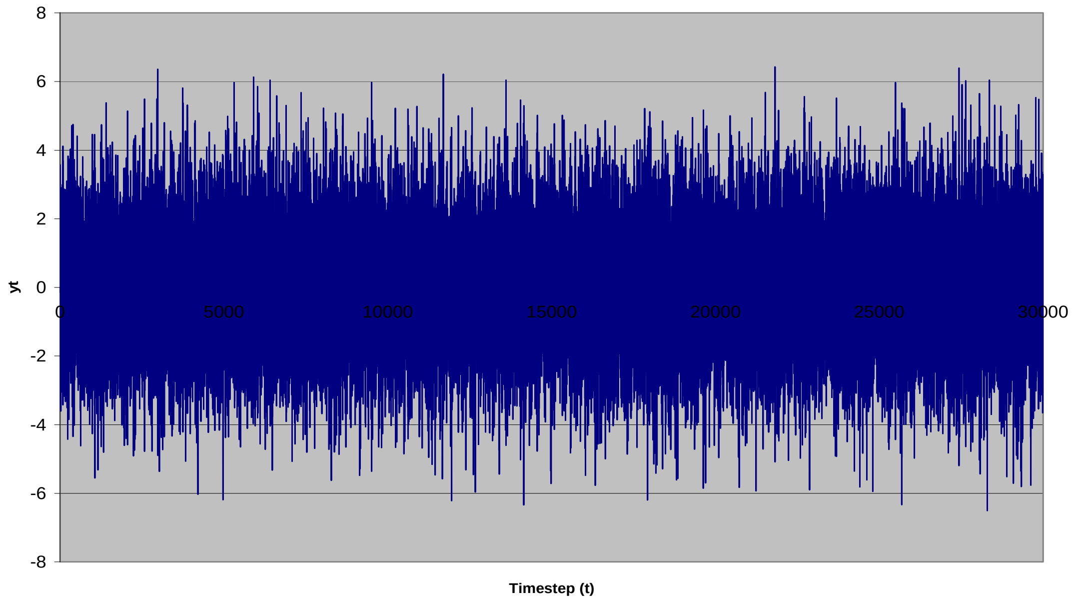

Figure 3 The first 30,000 realisations of the $\hat{y}_t$ variable from the 50,000 timestep series, generated from the *ARMA(1, 1)* base process.

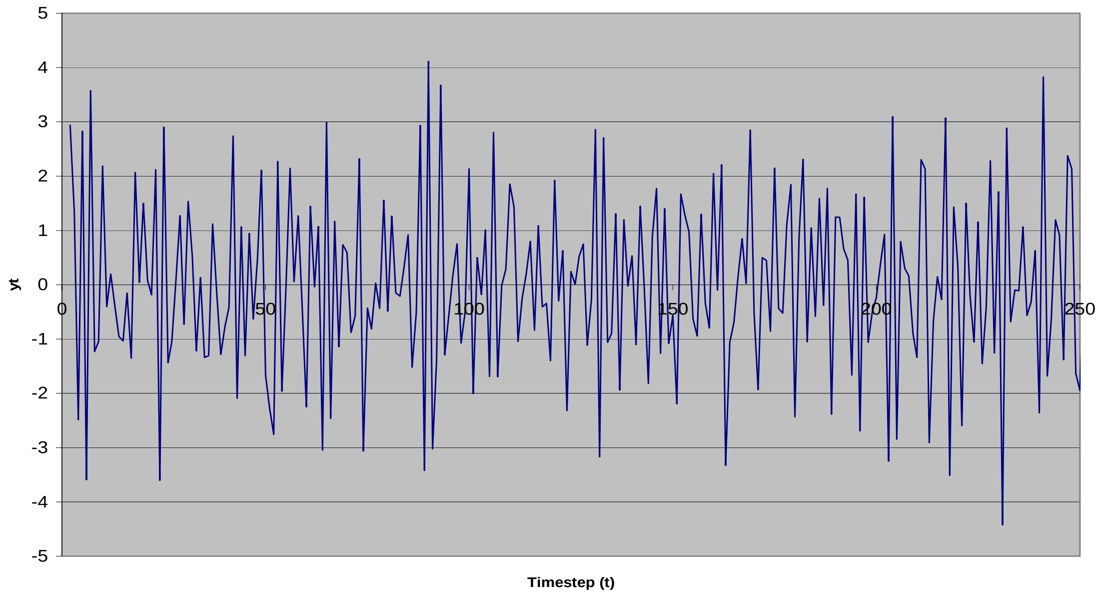

Figure 4 The first 250 realisations of the $\hat{y}_t$ variable from the 50,000 timestep series, generated from the *ARMA(1, 1)* base process.

Figure 5 shows a comparison of the total autocorrelation structure of the "observed" and final simulated data (limited to a lag of 20 time steps). It is evident both data sets have a similar structure, also lending confidence to the method.

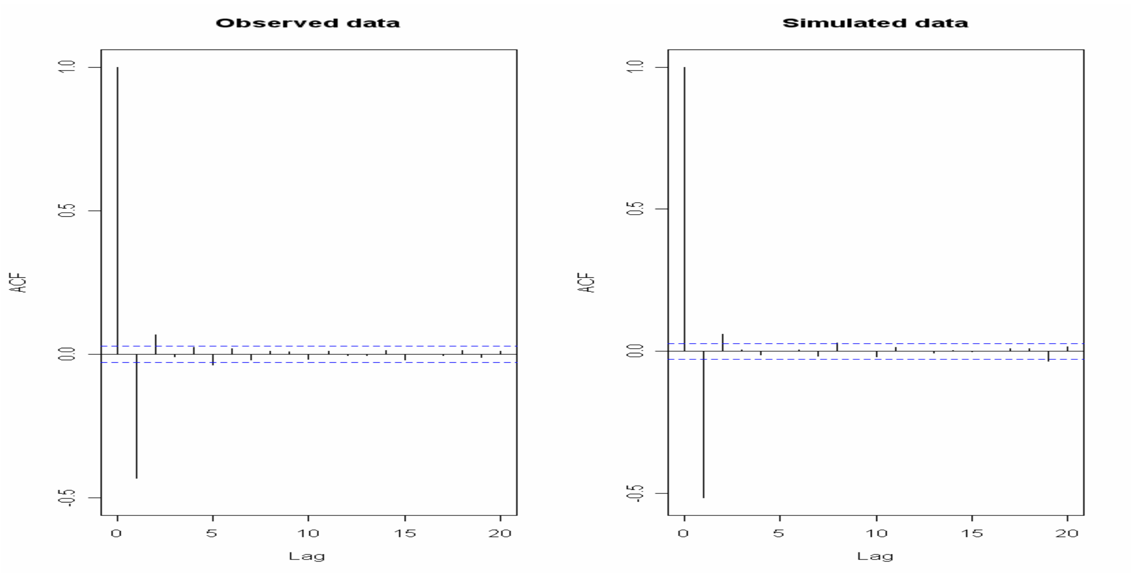

Figure 5. Autocorrelation function plots of the observed and simulated data up to 20 lags.

Further comparison of the marginal distributions of the "observed" and simulated data is shown in the Q-Q plot (Figure 6). The data sets show good agreement, with only minor deviations from the unit-gradient line.

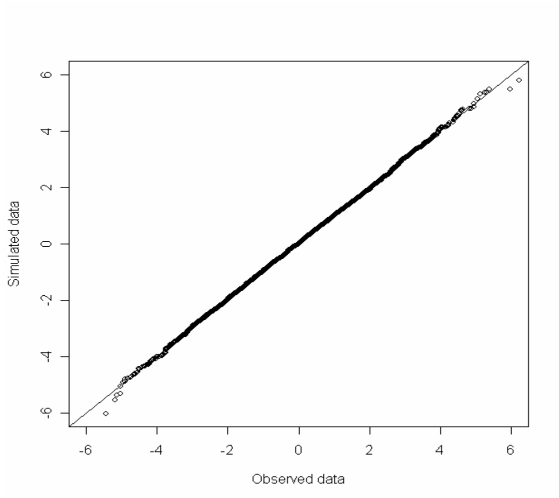

Figure 6. Q-Q plot of the observed and simulated data sets.

5. Comments and conclusions

We have presented a generalized method for simulation of a time series with the required correlation structure, up to $p+q$ lags, and desired marginal distribution functions. We have specifically shown a simulation study which demonstrates that the method works well.

The results obtained in this paper show that the estimated autocorrelation structure of the simulated time series is similar to that of the "observed" series, as are the marginal distributions. Our results also show that a base process *ARMA(p,q)* works very well.

Further work under this commission will involve the application of the method to real flood variables and extending the method to a multivariate case.

ACKNOWLEDGEMENT

The work described in this publication was supported by the European Community's Sixth Framework Programme through the grant to the budget of the Integrated Project FLOOD*site*, Contract GOCE-CT-2004-505420. This paper reflects the author's views and not those of the European Community. Neither the European Community nor any member of the FLOOD*site* Consortium is liable for any use of the information in this paper.

REFERENCES

BILLER, BAHAR and NELSON, BARRY L. (2003). Modeling and generating multivariate timeseries input processes using a vector autoregressive technique. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13, 211-237.

CALVER, A., LAMB, R. and MORRIS, S.E. (1999. Flood frequency estimation using continuous runoff modelling. *Proc. Institution of Civil Engineers, Water, Maritime & Energy*, 116, 225-234.

CARIO, MARNE C. and NELSON, BARRY L. (1998). Numerical methods for fitting and simulating autoregressive-to-anything processes. *Journal on Computing*, 10, 72-81.

DELER, BAHAR and NELSON, BARRY L. (2001). Modeling and generating multivariate time series with arbitrary marginals and autocorrelation structures. *Proceedings of the 2001 Winter Simulation Conference. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers*, 275-282

HAWKES, P.J. (2007). Joint probability analysis for estimation of extremes. Submitted to the Journal of Hydraulic Research for inclusion in a special issue based on FLOOD*site* Task 2.

HR WALLINGFORD (1990). Joint probability of waves and water levels on the North Wales coast. HR Wallingford Report EX 2331, copy available from HR Wallingford.