FLOOD FORECASTING USING TIME SERIES MODEL

Atan, I.B.[1], Aliman, A.A.[1], Jaafar, J.[1], Ashaari, Y.[1], Samsudin, M.B.[2], Mohamed, S.N.[5] and Baki, A.[2,3,*]
[1]Faculty of Civil Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[2]Envirab Services, P.O.Box 7866, GPO Shah Alam 40730, Selangor, Malaysia
[3]Faculty of Engineering, City University Malaysia, Seksyen 51A, 46100 Petaling Jaya, Selangor, Malaysia
[5]Faculty of Civil Engnineering, Universiti Teknologi MARA, Cawangan Johor Kampus Pasir Gudang, Johor, Malaysia
*Corresponding Author Email: aminbaki@msn.com

ABSTRACT

Flood forecasting is the study of rainfall patterns, catchment characteristics, and river hydrographs to predict the average frequency of flood occurrence in the future. Time-series model is used by representing a stochastic process, to estimate future values based on previously observed values. Flood may occur due to various causes such as inadequate design capacity of a storage to accommodate a certain amount of water, human activities like deforestation, lack of awareness in maintaining an early warning system for floods, and also heavy precipitation. In this study, discharge values of ten rivers were taken to be analysed by phases from collecting, manipulating, fitting the ARMA model, until the step of forecasting and analysing models. By using Mean Absolute Percent Error (MAPE), four of the models were analysed to be good models with percentage error below 20 percent, four more models are considered as tolerable models with percentage error between 20 and 40 percent, and another two models are rejected ones with value of percentage error exceeding 40 percent. By carrying out this study, it is recommendable for this method to be used in implementing other measures like early warning system and design of water storage. Flood forecasting is thus crucial in ensuring sustainability of human development.

Keywords:

ARMA Model, Flood Forecasting, Rainfall-Runoff, River Flow, Time Series Model.

INTRODUCTION

Over the years, floods are the most common and widespread of all natural hazards. Some floods develop over a period of days, while flash floods can result in raging waters in just a few minutes. Flash floods can be a deadly cargo of rocks, mud, and other debris and can occur without any visible sign of rainfall. Every region of every state is at risk from the hazards of flooding. A flooding situation is not a daily occurrence. However, flood forecasting operations must, of necessity, be a continuous activity. It is carried out from day to day even when the possibility of a flood is highly improbable. This mode of operation enables flood forecasters to pinpoint the beginning of a potential flood-generating situation.

Flood forecasting is the study of rainfall patterns, catchment characteristics, and river hydrographs to predict the future average frequency of occurrence of flood events. Flood predictions seek to estimate the probable discharge that, on average, will be exceeded any particular period. Flood may occurs due to various cause such as inadequate design capacity of a storage to accommodate a certain amount of water, human activities like deforestation, lack of awareness in maintaining an early warning system for floods, and also heavy precipitation.

In statistics, signal processing, econometrics and mathematical finance, a time series is a sequence of data points, measured typically at successive times spaced at uniform time intervals. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to forecast future events based on known past events to predict data points before they are measured.

Flood forecasting is an important component of flood warning, where the distinction between the two is that the outcome of flood forecasting is a set of forecast time-profiles of channel flows or river levels at various locations, while “flood warning” is the task of making use of these forecasts to make decisions about whether warnings of floods should be issued to the general public or whether previous warnings should be rescinded or retracted.

At the beginning of 1960's the formal development of stochastic modeling has started with the introduction and application of autoregressive models for annual and seasonal streamflows (Thomas and Fiering, 1962; Yevjevich, 1963). It is then leads to extensive research efforts toward improving those early concepts and models (Salas, 1993 Chatfield, 2000; Young, 2002; Damle, 2005; Young, 2006; Arselan, 2012; Vaghela and Vaghela, 2014; Meng *et al*., 2019); providing physical justification of some models, introducing alternative models and studying their impacts in water related to these various aspects is extensive and has been reviewed by hydrologists. Several stochastic models have been proposed in the past for modeling hydrologic time series. Although each model has its own merit and some of them can be successfully applied in operational hydrology, they do have limitations. They all have been criticized for one or more of the following reasons:

i. not being able to reproduce short-term dependence,

ii. not being able to reproduce long-term dependence,

iii. difficulty in estimating parameters,

iv. limitations for generating large samples of synthetic data,

v. lack of physical basis, and

vi. too many parameters.

Time series model is a mathematical model representing a stochastic process. It has a certain mathematical form or structure and a set of parameters. If X is normal with mean μ and variance σ_2 , the time series model can be conveniently written as

$$Xt = \mu + \sigma\, \varepsilon_t \,, t = 1,2, \ldots \quad (1)$$

Where εt is also normal with mean zero and variance one and $\varepsilon_1, \varepsilon_2, \ldots$ are independent. In Eq. (1) the model has the parameters μ and σ and since they are constants the model is stationary. The structure of the model is simple since the variable X_t is a function only of the independent variable ε_t and so X_t is also independent.

A time series model with dependence structure can be formed as

$$\varepsilon_t = \varphi\, \varepsilon_{t-1} + \xi_t \quad (2)$$

where ξt is an independent series with mean zero and variance $(1 - \varphi_2)$, εt is the dependent series an φ is the parameter of the model. In Eq. (2) ε_t is a dependent series because in addition to being a function of ξ_t, it is a function of the same variable ε at time t-1.

Since the parameters of the above models are constants, the models are stationary representing stationary time series or stationary stochastic processes. Non-stationary models would result if such parameters would vary with time.

Time series modeling is a process which can be simple or complex, depending on the characteristics of the available sample series, on the type for model to use and on the selected techniques of modeling. In general, time series modeling can be organized in the following stages (Box and Jenkins, 1970):

i. The selection of the type of model,

ii. The identification of the form of the model,

iii. The estimation of the model parameters, and

iv. The diagnostic check of the model.

As an example of the physical justification of autoregressive and moving average (ARMA) models for annual streamflow simulation, consider a watershed system as in Figure 3.0, where the variables are of annual values.

ISBN 1823-1039 @2020 FEIIC

Then the annual streamflow z_t is composed of groundwater contribution equal to c σ_{t-1} and surface runoff equal to dxt (Thomas and Fiering, 1962). That is

$$Z_t = c\ \sigma_{t-1} + d\ x_t \quad (3)$$

The continuity equation for the groundwater storage σ_t gives

$$\sigma_t = \sigma_{t-1} + a\ x_t - c\ \sigma_{t-1}$$

or

$$\sigma_t = (1\text{-}c)\ \sigma_{t-1} + a\ x_t \quad (4)$$

Combining Eqs. (3) and (4) Salas *et al.* (1980) showed that the model for the annual streamflow z_t can be written as Eq. (5) which has the form of an independent series.

$$z_t = (1\text{-}c)\ z_{t-1} + d\ x_t - [\ d\ (1\text{-}c) - ac\ ]\ x_{t-1} \quad (5)$$

Autoregressive moving average (ARMA) model is a forecasting model or process in which both autoregression analysis and moving average methods are applied to a well-behaved time series data. ARMA assumes that the time series is stationary-fluctuates more or less uniformly around a time-invariant mean. Non-stationary series need to be differenced one or more times to achieve stationarity. ARMA model forms a class of linear time series models which are widely applicable and parsimonious in parameterization. By allowing the order of an ARMA model to increase, one can approximate any linear time series model with desirable accuracy.

The notation ARMA(p, q) refers to the model with p autoregressive terms and q moving-average terms. This model contains the AR(p) and MA(q) models,

$$X_t = c + \varepsilon_t + \Sigma q_i{=}1\ \varphi i\ x_{t-i} + \Sigma q_i{=}1\ \theta_i\ \varepsilon_{t-i} \quad (6)$$

In statistics and signal processing, an autoregressive (AR) model is a type of random process which is often used to model and predict various types of natural phenomena. The autoregressive model is one of a group of linear prediction formulas that attempt to predict an output of a system based on the previous outputs. The notation AR(p) indicates an autoregressive model of order p. The AR(p) model is defined as

$$X_t = c + \Sigma q_i{=}1\ \varphi_i\ x_{t-i} + \varepsilon_t \quad (7)$$

where φ1 … φp are the parameters of the model, c is a constant (often omitted for simplicity) and ε_t is white noise. The constant term is omitted by many authors for simplicity.

In time series analysis, the moving-average (MA) model is a common approach for modeling univariate time series models. The notation MA(q) refers to the moving average model of order q:

$$X_t = \mu + \varepsilon_t + \theta_1\ \varepsilon_{t-1} + \ldots + \theta_q\ \varepsilon_{t-q} \quad (8)$$

where μ is the mean of the series, the θ_1, ..., θ_q are the parameters of the model and the ε_t, ε_{t-1},... are white noise error terms. The value of q is called the order of the MA model. That is, a moving-average model is conceptually a linear regression of the current value of the series against previous (unobserved) white noise error terms or random shocks. The random shocks at each point are assumed to come from the same distribution, typically a normal distribution, with location at zero and constant scale. The distinction in this model is that these random shocks are propagated to future values of the time series. Fitting the MA estimates is more complicated than with autoregressive models (AR models) because the error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares. MA models also have a less obvious interpretation than AR models.

Sometimes the autocorrelation function (ACF) and partial autocorrelation function (PACF) will suggest that a MA model would be a better model choice and sometimes both AR and MA terms should be used in the same model. In time series analysis, the partial autocorrelation function (PACF) or PARtial autoCORrelation (PARCOR) plays an important role in data analyses aimed at identifying the extent of the lag in an autoregressive

model. Given a time series z_t, the partial autocorrelation of lag k, denoted α(k), is the autocorrelation between z_t and $z_t + k$ with the linear dependence of $z_t + 1$ through to $z_t + k - 1$ removed; equivalently, it is the autocorrelation between z_t and $z_t - k$ that is not accounted for by lags 1 to $k - 1$, inclusive.

$$\alpha(1)=\mathrm{Cor}(z\_(t),z\_(t+1)) \quad (9)$$

$$\alpha(k)=\mathrm{Cor}[[z]\_(t+k)- P\_(t,k)(z\_(t+k)), z\_(t)- P\_(t,k)(z\_t)], \text{ for } k \geq 2, \quad (10)$$

where Pt,k(x) denotes the projection of x onto the space spanned by z_t+1 , … , z_t+k+1.

After a time series has been stationarized by differencing, the next step in fitting an ARIMA model is to determine whether AR or MA terms are needed to correct any autocorrelation that remains in the differenced series. Software like Minitab can be used to try some different combinations of terms and see what works best. By looking at the autocorrelation function (ACF) and partial autocorrelation (PACF) plots of the differenced series, you can identify the numbers of AR and/or MA terms that are needed. In ACF plot, it is merely a bar chart of the coefficients of correlation between a time series and lags of itself. The PACF plot is a plot of the partial correlation coefficients between the series and lags of itself.

In general, the "partial" correlation between two variables is the amount of correlation between them which is not explained by their mutual correlations with a specified set of other variables. For example, if we are regressing a variable Y on other variables X_1, X_2, and X_3, the partial correlation between Y and X_3 is the amount of correlation between Y and X_3 that is not explained by their common correlations with X_1 and X_2. This partial correlation can be computed as the square root of the reduction in variance that is achieved by adding X_3 to the regression of Y on X_1 and X_2. (Luce, 1992)

There are several steps need to be considered in processing this study. Figure 1 is the flowchart of the study showing from the starting point, onto the process, and the finishing point.

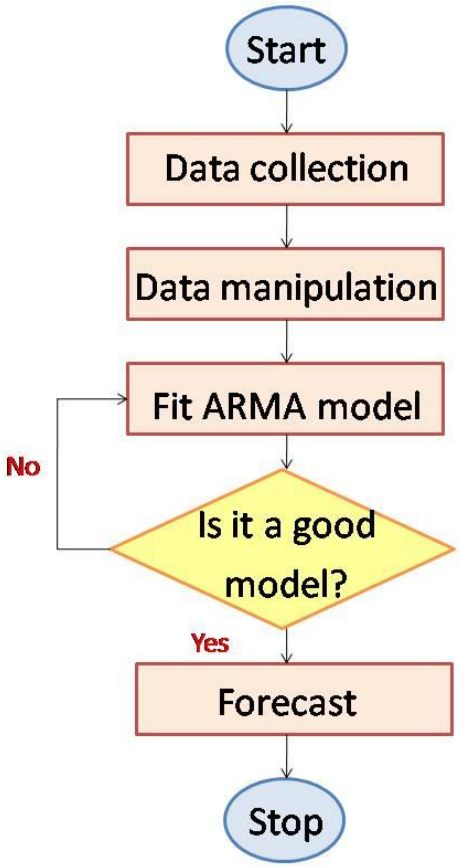


Figure 1: Flowchart of the study

In order to determine if the values are suitable and could fir the ARMA model, several stages were introduced to keep a smooth flow. A systematic approach to hydrologic time series modeling may be composed of six main phases (Salas *et al.*, 1980):

i. Identification of model composition,

ii. Selection of model type,

iii. Identification of model form,

iv. Estimation of model parameters,

v. Testing goodness of fit of the model, and

vi. Evaluation of uncertainties.

It is important to remember the difference between a model and a process. In practice which ARMA process has generated a given realization could not be observed, so a trial-and-error procedure was utilised. In the 'trial' part, the estimated acf and pacf calculated from the realization. Some ARMA generation mechanisms were generated to fit the available data adequately. A model is different from a process where a process is the true but unknown mechanism that has generated a realization, while a model is only an imitation or representation of the process. A model was selected based on the adequacy with respect to the available data.

Following are the characteristics to be considered in deciding if it is a good model. (Pankratz, 1983)

i. It is parsimonious (uses the smallest number of coefficients needed to explain the available data)

A parsimonious model fits the available data adequately without using any unnecessary coefficient. For example, if an AR(1) model and an AR(2) model are essentially the same in all other respects, we would select the AR(1) model because it has one less coefficient to estimate.

Parsimonious models generally produce better forecasts. It is a model which only approximates the true process as long as the model explains the behavior of the available realization in a parsimonious and statically adequate manner.

ii. It is stationary (has AR coefficients which satisfy some mathematical inequalities)

ARMA method applies only to a realization that is stationary, meaning it has a constant mean, variance and acf. Stationarity for a model can be check by seeing if the estimated AR coefficients satisfy some mathematical inequalities.

iii. It is invertible (has MA coefficients which satisfy some mathematical inequalities)

Invertibility is algebraically similar to stationarity. The invertibility can be check by seeing if the estimated MA coefficients satisfy some mathematical inequalities.

iv. It has estimated coefficients (Ø's and θ's) of high quality:

a) Absolute t-values about 2.0 or larger,

As we want to avoid a forecasting model which represents only a chance relationship, so each Ø or θ coefficient have an absolute t-statistic of about 2.0 or larger. This means each estimated Ø or θ coefficient should be about two or more standard errors away from zero.

b) Ø's and θ's not too highly correlated.

Estimated Ø and θ coefficients should not be too highly correlated with each other. If they are tend to be unstable even if they are statistically significant.

ISBN 1823-1039 @2020 FEIIC

v. It has uncorrelated residuals

A good model has statistically independent residuals. The assumption was that random shocks (a_t) are independent in a process. The random shocks were not observed, estimates of them can be made at the estimation stage. The α_t are called residuals of a model. The shocks were tested for independence by constructing an acf using the residuals as input data. If the residuals are statistically independent, this is important evidence that the model cannot be improved further by adding more AR or MA terms.

vi. It fits the available data (the past) well enough at the estimation stage:

a) Root-mean-squared error (RMSE) is acceptable,

b) Mean absolute percent error (MAPE) is acceptable.

No model can fit the data perfectly because there is a random shock element present in the data. These two measures of closeness of fit; root-mean-squared error (RMSE) and mean absolute percent error (MAPE) need to be assured that they are acceptable.

vii. It forecasts the future satisfactorily.

Above all, a good model has sufficiently small forecast errors. Although a good forecasting model will usually fit the past well, it is even more important that it forecast the future satisfactorily.

Mean Absolute Percent Error (MAPE) is commonly used in quantitative forecasting methods because it produces a measure of relative overall fit. MAPE is used in this study, to determine the fit of a model. It usually expresses accuracy as a percentage, and is defined by the formula:

$$\text{MAPE} = \Sigma \left|((y\_t-\acute{y}\_t))/y\_t\right|/n \times 100 \quad (y\_t \neq 0) \quad (11)$$

where y_t is the actual value and $\acute{y}_t$ is the forecast value. (Cryer, 1986)

The difference between y_t and $\acute{y}_t$ is divided by the actual value y_t again. The absolute value in this calculation is summed for every fitted or forecasted point in time and divided again by the number of fitted points n. multiplying by 100 makes it a percentage error. A good forecasting model would usually fit the calculation by having the percentage of not more than 20.

RESULT AND DISCUSSION

All ten rivers will be evaluated, from first, picking the smallest mean square (MS) value of all zero to five model possibilities, second, comparing the chi-square and degree of freedom value, and next, gaining the percentage error from MAPE later on. The models will then be categorized under three different classes which are the good models, tolerable models and rejected models.

There are several steps in evaluating the data of the rivers. For Ara River:

Differentiating mean square (MS) value of the models

The model with the smallest MS value will be taken into consideration. In order to identify which model to be picked, Minitab is used to generate the data, and compare the model possibilities starting from (0,1), (0,2), (0,3) … (5,5).

Comparing the value of chi-square, x_2 with degree of freedom (DF)

ISBN 1823-1039 @2020 FEIIC

Based on table of Percentage Point of the Chi-Square, x_2 Distribution, the value of chi-squared must always be assured to be less than DF.

The percentage error is to be obtained by using MAPE's formula. (Cryer, 1986)

MAPE = $\sum_{t=1}^{n} |((y_t - \hat{y}_t))/y_t|/n \times 100$ ($y_t \neq 0$)

A model is a good fit when the percentage error is not more than 20%.

A model is considered a good fit when it satisfies the mathematical inequalities. A good model also has statistically independent residuals. In this case, a model is perfectly fit when it has a low percentage error when generating it from the MAPE and the value should not be more than 20%. This means the model is able to forecast the future satisfactorily.

After analyzing the data from Ara River has measures of accuracy of 18.30% (as shown in Table 1).

CONCLUSION

It can be concluded that the combination of AR and MA model could form a perfect whole ARMA model with the help of a well-behaved time series data. There are several processed that need to be considered to determine if a model is perfectly or poorly fit. Models with measures of accuracy less than 20% makes good models, while tolerable models are those with measures of accuracy more than 20% but less than 40%. When it exceeds 40%, it is then categorized as rejected models. By generating a time-series model:

i. This method could effectively forecast the flood occurrence in the near future. Since this study forecast the annual river flow, it is applicable to predict the river flow up to the next 5 years. Thus it is recommended to bring improvements to the flood prediction.

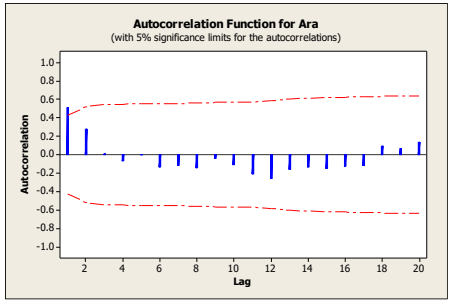
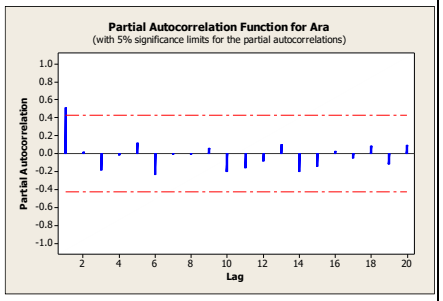
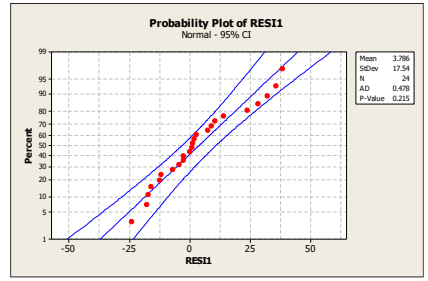
ii. By fulfilling the main objectives of this study, this method of time-series model could also be very useful in water resource planning, such as for water storage system e.g dams, channels, retention tanks and many others. The ability to accumulate a large amount of water could prevent it from overflowing that will then lead to the occurrence of flood.

iii. This method is recommended to achieve an adequate early warning system. When the flow or discharge of a certain river is already known, it is easier to prepare a system that could alert the community and provide more time to evacuate the area. In achieving all these conditions, it is hoped that the risk could be sufficiently controlled in the future.

By carrying out this study, it is recommended that this method be used in implementing other measures like early warning system and design of water storage. Flood forecasting is thus crucial in ensuring sustainability of human development.

ISBN 1823-1039 @2020 FEIIC

Table 1: Ara River

RIVER (MODEL)	AUTOCORRELATION FUNCTION (ACF)	PARTIAL AUTO CORRELATION FUNCTION (PACF)	MODEL ADEQUENCY	RESIDUAL
ARA (3,4)			Final Estimates of Parameters Type Coef SE Coef T P AR 1 0.3991 0.2896 1.38 0.187 AR 2 0.3965 0.2553 1.55 0.140 AR 3 -0.1694 0.2874 -0.59 0.564 MA 1 -0.3795 0.2983 -1.27 0.222 MA 2 0.0671 0.3377 0.20 0.845 MA 3 0.1840 0.3632 0.51 0.619 MA 4 0.9741 0.3008 3.24 0.005 Constant 21.423 1.708 12.55 0.000 Mean 57.317 4.569 Number of observations: 24 Residuals: SS = 7417.86 (backforecasts excluded) MS = 463.62 DF = 16 Modified Box-Pierce (Ljung-Box) Chi-Square statistic Lag 12 24 36 48 Chi-Square 9.3 * * * DF 4 * * * P-Value 0.055 * * *	

REFERENCES

[1] Arselan, C.A., (2012), Stream flow Simulation and Synthetic Flow Calculation by Modified Thomas Fiering Model, “Al-Rafidain Engineering”, Vol.20, No.4, pp.118-127.
[2] Box, G. and Jenkins, G. (1970) "Time series analysis: Forecasting and control". San Francisco: Holden-Day.
[3] Chatfield, C., (2000)"Time-series forecasting." Research Journal, Department of Mathematical Sciences, University of Bath, UK.
[4] Vaghela, C.R. and Vaghela, A.R. (2014) Synthetic Flow Generation, “Int. Journal of Engineering Research and Applications” www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 7 (Version 6), July 2014, pp.66-71.
[5] Cryer, J.D., (1986) “Time Series Analysis”, Duxbury Press.
[6] Damle, C., (2005)."Flood forecasting using time series data mining." Theses and Dissertations, Paper 2844.
[7] Luce, M.F., (1992), "Buying More Than We Can Use: Factors Influencing Forecasts Of Consumption Quantity", in Advances in Consumer Research Volume 19: 584-588.
[8] Meng, E., Huang, S., Huang, Q., Fang, W., Wu, L. and Wang, L. (2019) A robust method for non-stationary streamflow prediction based on improved EMD-SVM model, “Journal of Hydrology”, Vol.568, pp.462-478.
[9] Pankratz, A., (1983) "Forecasting With Univariate Box-Jenkins Models; Concepts and Cases", John Wiley & Sons, NY.
[10] Salas, J.D., (1993) "Analysis and Modelling of Hydrologic Time Series", Chapter 19:72 in The McGraw Hill Handbook of Hydrology, D.R. Maidment, Editor.
[11] Salas, J. D., Delleur, J. W., Yevjevich, V. and Lane, W. L. (1980) "Applied Modeling of Hydrologic Time Series", Water Resources Publications, Littleton, Colorado, 484.
[12] Thomas, H.A. & Fiering, M.B. (1962) Mathematical synthesis of stream flow sequences for the analysis of river basins by simulation. In Maas, A., Hufschmidt, M.M. and Dorfman, R. (eds.) “Design of Water Resources Systems”, Harvard University Press.
[13] Yevjevich, V.M., (1963) “Fluctuations of wet and dry years, I, Research data assembly and mathematical models”, Hydrology Papers (Colorado State University) no. I.
[14] Young, P.C., (2002). "Advances in real-time flood forecasting." Royal Society A (May 2002) 360: 1433-1450.
[15] Young, P.C., (2006). "Updating algorithms in flood forecasting." FRMRC Research Report UR5, 55-56.