한국방재학회논문집
제10권 1호 2010년 2월
pp. 81 ~ 88

하천방재

Synthetic Streamflow Generation Using Autoregressive Modeling in the Upper Nakdong River Basin

Rubio, Christabel Jane P.* · Oh, KukRyul** · Ryu, Jae H.*** · Jeong, SangMan****

Abstract

The analysis and synthesis of various types of hydrologic variables such as precipitation, surface runoff, and discharge are usually required in planning and management of water resources. These hydrologic variables are mostly represented using stochastic models. One of which is the autoregressive model, that gives promising results in time series modeling. This study is an application of this model, which aimed to determine the AR model that best represents the historical monthly streamflow of the two gauging stations, namely Andong Dam and Imha Dam, both located in the upper Nakdong River Basin. AR(3) model was found to be the best model for both gauging stations. Parameters of the determined order of AR model (ϕ_1, ϕ_2 and ϕ_3) were also estimated. Using several diagnostic tests, the efficiency of the determined AR(3) model was tested. These tests indicated the accuracy of the determined AR(3) model.

***Key words* : Autoregressive Model, Time Series, Monthly Streamflow**

요 지

수자원의 관리 및 계획시 강우, 유출, 유량과 같이 다양한 종류에 의한 수문사상의 합성 및 분석이 요구된다. 다양한 수문사상들은 대부분 추계학적모형에 의한 해석이 필요하며, 이중 적절한 시계열모의결과를 나타낼 수 있는 자기회귀모형 적용을 시도하였다. 본 연구에서는 낙동강 상류에 위치한 안동댐과 임하댐 두 관측소의 월유출량 자료를 이용하여 최적의 자기회귀모형을 검토하였으며, 분석결과 AR(3) 모형의 매개변수(ϕ_1, ϕ_2, and ϕ_3)가 가장 적합한 것으로 나타났으며, 다양한 분석 및 평가결과 AR(3)모형이 효과적이고 정확한 것으로 나타났다.

핵심용어 : 자기회귀모형, 시계열, 월유출량

1. Introduction

Time series analysis has become a major tool in hydrology. It is used for building mathematical model to generate synthetic hydrological records, to forecast hydrological events, to detect trends and shifts in hydrologic records, and to fill in missing data and extend records (Maidment, 1993).

Modeling can contribute to understanding the physical systems by revealing something about the physical process that builds persistence into the series. For example, a simple physical water-balance model consisting of terms for precipitation input, evaporation, infiltration, and groundwater storage can be shown to yield a streamflow series that follows a particular form of an autoregressive (AR) model. AR models can also be used as a baseline to evaluate possible importance of other variables in the system.

The modeling of streamflow time process has essentially followed two approaches: the deterministic or physical simulation of the hydrological system, and the statistical or stochastic simulation of the system. In the first approach, the hydrologic system is described and represented by theoretical and/or empirical physical relationships. On the other hand, in the stochastic approach, a type of model is assumed aimed to represent the most relevant statistical characteristics of the historical series (Salas, et al., 1980). Within this approach, the most widely used models have been the AR models (Thomas and Fiering, 1962; Yevjevich, 1963).

Autoregressive (AR) models have been extensively used in hydrology and water resources since the early 1960's. Salas et al. (1980) cited several reasons why hydrologists have been attracted in using autoregressive models: (1) the autoregressive form has an intuitive type of time depen-

*Graduate Studentrt · Ongju National University, Department of Civil and Environmental Engineering (E-mail: cjrubio@kongju.ac.kr)
**Ph. D. Student · Kongju National University, Department of Civil and Environmental Engineering
***Temporary Faculty · University of Idaho, Biological and Agricultural Engineering Department
****Member · Corresponding Author · Professor · Kongju National University, Department of Civil and Environmental Engineering

dence (the value of a variable at the present time depends on the values at previous times), and (2) they are the simplest models to use.

For this research, a synthetic streamflow is generated using an autoregressive (AR) model, applied to two time series, Andong Dam and Imha Dam transformed streamflow series.

2. Study Area

2.1 Study Area Description

Nakdong River system has a catchment area of 23,393.70 km^2 and a river length of 509.70 km. It is the largest river system in terms of river length and second in catchment area. There are a total of 7 large multipurpose dams under operation located within Nakdong River Basin and 22 streamflow gauging stations.

The AR models determined in this study were applied to two gauging stations, Andong Dam and Imha Dam. Andong Dam is located in the upper Nakdong River Basin with a total drainage area of 1,584.0 km^2, which is 20% of the total Nakdong River. It is the main supply of drinking water for metropolitan cities like Daegu and Busan, which are located in the lower Nakdong River and Kimhae, one of the large scale agricultural areas in South Korea. Imha Dam was built in 1992, located 18km downstream from the confluence of the main Nakdong River and tributary Banbyeon. Its total drainage area is 1,461 km^2 which supplied water, generates hydropower, and controls flood.

2.2 Data Collection

Development of the AR model was performed using the monthly long-term outflow series of two gauging stations in Nakdong River Basin namely Andong Dam and Imha Dam. The Water Management Information System (WAMIS) of the Ministry of Construction and Transportation (MOCT) provided the monthly streamflow for the Andong and Imha parallel reservoir group recorded from 1966 until 2005 (MOCT, 2008). Fig. 1 shows the location of the two gauging stations.

3. Data Preprocessing

3.1 Initial Data Analysis

The main purpose of this part is to check the normality of the original monthly series necessary to make appropriate transformations to normality if necessary.

Most probability theory and statistical techniques applied to hydrology in general and to hydrologic time series analysis in particular, are developed assuming the variables are

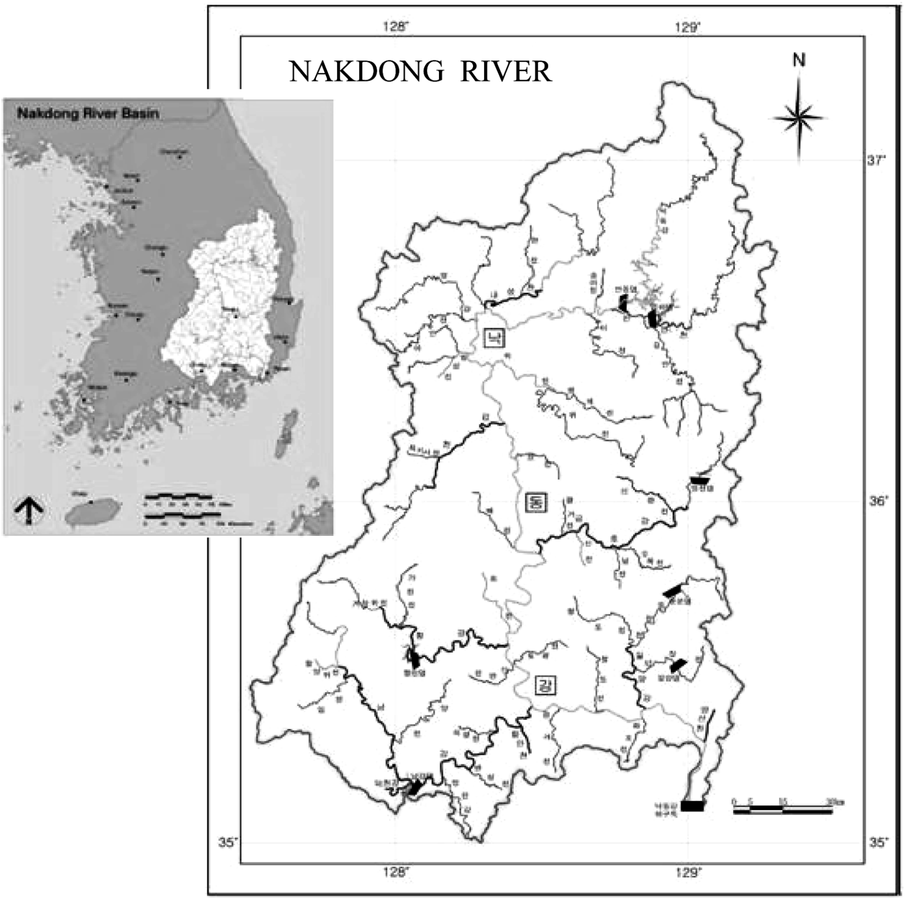


Fig. 1. Nakdong River Basin

normally distributed. Because most frequency curves of hydrologic variables are asymmetrically distributed, or are bounded by zero, it is often necessary to transform those variables to normal before carrying out the statistical analysis of interest.

To check the normality of the original monthly series, the skewness test of normality (Salas et al., 1980) was performed. The test indicated that both the skewness for the two gauging stations are outside the limit = ±0.759 (where N = 40; number of observation years). Therefore, both series needs to be transformed for a more efficient modeling. For this study, the logarithmic transformation was performed to reduce the skewness.

3.2 Data Transformation

The normalization process consists of transforming the original streamflow records into normal series since the original streamflow data are positively valued variables. This procedure will remove the skewness from the original records. The logarithmic transformation was used, given by

$$X_{vr} = \log(Q_{vr}) \qquad (1)$$

where Q_{vr} is the monthly inflow (m^3/sec) for month (v = 1,···, 12) and year (t = 1,···, N); N is the number of years of record of the series.

Next, to account for periodicity and/or tendencies, the resulting transformed series of Eq 1 will be standardized to improve the overall operation of the AR model. That is, each normalized series is converted into a series with mean zero and a standard deviation equal to one through the equation

$$Y_{v\tau} = \frac{X_{vr} - \bar{X}_\tau}{s_\tau} \qquad (2)$$

where $X_{v\tau}$ = normalized inflow for year v and month τ
$\bar{X}_\tau$ = sample mean for month τ
s_τ = standard deviation for month τ
where $Y_{v\tau}$ = standardized value for year v and month τ

Table 1 through Table 3 gives the summary of basic statistics, mean, standard deviation and skewness, of the original and the transformed monthly series of Andong Dam and Imha Dam. After the normalization, skewness was significantly reduced, except for some months (January, February, November and December for Andong Dam; August for Imha Dam). Additional tests were performed in order to further check normalilty.

Additional test for statistical confidence was performed to check the transformed series. The Kolmogorov-Smirnov test was performed which quantifies the distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, in this case, the normal distribution. Under the null hypothesis that the data are normally distributed, we can fail to reject the null hypothesis at the "usual" alpha = 0.05 (5% significance level) for Andong Dam with a p-value of 0.954. However, for Imha Dam the p-value is 0.0498, value less than 0.05 but could be considered as almost normal.

To show more evidence against normality, the quantile-quantile (Q-Q) plots for both the series are given in Fig. 2. A Q-Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the Q-Q plot

Table 1. Monthly Mean Summary

(unit: m^3/sec)

Month	ANDONG DAM			IMHA DAM		
	Raw ($Q_{v\tau}$)	Normalized ($X_{v\tau}$)	Standardized ($Y_{v\tau}$)	Raw ($Q_{v\tau}$)	Normalized ($X_{v\tau}$)	Standardized ($Y_{v\tau}$)
January	5.45	0.65	-0.002	5.15	0.56	0.001
February	7.51	0.68	-0.001	8.99	0.64	0.001
March	17.55	1.09	0.001	17.89	1.04	0.000
April	33.32	1.41	-0.002	29.24	1.31	0.001
May	26.45	1.27	0.000	25.97	1.26	-0.001
June	32.41	1.36	0.000	33.77	1.37	0.000
July	86.24	1.86	0.000	81.97	1.79	0.001
August	84.12	1.78	0.001	93.56	1.83	-0.001
September	58.83	1.61	-0.001	64.96	1.64	-0.001
October	16.54	1.13	-0.001	17.00	1.17	0.002
November	11.51	0.99	0.001	10.42	0.97	-0.001
December	7.06	0.79	0.000	6.40	0.73	0.000

Table 2. Monthly Standard Deviation Summary

(unit: m^3/sec)

Month	ANDONG DAM			IMHA DAM		
	Raw ($Q_{v\tau}$)	Normalized ($X_{v\tau}$)	Standardized ($Y_{v\tau}$)	Raw ($Q_{v\tau}$)	Normalized ($X_{v\tau}$)	Standardized ($Y_{v\tau}$)
January	5.05	0.24	1.001	6.16	0.34	1.000
February	9.82	0.37	0.999	14.27	0.50	1.000
March	16.48	0.39	0.999	18.30	0.46	0.999
April	24.94	0.33	0.999	25.94	0.39	1.000
May	22.01	0.36	1.000	21.91	0.39	0.999
June	28.68	0.37	1.000	31.07	0.40	0.999
July	47.12	0.29	1.001	56.92	0.37	1.000
August	67.76	0.39	0.999	71.27	0.41	1.000
September	49.91	0.40	1.000	56.09	0.42	1.001
October	13.05	0.27	0.999	9.04	0.23	0.999
November	8.88	0.24	1.000	5.21	0.22	0.999
December	4.95	0.21	1.001	5.01	0.25	1.001

Table 3. Monthly Skewness Summary

(unit: m^3/sec)

Month	ANDONG DAM			IMHA DAM		
	Raw ($Q_{v\tau}$)	Normalized ($X_{v\tau}$)	Standardized ($Y_{v\tau}$)	Raw ($Q_{v\tau}$)	Normalized ($X_{v\tau}$)	Standardized ($Y_{v\tau}$)
January	3.52	1.50	1.50	3.18	0.64	0.64
February	3.14	0.89	0.89	3.16	0.49	0.49
March	2.44	-0.40	-0.40	1.90	-0.33	-0.33
April	1.13	0.06	0.06	1.97	-0.35	-0.35
May	1.21	0.06	0.06	1.31	-0.35	-0.35
June	1.40	0.21	0.21	1.82	-0.22	-0.22
July	0.31	-0.69	-0.69	0.72	-0.61	-0.61
August	1.50	-0.42	-0.42	1.64	-0.92	-0.92
September	1.26	-0.14	-0.14	1.16	-0.48	-0.48
October	2.21	0.70	0.70	1.26	-0.19	-0.19
November	2.70	0.92	0.92	1.33	-0.27	-0.27
December	3.57	0.81	0.81	3.55	0.15	0.15

will approximately lie on the line $y = x$ (standard normal distribution line). In Fig. 2, the red line is the line passing through the first and third quantiles with its corresponding 95% confidence interval limit given by the red dashed line. A largely straight-line pattern, following the standard normal distribution line is shown in the Q-Q plot except for the extreme values, therefore supporting the assumption of a normally distribution.

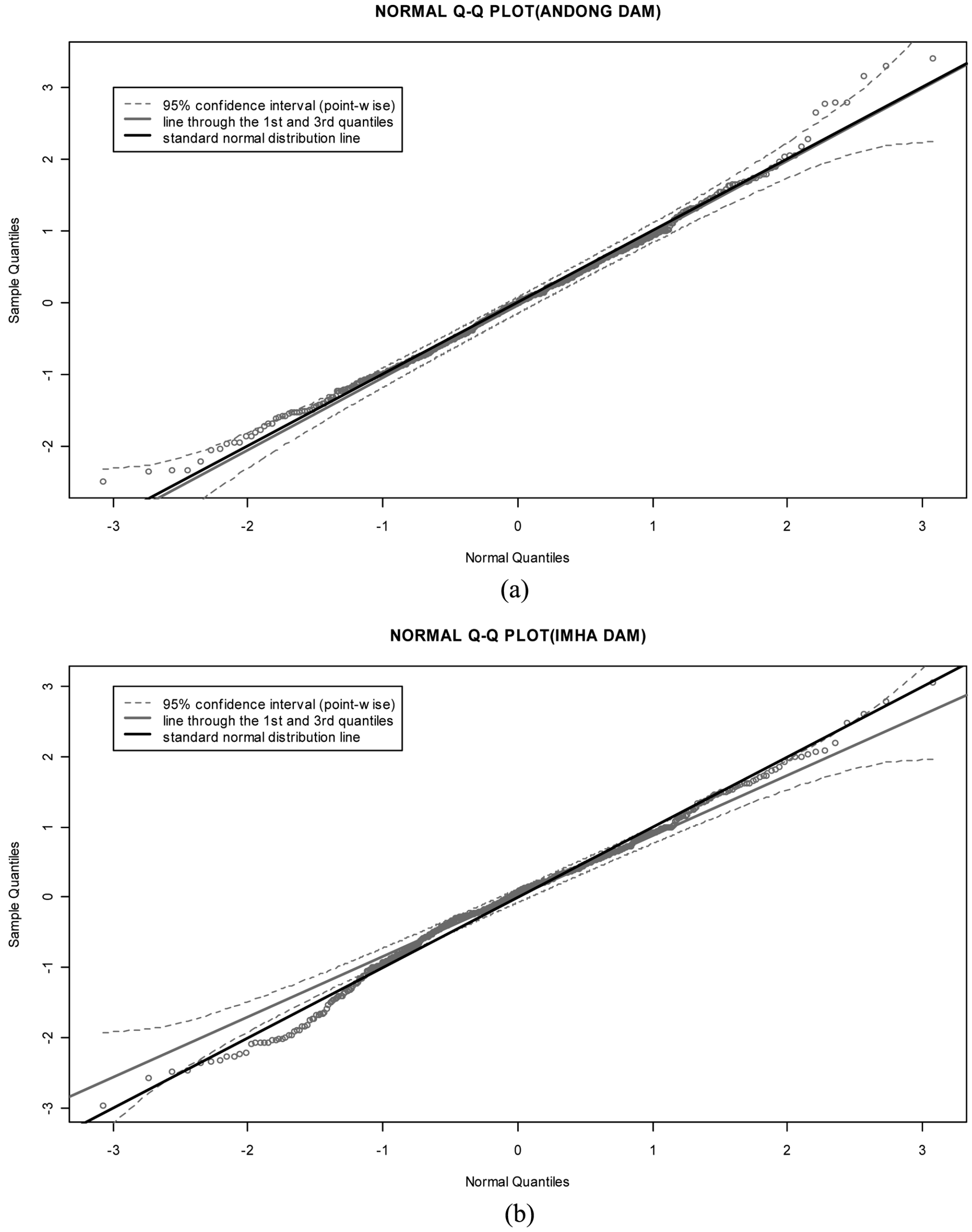
4. Model Development

An autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. The value of p is called the order of the AR model. Mathematically, it can be expressed as:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + e_t \qquad (3)$$

where $\phi_1, \phi_2, ..., \phi_p$ are the autoregressive coefficients, X_t is the time series and e_t is the residuals.

AR models can be analyzed with one of various methods, including standard linear least squares techniques. One of the most popular is the Box-Jenkins model (Box and Jenkins, 1970). It is a combination of AR and moving average (MA) models. This modeling proceeds by a series of well-defined steps, namely: (1) model identification; (2)model estimation; and (3) model validation.



(a)

(b)

Fig. 2. Q-Q Plot of the Transformed Series: (a) Andong Dam (b) Imha Dam

4.1 Model Identification

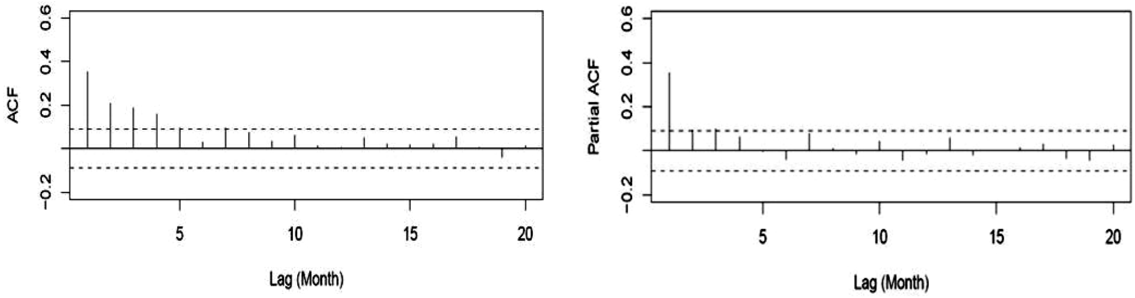
An important diagnostic tool for examining dependence and identifying the order of the AR model is the sample-autocorrelation function. Randomness is ascertained by computing autocorrelations for data values at varying time lags. If random, such autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.

Consider a time series, autocorrelation function for a variety of lags = 1, 2,..., *k* can be estimated by computing sample correlation between the pairs units apart in time. This can be modified slightly, taking into account that stationarity is assumed, which implies a common mean and variance for the series. Sample autocorrelation function (ACF), at lag can be computed by:

$$r_k = \frac{\sum_{t=k+1}^{n} [X_t - \overline{X}][X_{t-k} - \overline{X}]}{\sum_{t=1}^{n} (X_t - X)^2} \tag{4}$$

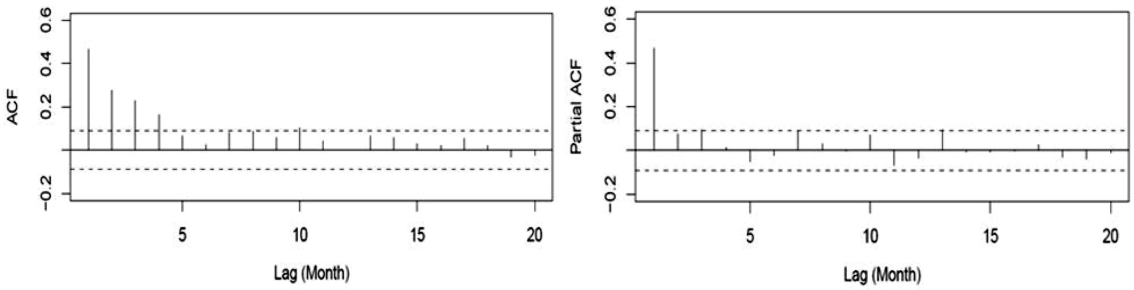
where $\overline{X}$ is the “grand mean”. For a variety of reasons, this has become the standard definition for the sample ACF (Cryer and Chan, 2008).

Fig. 3 shows the ACF and partial autocorrelation function (PACF) plots for Andong Dam, while Fig. 4 shows that of Imha Dam. The dashed horizontal lines plotted at $\pm 2\sqrt{n} = \pm 0.091$ (with = 480 months for both series), are intended to give critical values for testing whether or not the autocorre-



(a) (b)

Fig. 3. Andong Dam Correlogram : (a) ACF (b) PACF



(a) (b)

Fig. 4. Imha Dam Correlogram : (a) ACF (b) PACF

lation and partial autocorrelation coefficients are significantly different from zero, also known as the 95% confidence interval. These limits are based on the approximate large sample standard error that applies to a white noise process, namely $1/\sqrt{n}$.

For Andong Dam, PACF is significant at lags = 1, 2 and 3. Lags = 1 and 3 are significant for Imha Dam. Further analysis was performed to identify the order of the AR model.

A number of other approaches to model specification have been proposed since Box and Jenkin's work. One of the most studied is Akaike's Information Criteria (AIC) (Akaike, 1973). An Akaike Information Criterion (AIC) is used to select the best among the competing models. The AIC is computed as if the variance estimate were the maximum likelihood estimate (MLE), omitting the determinant term from the likelihood. The AIC for an AR(p) model takes the form:

$$\mathrm{AIC}(p) = \mathrm{N}\ln(\sigma_{\varepsilon}^{2}) + 2p \qquad (5)$$

where σ_{ε}^{2} is the MLE of the variance and p is the order of the model.

For this study, the AIC is computed as if the variance estimate were the MLE, omitting the determinant term from the likelihood. The best model was selected by choosing the order with lowest AIC. Therefore, from this criterion, a third order autoregressive model was used for both gauging stations.

4.2 Estimation of Parameters

The method of moments and maximum likelihood are among the most popular methods of parameter estimation. The method of maximum likelihood was used in this study to estimate the parameters.

For Andong Dam, the parameters were estimated as $A_1 = 0.3113$, $A_2 = 0.0599$, and $A_3 = 0.1005$. The noise variance was also estimated as 0.837. Noting the standard errors, the estimates of the lags 1,2 and 3 autoregressive coefficients are significantly different from zero, as is the intercept (mean) term. The stimated model for Andong Dam can then be written as:

Andong Dam :

$$Y_{tA} = 0.3113Y_{t\text{-}1} + 0.0599Y_{t\text{-}2} + 0.1005Y_{t\text{-}3} + \xi_t\ ;\ \sigma^2 = 0.837$$

In the case of Imha Dam, the estimated parameters were $A_1 = 0.4270$, $A_2 = 0.0310$, and $A_3 = 0.0963$ with a noise variance equal to 0.7499. Since autoregressive terms were significantly different from 0, the streamflow series of Imha Dam could be modeled as:

Imha Dam :

$$Y_{tI} = 0.4270Y_{t\text{-}1} + 0.0310Y_{t\text{-}2} + 0.0963Y_{t\text{-}3} + \xi_t\ ;\ \sigma^2 = 0.7499$$

4.3 Evaluation of Models' Performance

Synthetic monthly streamflows were computed using the determined AR(3) model for Andong and Imha Dam. 410 years of record had been generated. Since it is necesarry to use average values of the antecedent flows to start the simulation procedure in year one, the first few generated monthly flows will not necessarily occur randomly. The first and last five years were arbitrarily discarded to insure that the important statistical properties were retained.

It is necessary to generate data to insure that the method preserves the natural streamflow characteristics. Synthetic years 6 through 405 were divided into ten consecutive segments of record of the same length as the original streamflow record. The mean, standard deviation and skewness coefficient were computed of each month of the ten segments of the synthetic record. Table 4 and Table 5 shows the comparison of the historical and synthetic streamflow record statistics. From these table, it is evident that determined AR(3) model was able to reproduce fairly the historical mean and standard deviation. However, the skewness doesn't seem to be fairly reproduced primarily. Salas et al. (1980) mentioned that skewness is highly uncertain, so whether or not the model is able to reproduce precisely the estimated skewness depends on how long the sample was, and how important the skewness is for the model application.

Most hydrological time series are represented by an AR(2) model, however, for this study, AR(3) model was found to be suitable. This suggests that the historical record of streamflow for Andong and Imha Dam can be reproduced by a model which depends on the previous three months. Evaluation of the model accuracy which included residual analysis and overfitting and parameter redundancy proved that AR(3) model performs better in representing the Andong and Imha Dam streamflow.

Table 4. Basic Statistics of Historical and Synthetic Monthly Streamlfow for Andong Dam

Month	Mean (m^3/sec)		Standard Deviation (m^3/sec)		Skewness (unitless)	
	Historical	Synthetic	Historical	Synthetic	Historical	Synthetic
January	5.45	4.88	5.05	2.81	3.52	1.27
February	7.51	6.51	9.82	6.46	3.14	2.49
March	17.55	17.63	16.48	17.13	2.44	2.31
April	33.32	31.11	24.94	24.87	1.13	1.75
May	26.45	25.21	22.01	25.29	1.21	2.67
June	32.41	33.03	28.68	31.16	1.40	2.14
July	86.24	93.18	47.12	63.61	0.31	1.56
August	84.12	93.29	67.76	87.36	1.50	1.92
September	58.83	52.44	49.91	49.41	1.26	1.99
October	16.54	15.84	13.05	10.58	2.21	1.97
November	11.51	11.18	8.88	6.72	2.70	1.63
December	7.06	6.67	4.95	3.23	3.57	1.10

Table 5. Basic Statistics of Historical and Synthetic Monthly Streamlfow for Imha Dam

Month	Mean (m^3/sec)		Standard Deviation (m^3/sec)		Skewness (unitless)	
	Historical	Synthetic	Historical	Synthetic	Historical	Synthetic
January	5.15	4.96	6.16	4.19	3.18	2.02
February	8.99	8.90	14.27	11.83	3.16	2.52
March	17.89	18.67	18.30	23.81	1.90	2.62
April	29.24	29.80	25.94	27.31	1.97	1.85
May	25.97	24.73	21.91	22.75	1.31	2.10
June	33.77	33.58	31.07	35.67	1.82	2.74
July	81.97	88.18	56.92	90.36	0.72	2.48
August	93.56	99.92	71.27	111.18	1.64	2.28
September	64.96	67.96	56.09	70.96	1.16	1.93
October	17.00	17.53	9.04	9.50	1.26	1.53
November	10.42	10.71	5.21	5.34	1.33	1.02
December	6.40	6.40	5.01	3.77	3.55	1.26

5. Conclusion

In this study, the use of an autoregressive model for the historical data of two gauging stations, namely, Andong Dam and Imha Dam located at the upper most part of Nakdong River Basin was investigated. Several AR models where analyzed to determine the best fitting model. However, AR(3) model was found to be suitable for reproducing the historical streamflow record, both for Andong Dam and Imha Dam. This indicates that an autoregressive model using a 3-month lag can be used to estimate the current streamflow. The AR(3) model for Andong and for Imha Dam where evaluated using several model diagnostic tests such as residual analysis and overfitting and parameter redundancy. These tests proved the accuracy of the determined AR(3) models. For verification, synthetic streamflow were generated using the AR(3) models. The mean, standard deviation and skewness coefficient were also computed and compared to that of the historical record. Although the model reproduce the mean and standard deviation fairly, the model was not able to reproduce the skewness coefficient. This could be attributed to the short sample years used in the synthetic monthly streamflow generation.

With these findings, it is adequate enough to conclude that the determined AR(3) model is accurate and represents the time series efficiently. Findings from this study can contribute to understanding the physical systems by revealing something about the physical process that builds persistence into the seris. However, this study was based on historical data which assumes stationarity of the series. This indicates that nonstationarity such as the effects of climate change were not considered.

References

Akaike, H. (1973) Maximum likelihood identification of Gaussian auto-regressive moving-average models, *Biometrika*, Vol. 60, pp. 255-266.

Box, G.E.P. and Jenkins, G.M. (1970) *Time Series Analysis Forecasting and Control.* Haden-bay Press, San Francisco, California.

Box, G.E.P. and Pierce, D.A. (1970) Distribution of residual correlations in autoregressive-integrated moving average time series models, *Journal of American Royal Statistical Society B*, Vol. 65, pp. 1509-1526.

Cryer, J.D. and Chan, K.S. (2008) *Time Series Analysis with Application in R.* Second Edition, Springer, New York.

Maidment, D.R. (1993) Handbook of Hydrology, McGraw-Hill, Inc.

Ministry of Construction and Transportation (MOCT) (2008). Water Management Information System (WAMIS): Daily Streamflow Series for Andong Dam (W.Y. 1966-2005) and Imha Dam (1966-2005) Gauging Stations. Available at:http://www.wamis.go.kr/WKW/wkw_cms_lst.aspx

Salas, J.D., Delleur, J.W., Yevjevich, V. and Lane, W.L. (1980) *Applied Modeling of Hydrological Time Series.*2 Water Resources Publications, Littleton, Colorado.

Thomas, H.A. and Fiering, M.B. (1962) Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. *Design of Water Resources Systems*, Edited by Mass, A. et al., Harvard University Press, Cambridge, Massachusetts, pp. 459-493.

Yevjevich, V., (1963) Fluctuations of wet and dry years: Part I - Research data assembly and mathematical models. *Hydrology Paper 1*, Colorado State University, Fort Collins, Colorado.

◎ 논문접수일 : 09년 12월 17일
◎ 심사의뢰일 : 09년 12월 17일
◎ 심사완료일 : 10년 01월 05일