# Uncertainty Analysis for Synthetic Streamflow Generation

Dong-Jin Lee[1], Jose D. Salas[2], and Duane C. Boes[3]

[1] Graduate Student, Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, CO 80523, USA; PH (970) 491-4302; FAX (970) 491-7727; email: djlee@engr.colostate.edu

[2] Professor, Department of Civil & Environmental Engineering, Colorado State University, Fort Collins, CO 80523, USA; PH (970) 491-6057; FAX (970) 491-7727; email:jsalas@engr.colostate.edu

[3] Emeritus Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.

***Abstract***

Synthetic streamflow generation has been widely used in hydrology and water resources since the 1960's for a number of practical problems such as determining the capacity of a reservoir and assessing the long-term behavior of an existing reservoir. Synthetic streamflows can be obtained using parametric and non-parametric approaches. The former assumes that a certain mathematical model describes the stochastic behavior of the underlying process, e.g. streamflow. And the mathematical model hinges on a number of parameters that must be estimated from historical data. If the available historical data would be sufficiently long (e.g. hundreds of years), the model parameters could be estimated with a good precision, the synthetic samples produced from the model would reflect the expected variability of the process under consideration, and consequently the expected variability of the design variables obtained from them (e.g. the size of the needed storage capacity for a reservoir). However, the usual lengths of historical streamflow records are short which means that the model parameters are uncertain and consequently the variability of the design variables may be uncertain beyond what is expected. A number of approaches have been proposed in literature to tackle the problem of parameter uncertainty in simple stochastic models. In the paper described herein we take an

approach based on the asymptotic distribution of the parameter estimators of an AR(1) model and investigate in some detail the effect of the uncertainty in one or more parameters on the design variables such as the reservoir size and reliability. Our analysis has been conducted based on simulation studies of an AR(1) model for a wide range of parameters. The paper includes an example to illustrate the applicability of the concepts obtained in the study.

## *Introduction*

Uncertainties in water resources commonly arise from the random nature of hydrological process and from the limited information (data) that is available regarding true nature of the underlying process. Since the parameters of stochastic models are estimated using the limited historical records, these estimates are uncertain quantities. These uncertainties in the paprame6ers of stochastic models are translated into the uncertainty of decision variables of planning and management of water resources. Perhaps the simplest example may be case of designing a flood related structure for the 100-yr flood. The flood frequency distribution is an expression of the natural uncertainty of the underlying extreme floods but the magnitude of a specific flood quantile, e.g. the 100-yr flood quantile, is uncertain. It is well known that such uncertainty is commonly expressed by determining the confidence limits of the population quantile (e.g. Stedinger et. al., 1993).

Likewise, conventional approaches for designing the capacity of a reservoir generally consider the effect of the natural uncertainty of streamflows. For example this is done by building a stochastic model and simulating synthetic flow records from which the frequency distribution of the needed reservoir storage capacity can be obtained. However, as is the case for the flood protection design problem illustrated above, the stochastic model of streamflows has a parameter set that is uncertain because of the limited data available and as a result the distribution of decision variables related to them, e.g. reservoir capacity, is also uncertain. Although this issue has been recognized in the past and some procedures have been suggested (e.g. Vicens et al., 1975; Wood, 1978; Valdes et al., 1977; McLeod and Hipel, 1978; Salas et al., 1980; Klemes et al., 1981; Grygier and Stedinger, 1990) unfortunately the problem remains perhaps because of its complexity and the lack of understanding of the many factors involved.

In this paper we report results of a systematic study of the effect of parameter uncertainty of a lag-1 autoregressive model, which is a commonly used stochastic model for generating synthetic streamflow data, on the size of the storage capacity of a reservoir.

***Stochastic Model and Parameter Uncertainty***

We assume that the underlying annual streamflows, denoted by $X_t$, is lognormal distributed such that $Y_t = \log_e(X_t)$ is normally distributed. The mean $\mu_Y$ and variance $\sigma_Y^2$ of $Y_t$ may be expressed as a function of the mean $\mu_X$ and coefficient of variation $\eta_X$ of $X_t$ as (e.g. Yevjevich, 1972);

$$\mu_Y = \frac{1}{2}\ln\left(\frac{\mu_X^2}{1+\eta_X^2}\right) \tag{1}$$

$$\sigma_Y^2 = \ln\left(1+\eta_X^2\right) \tag{2}$$

Then assuming a lag-1 autoregressive model, AR(1), the variable $Y_t$ may be written as

$$Y_t = \mu_Y + \phi\left(Y_{t-1} - \mu_Y\right) + \varepsilon_t \tag{3}$$

where $\phi_Y$ is the autoregressive coefficient that satisfies the causal condition in the range of $-1 < \phi_Y < 1$ and corresponds to the lag-1 serial correlation of $Y_t$ and $\varepsilon_t$ is the time independent innovation term with mean 0 and variance $\sigma_\varepsilon^2$. The relationship between $\phi_X$ and $\phi_Y$ is given by (Matalas, 1967)

$$\phi_Y = \frac{1}{\sigma_Y^2}\log_e\left[\phi_X \exp\left(\sigma_Y^2\right) - \phi_X + 1\right] \tag{4}$$

and the variance of the innovation term is given by

$$\sigma_\varepsilon^2 = \sigma_Y^2\left(1-\phi_Y^2\right) \tag{5}$$

The asymptotic distribution of the maximum likelihood (ML) estimators of the parameters of an autoregressive moving average (ARMA) model is available (e.g. Box and Jenkins, 1976; McLeod and Hipel, 1978) and the uncertainty of the parameters can be quantified in terms of their variances. Asymptotic theory might be applicable when the sample size is at least greater than 50 (e.g. Haugh, 1976). For the AR(1) model, the asymptotic distributions of the estimators of the parameters $\hat{\phi}_Y$, $\hat{\mu}_Y$, and $\hat{\sigma}_\varepsilon^2$ are given by (Box and Jenkins, 1976)

$$\hat{\phi}_Y \sim Nor\left(\phi_Y, \frac{1}{N}\left(1-\phi_Y^2\right)\right) \tag{6}$$

$$\hat{\mu}_Y \sim Nor\left( \mu_Y , \frac{1}{N}\left( \frac{\sigma_\varepsilon}{1-\phi_Y} \right)^2 \right) \tag{7}$$

and

$$\hat{\sigma}_\varepsilon^2 \sim N\left( \sigma_\varepsilon^2 , \frac{2\sigma_\varepsilon^4}{n} \right) \tag{8}$$

where, $\hat{\mu}_Y$, $\hat{\phi}_Y$, $\hat{\sigma}_\varepsilon^2$ are ML estimators of $\mu_Y$, $\phi_Y$, $\sigma_\varepsilon^2$ and N represents sample size.

***Effect of Parameter Uncertainty on Synthetic Flow Statistics***

To evaluate the effect of parameter uncertainty we examined six cases: (a) no uncertainty is considered, i.e. all parameters are assumed as constant values, (b) only uncertainty of $\mu_Y$ is considered, (c) only uncertainty of $\phi_Y$ is considered, (d) only uncertainty of $\sigma_\varepsilon^2$ is considered, (e) the uncertainties of $\phi_Y$ and $\sigma_\varepsilon^2$ are considered, and (f) the uncertainties of all parameters $\mu_Y$, $\phi_Y$, and $\sigma_\varepsilon^2$ are considered. Also, we evaluate the effect of uncertainty in terms of the coefficient of variation $\eta_X$ and lag-1 serial correlation $\phi_X$ of the original (non-transformed) series and different values of $\eta_X$ and $\phi_X$ are employed; $\eta_X$ =0.1- 2.0, $\phi_X$ = 0 - 0.9, with increment of 0.02, respectively.

We conduct simulation experiments assuming $\mu_X$ = 10 and the AR(1) model parameters of $Y_t$ are estimated using Eqs. (1)-(2) and (4)-(5) for the assumed values of $\eta_X$ and $\phi_X$. Four sample sizes are considered: $N$ = 25, 50, 75, and 100, and the uncertainty of the parameters $\mu_Y$, $\phi_Y$, and $\sigma_\varepsilon^2$ are determined from (6)-(8) and the4 parameters are sampled fro the respective distribution according to the cases (a) – (f). Thus for a given parameter set 10,000 samples of synthetic streamflow series are generated and for each set the storage capacity is calculated. The storage capacity is calculated using the sequent peak algorithm (SPA) as (Loucks et al., 1981)

$$S_t = \max(0, S_{t-1} + D_t - X_t) \quad , \quad t = 1,....,N \tag{10a}$$

where $D_t$ = water demand, $X_t$ = reservoir inflow, and $S_0 = 0$. Then the storage capacity becomes

$$S_c = \max(S_0, S_1, ...., S_N) \qquad (10b)$$

For the purpose of this study we assumed that the demand level $D_t$ is equal to the mean. However, we considered three options: (i) the historical sample mean, which is assumed to be fixed for all generated samples, this option is labeled FM; (ii) the sample means obtained from each generated sample is utilized as the demands (SM); and (iii) the sample mean is obtained (sampled) from the asymptotic distribution (7) and remains fixed for all the generated samples, this option is labeled as PM.

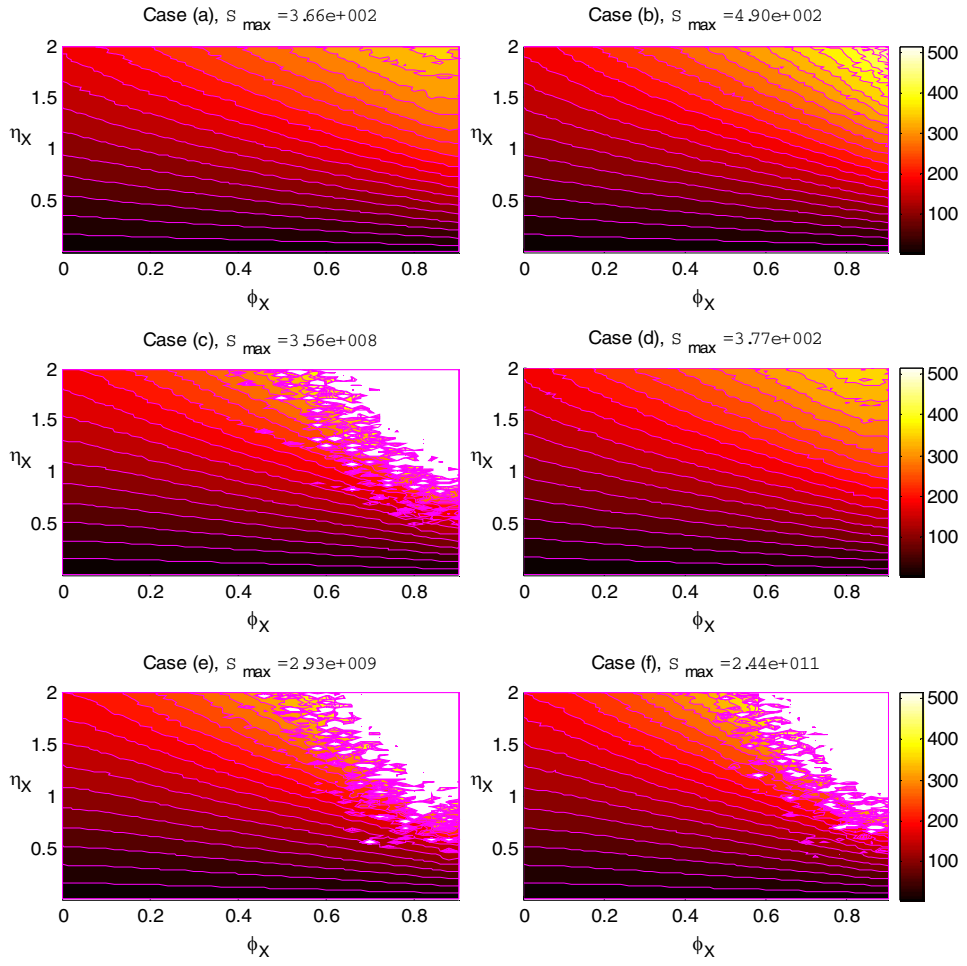



Figure 1: Expected value of storage capacity computed from 10,000 generated traces for each combination of the coefficient of variation $\eta_X$ and autoregression coefficient $\phi_X$. The storage capacity was calculated with demand $D_t$ as option (ii), i.e. $D_t$ equal to the sample mean (SM) obtained from each generate sample. The cases (a) – (f) correspond to the various alternatives of uncertainty considered and the value $S_{max}$ shown in each case is the maximum value of the storage capacity obtained from all combinations of $\eta_X$ and $\phi_X$ utilized.

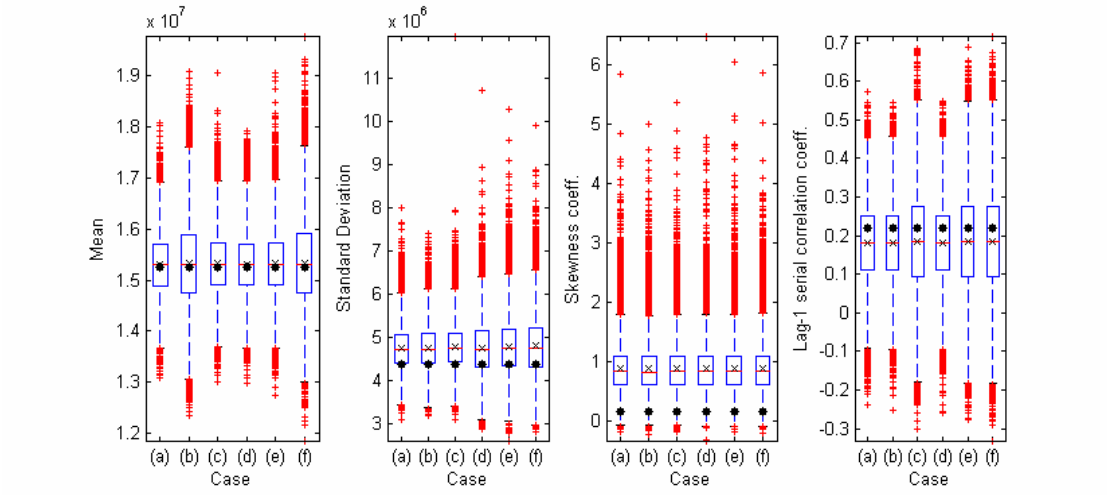
Figure 1 illustrates the expected value of storage capacity obtained over the specified ranges of values of $\eta_X$ and $\phi_X$ for a sample size of N=100 and demand option *SM*. For $\eta_X \leq 0.5$, about similar patterns are observed for the various cases analyzed. Overall, the expected value of storage capacity increases as $\phi_X$ increases or $\eta_X$ increases; also the effect of $\eta_X$ on the expected storage capacity is larger than that of $\phi_X$. Figure 1 (a)-(f) shows that the effect of the uncertainty of $\sigma_\varepsilon^2$ seems to be less important that the effect of the uncertainty of $\mu_Y$ or $\phi_Y$. Remarkably, the effect of the uncertainty of $\phi_Y$ seems very significant.

Figure 1 (c), (e), and (f) illustrate the cases where the effect of $\phi_Y$ is included. The figure suggests that the effect of the uncertainty of $\phi_Y$ dominates over the effect of the other two. Also in all three cases for a combination of high value of $\eta_X$ (say in the range 1-2) and high value of $\phi_X$ (say bigger than 0.5) very large values of the storage capacity may occur. This could occur for example if $\hat{\phi}_Y$ and N=50 so that $\hat{\phi}_Y \sim Nor(0.7321, 0.0963^2)$. Sampling from this distribution may lead to values of $\phi_Y$ close to 1, which in turn may produce very large values of streamflows, a large value of the sample mean, and consequently a very large value of the storage capacity. Note the very large values of the storage capacity obtained for cases (c), (e), and (f). On the other hand, this possibility of very large values of the storage capacity are not found for the other cases, i.e. (b) and (d), where the uncertainties of $\mu_Y$ or $\sigma_\varepsilon^2$ are considered. The foregoing results correspond to the demand option SM. On the other hand, for the demand options *FM* and *PM* no extremely large storage capacities are obtained regardless of the number of generated traces. For all cases considered, *PM* gives larger storage capacities than *FM*. Storage capacities are shown to increase with larger variation and larger serial correlation. It shows consistency with the serial correlation when the variation is small; i.e. $\eta_X \leq 0.5$. *PM* produces larger storage capacities over the whole ranges of $\eta_X$ and $\phi_X$ when compared with *FM*. As expected, the case (f), which includes the effect of the uncertainties in all parameters gives the larger storage capacities as compared with other cases.

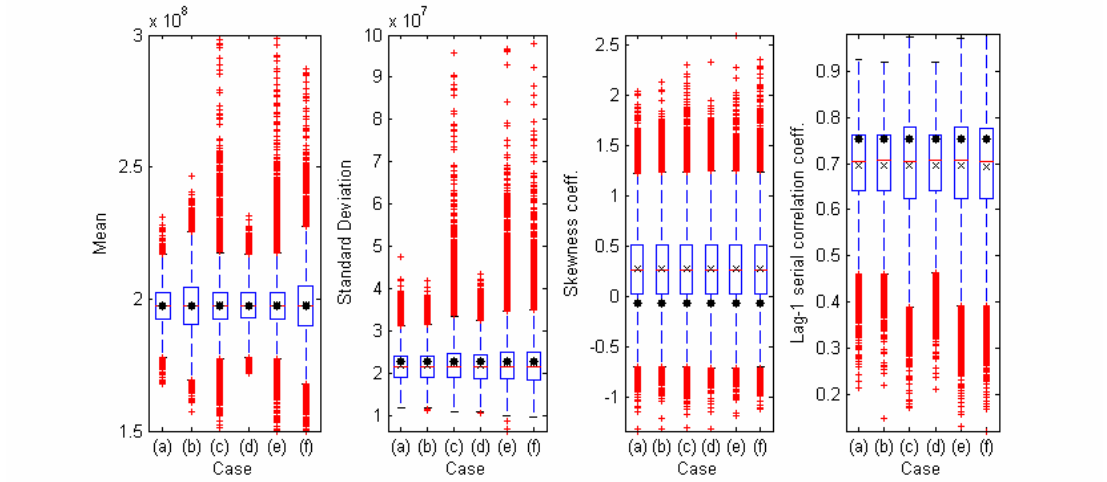
***Uncertainty analysis using real streamflow data***

Two annual streamflow series one with high lag-1 serial correlation and another with low serial correlation are chosen to further study the effect of parameter uncertainty on synthetic streamflows. The annual flows of the St. Lawrence River at Cornwall, Ontario near Massena, NY is selected, it has the following characteristics: $\eta_X$=0.11, $\phi_X$=0.75, skewness coefficient equal to -0.06, and kurtosis coefficient equal to 2.29. Additionally, the annual flows of the Colorado River at Lee's Ferry site is analyzed where $\eta_X$=0.28, $\phi_X$=0.22, skewness coefficient equal to 0.16, and

kurtosis coefficient 2.36. Log-transformation without location parameter has been employed as needed to transform the original data to the normal domain. Six cases of parameter uncertainty are considered (as discussed above), and 30,000 synthetic traces were generated using the AR(1) model. The results have been compared in terms of basic statistics and storage capacity.




(a) Colorado River at Lee's Ferry

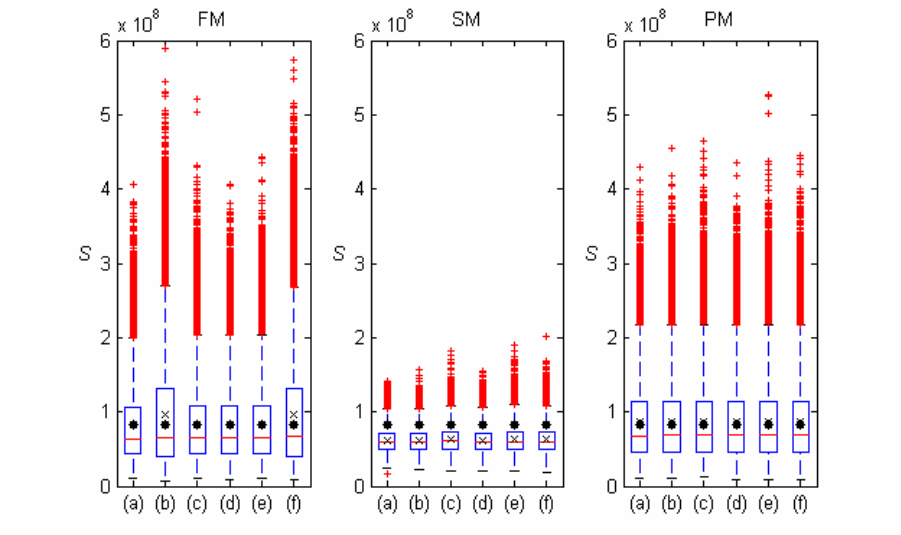



(b) St. Lawrence River

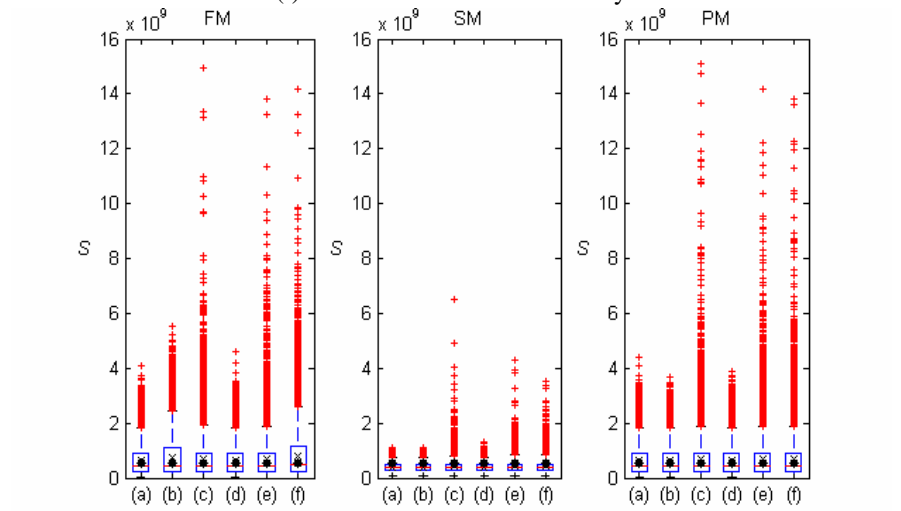
Figure 2. Box plots for the mean, standard deviation, skewness, and lag-1 serial correlation coefficients obtained from 30,000 generated annual streamflow series for (a) Lee's Ferry and (b) St. Lawrence River. The cases (a)-(f) in each plot refer to the 6 cases of parameter uncertainty considered. The symbols '•' and 'x' denote historical value and averaged value from the 30,000 synthetic traces.

Parameter uncertainty effects on simulated streamflows are shown in Figure 2 based on generated mean, standard deviation, skewness, and lag-1 serial correlation coefficient. For low correlated streamflows, the uncertainty of the mean parameter has a great effect on the mean variability. And, the uncertainty of the lag-1

correlation coefficient produces a considerable dispersion of the generated lag-1 correlation coefficients as well. For the highly correlated streamflows the effect of the uncertainty of the serial correlation is quite significant on the mean and standard deviation. As expected the case (f) which includes the uncertainties in all parameters is shows the major effects in the variability of the mean, standard deviation, skewness, and lag-1 correlation coefficients.

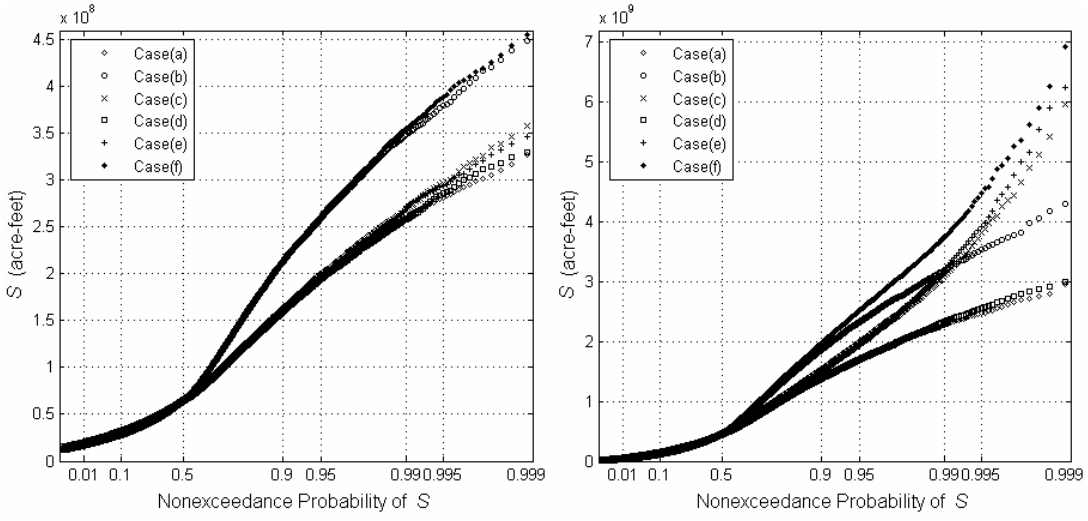



(i) Colorado River at Lee's Ferry




(ii) St. Lawrence River

Figure 3. Box plots of generated storage capacities for (i) Lee's Ferry data and (ii) St. Lawrence River. The symbols '•' and 'x' denote historical value and averaged value from the 30,000 synthetic traces, respectively. (unit: acre-feet).




(a) Colorado River at Lee's Ferry (b) St.Lawrence River

Figure 4. Frequency distribution of storage capacities plotted on Gumbel probability paper for the demand option FM for (i) Colorado River at Lee's Ferry and (ii) St. Lawrance River.

Table 1. Storage capacities calculated from the generated data using the demand option *FM* and 30,000 synthetic traces for (i) Colorado River at Lee's Ferry and (ii) St. Lawrance River. The % column (shaded) next to the estimated storage capacity column represents the percentage difference with respect to storage capacity of Case(a) in each nonexceedance probability section.

| Case | Nonexceedance probabilities | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0.9 | % | 0.95 | % | 0.99 | % | 0.995 | % | 0.999 | % |
| Lee's Ferry (unit: $10^8$ acre-feet) | | | | | | | | | | |
| (a) | 1.60 | 100 | 1.95 | 100 | 2.6 | 100 | 2.82 | 100 | 3.28 | 100 |
| (b) | 2.13 | 133 | 2.59 | 133 | 3.49 | 134 | 3.79 | 135 | 4.49 | 137 |
| (c) | 1.62 | 101 | 1.99 | 102 | 2.69 | 103 | 2.97 | 105 | 3.58 | 109 |
| (d) | 1.60 | 100 | 1.95 | 100 | 2.59 | 99 | 2.86 | 102 | 3.30 | 101 |
| (e) | 1.62 | 101 | 1.97 | 101 | 2.72 | 105 | 2.94 | 105 | 3.47 | 106 |
| (f) | 2.12 | 132 | 2.61 | 134 | 3.55 | 137 | 3.89 | 138 | 4.59 | 140 |
| St. Lawrence River (unit: $10^9$ acre-feet) | | | | | | | | | | |
| (a) | 1.40 | 100 | 1.70 | 100 | 2.28 | 100 | 2.47 | 100 | 2.95 | 100 |
| (b) | 1.87 | 133 | 2.33 | 137 | 3.20 | 140 | 3.54 | 143 | 4.30 | 146 |
| (c) | 1.52 | 108 | 1.97 | 116 | 3.11 | 136 | 3.81 | 154 | 5.96 | 202 |
| (d) | 1.39 | 99 | 1.70 | 100 | 2.34 | 102 | 2.55 | 103 | 3.00 | 102 |
| (e) | 1.50 | 107 | 1.95 | 114 | 3.20 | 140 | 3.93 | 159 | 6.27 | 212 |
| (f) | 1.97 | 141 | 2.51 | 148 | 3.75 | 164 | 4.49 | 182 | 6.99 | 237 |

Figure 3 shows the frequency distribution of the storage capacity obtained from the 30,000 synthetic flow traces for both the Colorado River at Lee's Ferry and the St. Lawrence River, with demand options *FM*, *SM*, and *PM*. As discussed previously the effect of the uncertainties of the serial correlation are significant for the highly

correlated streamflows for all demand options *FM*, *SM* and *PM*. For the options *SM* and *PM* the effects of the uncertainty of the mean are found to be negligible. The frequency distributions of the storage capacities calculated from the 30,000 synthetic traces are plotted on a Gumbel probability paper in Figure 4. It is found that the uncertainty of the mean parameter causes the larger storage capacity for given nonexceedance probability for the case of low serial correlated streamflow series (Colorado River at Lee's Ferry) while the uncertainty of the lag-1 correlation coefficient makes a significant effect on the amount of storage capacities for highly correlated series (St. Lawrence River). Table 1 illustrates the estimated storage capacities for nonexceedance probabilities $q$=0.9, 0.95, 0.99, 0.995, 0.999 (risk of 0.1, 0.02, 0.01, 0.005, 0.001, respectively) which correspond to return period of 10, 20, 100, 200, and 1000 years. The shaded columns next to estimated storage capacity columns represent the percentage difference with respect to storage capacity of case (a). For example, based on 0.05 risk over the next design life of 95 years, the same as the sample size of Lee's Ferry, storage capacities can be determined as $1.95\times10^8$, $2.59\times10^8$, $1.99\times10^8$, $1.95\times10^8$, $1.97\times10^8$, and $2.61\times10^8$ acre-feet for each parameter uncertainty consideration case, respectively. That is, 33% and 34% increased storage capacities are expected when uncertainties of mean parameter are incorporated in the simulation of annual streamflows based on Lee's Ferry streamflow data with consideration of 5% risk in the future.

***References***

Box, G.E.P. and G.W. Jenkins (1976) *Time Series Analysis Forecasting and Control*, revised edition, Holden-Day, San Francisco, USA.

Burges, S.J. (1970). *Use of Stochastic Hydrology to Determine Storage Requirements of Reservoirs – A Critical Analysis*, Ph.D Dissertation, Stanford University, CA, USA.

Grygier, J.C. and J.R. Stedinger (1990). *SPIGOT*, *A synthetic Streamflow Generation Software Package*, *technical description*, *version 2.5*, School of Civil and Environmental Engineering, Cornell University, Ithaca, N.Y., USA

Hashimoto T., J.R. Stedinger, and D.P. Louck (1982). "Reliability, resiliency and vulnerability criteria for water resource system performance evaluation", *Water Resources Research*, 18(1), pp. 14-20.

Haugh, L.D. (1976). "Checking the independence of two covariance-stationary time series: A univariate residual cross-correlation approach", *Journal of the American Statistical Association*, 71(354), pp. 378-385.

Klemes, V., R. Srikanthan and T.A. McMahon (1981). „Long-memory flow models in reservoir analysis: What is their practical value?", *Water Resources Research*, 17(3), pp. 737-751.

Matalas, N.C. (1967). "Mathematical assessment of Synthetic Hydrology", *Water Resources Research*, 3(4), pp. 937-945.

McLeod, A.I. and K.W. Hipel (1978). *Simulation Procedures for Box-Jenkins Models, Water Resources Research*, 14(5), 969-975.

McMahon, T.A. and A.J. Adeloye (2005). *Water Resources Yield*, Water Resources Publications, Littleton, Colorado, USA.

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

Valdes, J.B., I. Rodriquez-Iturbe, and G.J. Vicens (1977). "Bayesian Generation of Synthetic Streamflows: 2. The Multivariate Cases", *Water Resources Research*, 13(2), pp. 291-295.

Vicens, G.J., I. Rodriguez-Iturbe, and J.C. Schaake, Jr. (1975). "Bayesian generation of synthetic streamflows", *Water Resources Research*, 11(6), pp. 827-838.

Vogel, R.M. and R.A. Bolognese (1995). "Storage-reliability-resistance-yield relations for over-year water supply systems", *Water Resources Research*, 31(3), pp. 645-654.

Wood, E.F. (1978). "Analysing hydrologic uncertainty and its impact upon decisionmaking in water resources", *Advances in Water Resources*, 1(5), pp. 299-305.

Yevjevich, V. (1972). *Probability and Statistics in Hydrology*, Water Resources Publications, Fort Collins, Colorado, USA.