Principles of Stochastic Generation of Hydrologic Time Series for Reservoir Planning and Design: A case study

Rita Guimarães[1*] and Emídio Gil Santos[2]

1 ICAAM, Depto de Engenharia Rural, Universidade de Évora, Núcleo da Mitra, Apartado 94, 7002-554 Évora, Portugal.

2 Depto de Engenharia Civil e Arquitectura, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001, Lisboa, Portugal.

* Corresponding author. Tel: +351-266-760-823, Fax: +351-266-760-911 E-mail: rcg@uevora.pt

Abstract: Simulation has been an important tool for planners in many fields of knowledge. In the field of water resources the uncertainties due to unknown data population and the short length of the records work together to make the simulation especially important. The major utilization of water resources at the level needed in modern society makes water storage essential for satisfying the demand. Therefore, the need to reduce the uncertainty in the design of water storage capacity is an important problem in the field of water resources utilization. This problem can only be satisfactorily solved with the aid of simulation. On the other hand, the implementation of adequate exploration politics also needs to use simulations in order to obtain results with low uncertainty. In simulation of water resources systems the use of synthetic time series is a current practice. In this study the number of generated time series to use in the problems described are analyzed. Annual and monthly synthetic flows were generated preserving the relevant statistics of the available historical data and then the number of time series to generate was determined. The results of the case studies indicate that the proposed methodology is a plausible approach to solve the analyzed problem and lead to the conclusion that the number of time series to generate should be *1200.*

Key Words: Simulation; Stochastic hydrology; Planning of water resources systems; Number of time series to generate.

Introduction

Synthetic hydrology is an accepted practice for the planning and management of water resources systems. Keeping in mind that an observed flow sequence most likely will not reoccur in the future and knowing that such sequence represents a single sample among all possible events, it is easy to understand the importance of data flow generation in designing and managing water resources systems. The generated time series are alternative flows values with statistical properties similar to the historical data and with the same probability of occurrence and as such they allow the simulation of various possible scenarios.

In analysis that involves data generation two main questions arise:

(i) What should be the number of time series to generate?

(ii) What should be the length of the generated time series?

A brief review of literature shows that there is no consensus in this matter. For example, in studies involving storage capacity, Wallis and Matalas (1972) utilized 500 time series with length equal to 100; Hoshi and Burges (1978) used 1000 time series each 40 years long; Klemes *et al.* (1981) utilized 1000 time series with length equal to 30 years; Vogel and Stedinger (1987) utilized 1000 time series with length equal to 20 and 60 years; Vogel and Stedinger (1988) used 10000 time series with lengths of 20, 40 and 80 years; Phien (1993) used 2000 time series with lengths ranging from 20 to 50 years and Adeloye *et al.* (2001) utilized 1000 time series with length varying from 10 to 1000 years. In other kind of studies involving hydrologic time series, Yue *et al.* (2002a) used 2000 time series with lengths of 10,…,1000 (step10); Yue *et al.* (2002b) used 1000 time series with lengths of 20, 50, 100, 150 years and Yue and Wang (2002) utilized 2000 time series with lengths of 20, 40, 60, 80 years.

In an attempt to identify components that contribute to uncertainty in estimating the required storage distribution Burges (1970) recommends that the number of series to generate should be equal to 1000 and the length of the generated series should be equal to 40 years. Salas (1993) states that the answers to the questions (i) and (ii) depend on the problem under consideration and as practical guidelines recommends that “when data generation is required for designing a reservoir, if annual data are used, as many as 1000 samples may be needed to accurately define the probability distribution of the maximum storage required. On the other hand, if monthly flows are used, fewer samples may be adequate”. Regarding the length of the generated series, Salas (1993) asserts that the length “must be equal to the planning horizon or economic life of the reservoir being designed”. McMahon and Adeloye (2005) wrote that “the number of replicates depends on the application and on the streamflow

variability…1000 replicates usually are sufficient" and recommend that the length of the generated series should be equal to the length of historical records.

The objective of the present study is to answer the above first question when the data generation is performed for the purpose of designing a reservoir with the necessary capacity to supply the demand. With this objective in mind a methodology, summarized as follows, is proposed. First, several sets of synthetic annual streamflows are generated and then they are disaggregated in monthly time series. Next, for each set of synthetic time series, a set of storage capacities are found and for each set of storage capacity the storage with a specified reliability is determined by fitting an appropriate probability distribution. In the last step, the number of time series that leads, for a given reliability, to a stable estimate of the reservoir capacity, is determined.

Proposed Methodology

When water storage is needed to meet a specific demand value, a question arises: "How large does the reservoir capacity need to be to provide a given controlled release with an acceptable level of reliability?" (McMahon 1978). Knowing that the reservoir inflows are represented by the observed streamflow sequence and knowing that the streamflow process is a stochastic process, meaning that it is unpredictable, it is easy to understand that the sizing of a reservoir to meet a given demand based on the historical sequences is manifestly insufficient. For this reason the use of generated synthetic streamflow sequences is common practice to assess the probability of failure (or reliability) of the reservoir. A specific number of generated synthetic streamflow time series allows the determination of equal number of storage capacity which can be represented by a probability distribution and, in turn, this distribution allows the evaluation of the reliability of the reservoir.

In this study, reservoir reliability means that for all the possible inflows sequences to a reservoir, the storage having reliability *x* per cent is such that *x* per cent of the inflow sequences will meet the demand sequences with no shortage (Burges 1970). The reservoir reliability is assessed with the sequent peak algorithm (Thomas and Burden 1963) in conjunction with monthly synthetic streamflow time series, which are obtained by disaggregation of annual synthetic streamflow time series by the fragment method (Svanidze 1980). The annual time series are obtained by a stochastic model based on the two parameters Log-Normal distribution combined with the Wilson-Hilferty transformation (Wilson and Hilferty 1931). The demand is considered constant and established as a percentage

(*75%*) of the annual mean flow. The methodology proposed to annual time series arises if the annual series are independent. If annual series are dependent an appropriate stochastic model must be utilized.

The problem to be addressed is the evaluation of the number of the generated time series for the determination of the reservoir capacity necessary to meet the demand with a given reliability. In order to answer this question the following methodology was utilized:

1 – Set up a stochastic model that represents the historical annual flow sequence;

2 – Generate sets of, $s = 50,...,3000$ (step *50*), synthetic annual flow sequence with length (*n**) equal to the historical sequence length (*n*);

3 – Disaggregate, by the fragment method, each set of annual flow sequences into monthly flow sequences;

4 – Run these sets of monthly flow sequences through the sequent peak algorithm to find the sets of storage capacity sequences, $c_m^s,\ m = 1,2,...,s$;

5 – Fit an appropriate distribution to the storage capacity sequences;

6 – Find the storage capacity for various levels of reliability (*80, 85, 90, 95, 98 e 99%*) using the inverse of the probability distribution;

7 – Find the number *s* of generated time series that leads, for a given reliability, to a stable estimate of the reservoir capacity;

This methodology was implemented using five time series observed, in hydrological years, for four Portuguese rivers whose drainage basin localizations are presented in Figure 1. In Table 1 the characteristics of the gauging stations used in this study are presented and Table 2 contains the annual basic statistics of the time series. In this paper, the results for the station Fragas da Torre are presented, as an example, since the results for the other time series lead to the same conclusions.

Annual flows generation

A preliminary analysis of the historical time series was performed in order to check for trends, shifts and to examine the time series dependence. The application of the Mann-Kendall test as recommended by the World Meteorological Organization (1998) and the Mann-Whitney test (Salas 1993) revealed that the time series does not show trends neither shifts. The correlogram r_k, for

$k = 1,2,...,12$, with the confidence limits defined by Anderson (1941) shows that the time series is independent (Figure 2).

Taking into account these results, a model based on two parameters Log-Normal distribution combined with the Wilson – Hilferty transformation (Wilson and Hilferty 1931) was employed to generate annual flows. The use of the Wilson – Hilferty transformation was chosen to preserve the skewness of the historical streamflows that is often retained after the log transformation of the data.

The generation scheme is described below:

i) Take the logarithm of the observed annual flows, $y_i = ln\, x_i$;

ii) Calculate the mean ($\overline{y}$), standard deviation (s_y) and coefficient of skewness (g_y) of the transformed time series $y_{i*;}$

iii) Generate random numbers (t_{i*}) with Normal distribution $\mathrm{N}(0,1)$;

iv) Apply the Wilson-Hilferty transformation on t_{i*},

$$v_{i*} = \frac{2}{g_y}\left(1 + \frac{g_y}{6}t_{i*} - \frac{g_y^2}{36}\right)^3 - \frac{2}{g_y}, \quad (1)$$

where v_{i*} is a random variable with zero mean, unit variance and skewness g_y;

v) Obtain synthetic values of annual flows ($\hat{x}_{i*}$) via,

$$\hat{x}_{i*} = exp\left(\overline{y} + s_y v_{i*}\right). \quad (2)$$

With this generation scheme a set of $s = 50,...,3000$ (step *50*) was generated and the results show that the historical statistics are well preserved in all the generated time series. The quality of the generated series was evaluated by the comparison between historical statistic θ and the statistic of the generated times series $\hat{\theta}$. The mean $\mathrm{M}(\hat{\theta})$ and the standard deviation $\mathrm{S}(\hat{\theta})$ of $\hat{\theta}$ was calculated and then the confidence intervals $(1-\alpha)$ for θ was established by $\left]\mathrm{M}(\hat{\theta}) - z_{1-\alpha/2}\mathrm{s}(\hat{\theta}); \mathrm{M}(\hat{\theta}) + z_{1-\alpha/2}\mathrm{s}(\hat{\theta})\right[$. The historical statistic θ must be contained in the evaluated interval if that statistic is preserved by the model at the given level of confidence. As an example, we present in Table 3 the comparison between the statistics of the historical time series with the statistics of the 50, 1000, 2000 and 3000 generated time series. In this table is possible to see that all the historical statistics lies inside on the confidence interval for the 95% confidence level, meaning that the historical statistics are well preserved.

Monthly flows generation

To generate monthly flows the fragment method developed by Svanidze in 1961 was used. This method allows the simulation of monthly flows time series which accounts for the within-the-year runoff distribution and the stochastic dependence between the runoff values for individual months (Svanidze 1980). In this method, the observed monthly flows are first standardized year by year by dividing the monthly flows in a certain year by the corresponding average annual flow volume. The resulting set of standardized monthly flows in each year is referred as a fragment. After multiplying the mean annual volume, obtained in annual flow generation, by the fragment, the monthly distribution is obtained. This method has two major advantages in the generation of monthly flows. The first is that it implicitly preserves the skewness of the monthly flows because it takes into account the runoff monthly distribution, whatever it is. The second advantage is the way it can deal with zero flows, which, in climates like the one in Portugal, is very important. The fragments are obtained by dividing the observed monthly flows in a year by the corresponding average annual flow volume and if in a given month of the year the runoff is zero, the result of such division would be, obviously, also zero. To obtain the generated monthly flows, the mean annual volume is multiplied by the fragment and the result regarding the month with zero flow would be also zero. The method also preserves the mean and the variances and covariances of the monthly flows.

In order to apply the method of fragments the following methodology is proposed, having available a generated annual value ($\hat{x}_{i^*}$):

i) Sort the historical annual values (x_i) by ascendant order of magnitude;

ii) Determine the fragment for each value of the historical annual time series by,

$$F = \left[\frac{y_{i,j}}{\overline{x}_i}\right] = \begin{bmatrix} \frac{y_{1,1}}{\overline{x}_1} & \frac{y_{1,2}}{\overline{x}_1} & \cdots & \frac{y_{1,12}}{\overline{x}_1} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{y_{n,1}}{\overline{x}_n} & \frac{y_{n,2}}{\overline{x}_n} & \cdots & \frac{y_{n,12}}{\overline{x}_n} \end{bmatrix}, \quad (3)$$

where, **F** is an (*nx12*) matrix which contains the *n* fragments of the historical sequence. The indices *i* ($i = 1,2,...,n$) denotes the year and *j* (n) denotes the month. $y_{i,j}$ is the flow in month j in year i and $\overline{x}_i$ is the annual mean flow in year *i*, that is $\overline{x}_i = \frac{x_i}{12}$.

iii) Classify all the observed fragments according to the total runoff into r classes. For the first and last classes only the upper and lower limits are defined, respectively. The classification into classes is done by trial and the classes usually do not encompass the same range.

iv) Disaggregate each value of generated annual flow ($\hat{x}_{i^*}$) into monthly values by,

$$\hat{\mathbf{y}}_{i^*} = \hat{\bar{x}}_{i^*} \times \mathbf{F}_i \,, \tag{4}$$

Where $\hat{\mathbf{y}}_{i^*}$ is a vector that contains the generated monthly flows in year i^*, that is, $\hat{\mathbf{y}}_{i^*} = \left[\hat{y}_{i^*,1}, \hat{y}_{i^*,2}, ..., \hat{y}_{i^*,12}\right]$. $\hat{\bar{x}}_{i^*}$ is the generated annual mean flow obtained by $\hat{\bar{x}}_{i^*} = \hat{x}_{i^*}/12$ and $\mathbf{F_i}$ is the fragment to use, that is, a line of the matrix (3). The choice of $\mathbf{F_i}$ to use for the disaggregation of each generated annual value ($\hat{x}_{i^*}$) is as follow: identify the class in which $\hat{x}_{i^*}$ falls and if in that class more than one fragment is present then the fragment is drawn randomly without replacement. After all fragments in a given class are used, they are all replaced and can be used randomly again, if necessary.

Following this methodology the $s = 50,...,3000$ (step *50*) annual flows were disaggregated in monthly flows. As an example, Figures 3, 4, 5 and 6 show the comparison between the statistics of the historical time series and the statistics of the *3000* generated time series where it can be seen that the historical statistics are well preserved. As these figures demonstrate, the historical monthly mean, monthly standard deviation, monthly coefficient of skewness and the monthly lag-one correlation lies down on the established confidence intervals showing that those statistics are well reproduced by the model.

Storage capacity

The storage capacity associated with a specific reliability is found by the sequent peak algorithm (Thomas and Burden 1963) in conjunction with monthly synthetic streamflow time series. The following methodology was applied:

i) Calculate the residual mass curve ($Z_{i^*,j}$) by,

$$Z_{i^*,j} = \sum_{i^*=1}^{n^*} \sum_{j=1}^{12} \left(\hat{y}_{i^*,j} - q\right), \tag{5}$$

where n^* is the number of years of the generated time series, $\hat{y}_{i^*,j}$ is the generated flow in month j in year i^*, and q is the demand;

ii) Identify the first peak (M_1) in the curve;

iii) Identify the sequent peak (M_2) which is the next peak of grater magnitude than the first ($M_2 > M_1$);

iv) Identify the minimum (m_1) located between the two peaks;

v) Calculate the difference $D_1 = M_1 - m_1$;

vi) Repeat the steps ii to v until all peaks have been found;

vii) Calculate the storage capacity $c = max(D_p)$ with $p = 1,2,...,np$ where np is the number of peaks;

This storage capacity (c) is the minimum storage capacity necessary to meet the demand (q), without failure, for all the extension (n^*) of the generated time series ($\hat{y}_{i^*,j}$). If s generated time series are used s values of storage capacity ($c_m^s,\ m = 1,2,...,s$) are obtained. Then, the Gumbel distribution (Gumbel 1958) was fitted to each set of s values of storage capacity which allows the determination of the storage capacity c^s associated with a given reliability $F\left(c^s\right)$ by the inverse of Gumbel distribution,

$$c^s = \beta^s - \frac{1}{\alpha^s} ln\left\{-ln\left[F\left(c^s\right)\right]\right\}, \tag{6}$$

where α^s and β^s are the Gumbel distribution parameters estimates by the method of moments with s values of storage capacity.

This methodology was applied to find the values of storage capacity ($c^{50}, c^{100}, c^{150}, ..., c^{3000}$) for various levels of reliability (*80, 85, 90, 95, 98 e 99%*) to meet the monthly demand.

Number of time series to generate

Figure 7 shows the graphic representation of the $c^{50}, c^{100}, c^{150}, ..., c^{3000}$ storage capacity as function of the $s = 50,...,3000$ (step *50*) time series. The analysis of the figure reveals that for each reliability level, there is a great oscillation in the values of storage capacity determined with less than *300* generated time series. These oscillations are somewhat reduced when *300* to *900* generated time series are used and with more than *900* generated time series the oscillations decrease considerably.

In order to find an objective criterion to determine the number of time series to generate the following analysis was performed: If there is a number s that leads to a reasonably accurate value of

the storage capacity, then there is no increase in the accuracy in the calculation of the storage capacity if more than *s* time series were to be used. To find the number *s* we can try to identify the cut-point of the graphic in Figure 7. To do so, we analyzed the variance of the right-hand tail of the storage capacity distribution and drawn the correspondent graphic (Figure 8).

The analysis of Figure 8 indicates that the variance decreases as the number *s* increases and it tends asymptotically to a constant value with approximately s = *1200* approximately. That indicates that, for *s* = *1200* the variance of the estimate is already close enough to the asymptotic value and therefore we suggest that the number of time series to generate is *s* = *1200*.

Summary and Conclusions

The main goal of this study was to find the number (*s*) of time series of the same length of the historical series to generate for the calculation of the storage capacity needed to meet the demand with a specific reliability. For that purpose a methodology was presented and applied using five time series observed in four Portuguese rivers. The results obtained allow for the following conclusions:

i) The number (*s*) of time series to generate should be greater or equal to *1200*. This was the number found for all the cases studied, exemplified here in detail for one gauge station, namely Fragas da Torre. We believe that these cases are representative of Portuguese rivers and so we conclude that, without loss of generality, the number of time series to generate for sizing the storage capacity should be equal to *1200*.

Complementarily it was found that:

ii) The modeling of the skewness of the independent historic annual time series can be satisfactorily performed using log-transform of the annual data combined with the Wilson-Hilferty transformation. This can be explained by the fact that using the log-transformation alone seems to be insufficient to obtain a time series with null skewness which is required to apply the Normal distribution. However, when using the Wilson-Hilferty transformation after the log-transformation, the conditions to apply the Normal distribution are met.

iii) The disaggregation of the annual flow in monthly flows by the fragment method reveals to be quite good in preserving the monthly statistics. This method allows the modeling of monthly flows that are asymmetric because it takes into account the within-the-year runoff distribution. Furthermore this

method deals quite well with zero flows which, considering the climate conditions of Portugal is an especially important feature.

Acknowledgments

The writers wish to acknowledge the editors and associate editor of the journal and the four unknown reviewers for the insightful comments, remarks and suggestions that improved the paper.

References

Adeloye, A.J., Montaseri, M. and Garmann, C. (2001). Curing the misbehaviour of reservoir capacity statistics by controlling shortfall during failures using the modified sequent peak algorithm. *Water Resources Research*, 37(1), 73-82.

Anderson, R. L. (1941). Distribution of the serial correlation coefficients. *Annals of Math*. Statistics, 8(1), 1-13.

Burges, S.J. (1970). *Use of Stochastic Hydrology to Determine Storage Requirements of Reservoirs - A Critical Analysis*. Ph.D. Thesis, Department of Civil Engineering, Stanford University, Stanford.

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

Hoshi, K. and Burges, S.J. (1978). The impact of seasonal flow characteristics and demand patterns on required reservoir storage. *Journal of Hydrology*, 37(3-4), 241-260.

Kirby, W. (1974). Algebric boundedness of sample statistcs. *Water Resources Research*. 10(2), 220-222.

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

McMahon, T. A. (1978). *River Capacity and Yield*. Elsevier, Amsterdam.

McMahon, T.A. and Adeloye, A.J. (2005). *Water Resources Yield*. Water Resources Publications, LLC, Highlands Ranch, Colorado.

Phien, H.N. (1993). Reservoir storage capacity with gamma inflows. *Journal of Hydrology*, 146, 383-389.

Salas, J. D. (1993). Analysis and modeling of hydrologic time series. in: D.R. Maidment (Ed.), *Handbook of Hydrology*. McGraw - Hill, New York.

Svanidze , G. G. (1980). *Mathematical Modeling of Hydrologic Series*. Water Resources Publications, Fort Collins, Colorado.

Thomas, H. A. and Burden, R. P. (1963). *Operations Research in Water Quality Management*. Division of Applied Physics, Harvard University, Cambridge, Massachusets.

Vogel, R.M. and Stedinger, J.R. (1987). Generalized storage-reliability-yield relationships. *Journal of Hydrology*, 89, 303-327.

Vogel, R.M. and Stedinger, J.R. (1988). The value of stochastic streamflow models in overyear reservoir design applications. *Water Resources Research*, 24(9), 1483-1490.

Wallis, J.R. and Matalas, N.C. (1972). Sensitivity of reservoir design to the generating mechanism of inflows. *Water Resources Research*, 8(3), 634-641.

Wilson, E. B. and Hilferty, M. M. (1931). Distribution of Chi-square. *Procedings National Academy of Science*, 17, 648-688.

World Meteorological Organization (1988). *Analyzing Long Time Series of Hydrological Data With Respect to Climate Variability.* Wcap-3, WMO/TD 224.

Yue, S., Pilon, P. and Cavadias, G. (2002a). Power of the Mann-Kendall and Spearman's rho testes for detecting monotic trends in hydrological series. *Journal of Hydrology*, 259, 254-271.

Yue, S., Pilon, P., Phinney, B. and Cavadias, G. (2002b). The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrologic Processes*, 16, 1807-1829.

Yue, S. and Wang, C.Y. (2002). The influence of serial correlation on the Mann-Whiyney test for detecting a shift in median. *Advances in Water Resources*, 25, 325-333.