A STOCHASTIC MODEL FOR DAILY PRECIPITATION

BY
J. KELMAN[1]

SYNOPSIS -- A general model for description and sample generation of daily precipitation is presented. The basic assumption is that precipitation is a result of censoring a non-intermittent continuous-value process. Classical techniques for modeling the persistence in this latter process can then be applied. The continuous-value process admits an immediate extension to multivariate cases. The model was tested with series of several gauging stations in USA. Results have been found satisfactory.

INTRODUCTION

Stochasticaly generated rainfall sequences may be used for the design and operation of several water resources systems. For example, these sequences may be routed through some deterministic model of the hydrologic cycle, yielding in this way a synthetic streamflow series. It is conceivable that due to the better quality and quantity of rainfall data one may choose the uncertainty in the transfer function, rather than generating new sequences of streamflow from the unreliable historic records. These are often non-homogeneous due to man-made structures, while climate is in general stationary. Furthermore, generated rainfall sequences may be important by themselves, and not merely to be used to produce streamflow sequences, as would be the case in irrigation and drainage studies. In the ensuing text a rainfall model will be briefly described and then its application to some rainfall series will be shown in some detail. Meteorological factors related to the precipitation process, for example cloud type, temperature, winds, humidity, etc, are not considered. The observed record is examined merely as a realization of a stochastic process. No physical explanation of precipitation ocurrence can be derived from the statistical description of the observations presented herein.

THE MODEL

Let us assume that a stochastic process follows a first-order autoregressive model. Furthermore, let us admit that the marginal distribution is normal, namely

$$Z_t = \mu + \rho(Z_{t-1}-\mu) + \sigma\sqrt{1-\mu^2}\ \xi_t \qquad (1)$$

where $\xi_t \sim N(0,1)$, and $Z_t \sim N(\mu,\sigma^2)$.

Obviously, the Z_t-process is far from resembling an intermittent record such as daily rainfall. Therefore, some filtering is necessary, at least to eliminate the negative values of Z_t.

Define a Y_t-process as:

---

[1] Researcher, Centro de Pesquisas de Energia Elétrica, Rio de Janeiro, Brazil
Associate Professor, COPPE-UFRJ, Rio de Janeiro, Brazil

$$Y_t = Z_t, \text{ if } Z_t > 0$$

$$Y_t = 0, \text{ if } Z_t \leq 0 \qquad (2)$$

A realization of the Y_t-process can be considered as a censored sample of Z_t. A censored sample is such sequence for wich the values of the process that fall in a specific interval are not known. For example, all zero values in a realization of the Y_t-process represent negative but unknown observations of Z_t. In this case the censoring interval is $(-\infty,0)$. For this example, the resulting sample would be truncated, if the negative values of Z_t were not censored but also deleted from the record. In this case even the number of negative outcomes would not be known.

It is clear that Y_t is an intermittent process, provided with a mechannism of persistence. It remains to be seen whether this mechanism is appropriate in modeling and whether the marginal distribution of the positive observations obtained through the Y_t model, namely $P(Y_t<y|Y_t>0)$ fits the sample distribution well. In fact this last condition is not satisfied, because quite often the marginal distributions, in case the positive observations of the process are only studied, are characterized by a high skewness (higher than the one obtained by the truncated normal). Incidently, the truncated normal is the name given to the cumulative distribution function (c.d.f.)

$$P(Y<y) = \frac{\Phi[(y-\mu)/\sigma]}{\Phi[\mu/\sigma]} I_{(0,\infty)}(y) \qquad (3)$$

where $\Phi(\cdot)$ is the c.d.f. for the standard normal distribution. The positive values of Y_t might then be considered as a sample of this truncated normal distribution.

An examination of a typical case will help to explain why Y_t is not sufficient to represent the precipitation process. The histogram of the positive observations of daily rainfall at Austin for 70 years during the period May 1-June 1 is plotted in figure 1. For comparison the probability density functions (p.d.f.) which correspond to the truncated normal, and to the exponential distributions are also plotted in figure 1. The exponential distribution is included because it is often used to model the precipitation. The p.d.f. of the exponential distribution is:

$$f_X(x) = \psi e^{-\psi x} I_{(0,\infty)}(x) \qquad (4)$$

The parameter ψ is routinely estimated as the inverse of the arithmetic mean of the positive observations. For the Austin example $\hat{\psi} = 1.898$. The p.d.f. of the truncated normal distribution is

$$f_X(x) = \frac{1}{\Phi(\frac{\mu}{\sigma})\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2}(\frac{x-\mu}{\sigma})\} \; I_{(o,\infty)}(x) \qquad (5)$$

The parameters μ and σ are in principle estimated following

the procedure proposed by Cohen (1959). However, Cohen was mostly concerned with cases in which the number of censored elements is small compared with the total number of observations. In precipitation data there is a large number of zeros (censored observations). It turns out that graphs and tables supplied by Cohen are not sufficiently complete to handle this situation. Alternatively, an estimation procedure presented elsewhere by this writer (1976) was employed yieding the estimates $\hat{\mu}=0.627$ and $\hat{\sigma}=0.951$. The exponential one-parameter distribution was fitted only to positive observations, while the two-parameter truncated normal was fitted to the censored sample, in which the number of zeros was important. Since the probability of a zero outcome depends on the ratio μ/σ, it can be said that both distributions, exponential and truncated normal, had one degree-of-freedom to fit the data.

The inspection of figure 1 leads to the conclusion thar none of the two distributions produces a good fit. The form of the histogram suggest that a better fit could be obtained by using a p.d.f. which is asymptotic to the vertical axis.

Suppose that the Y_t-process is filtered according to

$$X_t = Y_t^{1/\alpha}, \tag{6}$$

with α = a real number. In this case the marginal distribution of positive observations of the X_t-process is the power-transformed truncated normal distribution (p.t.t.n., for short), namely

$$F_X(x) = \frac{\alpha x^{\alpha-1}}{\Phi(\mu/\sigma)\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(\frac{x^\alpha-\mu}{\sigma})\} I_{(o,\infty)}(x) \tag{7}$$

Notice that when $\alpha < 1$, $\lim_{x\to 0} f_X(x) = \infty$. For the Austin rainfall example, the estimate is $\hat{\alpha} = 0.595$. The corresponding p.d.f. is plotted in figure 1. From visual inspection, without any test, it is apparent that the p.t.t.n. does fit better the frequency histogram than the other two p.d.f.

The estimation procedure required to find out, in each case, the values of $\hat{\mu}$, $\hat{\sigma}$, $\hat{\rho}$ and $\hat{\alpha}$ will not be explained here. The reader is referred to the above mentioned paper.

DATA SELECTION

Choosing a particular precipitation record to be one of the cases studied here has been conditioned by the two requirements:

(i) The climatological description of the station location should be easily available; and

(ii) The stations should be sufficiently apart to possess different climatological conditions. However, a few stations should be sufficiently close in order to display some dependence, in this way serving as an illustration for the multivariate case for which the model is also applicable.

The first requirement was satisfied by imposing that a station would only qualify if it had been selected to receive a detailed description in WIC (1974). This publication gives a

summary of climatological data of a large number of precipitation stations in USA, furnished on a state by state basis. Among those, only a few are chosen to receive a complete description. The stations herein selected for study belong to this second category. They are given in Table 1.

Table 1. List of Stations Used for the Study

Station	Period of Record	Location LATIT.	Location LONG.	Elevation (ft.)	Average Precip. (in.)	Annual Days w/ Precip
Columbia (MO)	1951-1968	38°58'	92°22'	778	33.66	107
Kansas City (MO)	1946-1968	39°07'	94°36'	742	33.04	98
Springfield (MO)	1946-1968	37°14'	93°23'	1268	38.46	106
Raleigh-Durham (NC)	1951-1971	35°52'	78°47'	434	41.35	113
Austin (TX)	1898-1967	30°18'	97°42'	597	33.02	81
Rapid City (SD)	1951-1968	44°02'	103°03'	3165	16.39	93
Flagstaff (AZ)	1953-1970	35°08'	111°40'	6993	19.82	72
Seattle-Tacoma (WA)	1950-1970	47°27'	122°18'	386	39.95	164

The periods of record given in Table 1 were selected on the basis of the availabitily of data. They do not necessarily coincide with the periods in the WIC (1974) publication. Figure 2, with the locations of eight stations shows that the second requirement is also satisfied, namely that the stations are scattered throughout USA, with the exception of the three stations located in the State if Missouri, used to illustrate the multivariate case.

THE UNIVARIATE CASE

A possible application of the model may be in generating the new samples related to a specific short interval of time during the year, say a particular month. For this case one is tempted to assume the stationarity in the data. To study the applicability of the model for this case the data of each station series is divided im twelve seasons. The seasons have alternating lenghts of 32 and 28 days, adding up to a total of 360 days. As an example the estimates of μ, σ, ρ and α for the Columbia Station are given in table 2.

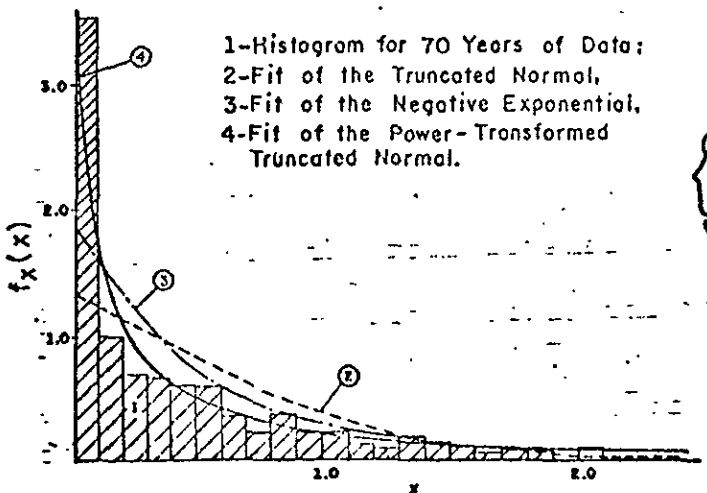


Fig.1 FITTING p.d.f. TO SAMPLE DATA (Austin; June)

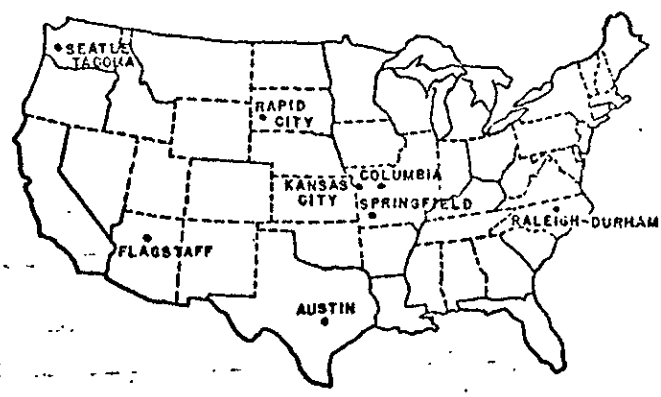


Fig. 2 LOCATION OF STATIONS

Table 2. Estimates of parameters for the Columbia Station

	Parameters			
	μ	σ	ρ	α
1	-0.4109	0.5475	0.3848	0.6121
2	-0.3291	0.5370	0.3584	0.6655
3	-0.2170	0.5383	0.1928	0.6249
4	-0.1947	0.5578	0.2295	0.7106
5	-0.3110	0.7080	0.3169	0.7143
6	-0.2939	0.7182	0.1900	0.6052
7	-0.3846	0.7701	0.2641	0.6353
8	-0.5526	0.7812	0.2158	0.6304
9	-0.6331	0.9114	0.3958	0.6065
10	-0.6799	0.8538	0.3706	0.6254
11	-0.5193	0.6632	0.2948	0.6576
12	-0.3301	0.5219	0.3451	0.6521

GOODNESS OF FIT

The chi-square goodness of fit statistic was evaluated for all the 8 x 12 = 96 "stations-seasons". The results are displayed in table 3, which is self-explanatory. An examination of this table shows that the good performance of the model with respect to reproducing the marginal distribution for each season-station is remarkable. Indeed out of 96 cases, only 13 had the hypothesis of correct fit rejected at the 5 percent significance level. At the 1 percent significance level only four cases. No null hypothesis stating the universality of the model applications is tested here. If this was the case, and if the studied time series were spatially and serially uncorrelated, then one would expected to have no more than 5 season-stations rejected at a 5 percent level or no more than 1 at a 1 percent level. The purpose of this particular study is rather to identify cases for which it is not advisable to apply the model. For instance, the four seasons that roughly span from December to March for Austin station should not be modeled by this approach.

TEST OF SERIAL INDEPENDENCE

One might wonder whether the model assumed for the continuous process, namely the first-order-Markov, may be excessively sophisticated for the problem at hand. This can be put in another way, whether it is possible that the continuous process is in fact serially independent, therefore with $\rho = 0$. If this is the case, any positive value estimated for ρ would be due to sample fluctuations. Hence, a test of the null hypothesis that $\rho = 0$ may be appropriate. Let $\bar{\Theta}$ be the four dimensional parameter space, namely $\bar{\Theta} = \{(\mu,\sigma,\rho,\alpha);\ -\infty < \mu < \infty,\ 0 < \sigma,\ 0 \leq \rho \leq 1,\ -\infty < \alpha < \infty\}$. Let us define the three-dimensional parameter subspace by $\bar{\Theta}_0 = (\mu,\sigma,\rho,\alpha);\ -\infty < \mu < \infty,\ 0 < \sigma,\ \rho = 0,\ -\infty < \alpha < \infty\}$. We want to test the null hypothesys H_0: $\Theta = (\mu,\sigma,\rho,\alpha) \in \bar{\Theta}_0$ versus the alternative hypothesis H_A: $\Theta = (\mu,\sigma,\rho,\alpha) \in \bar{\Theta} - \bar{\Theta}_0$. The generalized likelihood-ratio, denoted by λ is defined to be

Table 3. Chi-square goodness of fit statistic

Season \ Station	Columbia	Kansas	Spring-field	Raleigh	Austin	Rapid	Flagstaff	Seattle
1	8.979 (4) A	10.670 (5) A	5.035 (6) A	8.118 (9) A	34.283 (11) C	1.033 (1) A	5.492 (5) A	9.960 (10) A
2	7.015 (5) A	10.465 (5) A	10.794 (7) A	24.353 (9) C	27.749 (13) B	2.131 (1) A	3.735 (5) A	11.589 (8) A
3	10.857 (6) A	21.586 (9) B	16.871 (8) B	6.417 (8) A	23.074 (13) B	4.347 (2) A	6.226 (5) A	4.708 (6) A
4	10.427 (7) A	10.649 (8) A	6.327 (9) A	11.755 (7) A	21.488 (18) A	5.330 (4) A	3.546 (4) A	8.122 (5) A
5	16.416 (9) A	11.040 (11) A	17.041 (12) A	6.529 (9) A	26.883 (20) A	5.591 (6) A	0.179 (1) A	7.048 (4) A
6	6.570 (9) A	16.252 (12) A	15.006 (11) A	4.845 (9) A	26.069 (16) A	4.235 (7) A	3.165 (2) A	4.547 (3) A
7	7.825 (9) A	8.247 (12) A	13.923 (10) A	17.880 (12) A	16.346 (15) A	7.339 (5) A	5.409 (6) A	2.147 (2) A
8	7.823 (7) A	6.026 (10) A	5.182 (8) A	17.650 (10) A	8.120 (12) A	3.693 (3) A	12.934 (6) B	2.505 (4) A
9	5.637 (9) A	17.592 (11) A	12.114 (10) A	17.382 (9) A	25.744 (18) A	3.249 (4) A	14.631 (6) B	1.397 (5) A
10	6.093 (7) A	10.477 (8) A	2.574 (8) A	6.383 (8) A	27.779 (17) A	2.984 (1) A	3.551 (3) A	9.338 (8) A
11	10.577 (5) A	11.950 (6) A	13.029 (9) A	19.083 (8) B	11.971 (14) A	7.061 (2) B	5.143 (5) A	10.982 (10) A
12	4.587 (4) A	6.931 (5) A	15.226 (7) B	4.237 (7) A	32.065 (14) C	5.039 (2) A	6.445 (6) A	25.928 (10) C

Observation: For the Columbia Station, season 1, the chi-square statistic is 8.979 with 4 degress of freedom. Therefore the test could not be rejected at 5% significance level, classification: A. When the test is rejected at 5% significance level, classification: B. When the test is rejected at 1% significance level, classification: C

$$\lambda = \frac{\sup_{\theta \in \bar{\Theta}_o} L}{\sup_{\theta \in \bar{\Theta}} L} \qquad (8)$$

with sup (·) meaning the supremum and L the likelihood function

Notice that λ is a function only of the observations and therefore is a statistic. When the observations are replaced by their corresponding random variables then λ is itself a random variable. It is known, for example from Mood et al. (1974)., that for large sample $-2 \log \lambda$ is approximately distributed as chi-square with one degree of freedom, for this particular case.

Defining LL as log L, we have, from Eq. (8)

$$2\left[\sup_{\theta \in \bar{\Theta}} LL - \sup_{\theta \in \bar{\Theta}_o} LL\right] \approx \chi^2(1) \qquad (9)$$

Let

$$LL^* = \sup_{\theta \in \bar{\Theta}} LL \qquad (10)$$

Therefore, one should reject the null hypothesis, for the size of the test equal to γ, if

$$2(LL^* - \sup_{\theta \in \bar{\Theta}_o} LL) > \chi^2_{1-\gamma}(1), \qquad (11)$$

where γ is the probability that a wrong decision is reached, if the null hypothesis is rejected (Type I error). For $\gamma = 0.05$ one can reject the hypothesis whenever the test statistic takes a value greater than 3.84. Table 4 gives the values obtained for all the stations. It can be seen that for all but two of the 96 vases the null hypothesis were rejected. The only exceptions ocurred for the 12th season of Rapid City station, where $\chi^2 = 3.81$, and 4th season of Flagstaff, with $\chi^2 = 3.79$. These two cases may be results of pure chance variations.

This overwhelming rejection of the hypothesis of serial independence in the analysed precipitation series makes one wonder about the reality of several models, described in the literature, that neglect the time dependence in daily precipitation.

Table 4. Test of Serial Independence
Critical Value =3.84

Season \ Station	Columbia	Kansas	Spring-field	Raleigh	Austin	Rapid	Flagstaff	Seattle
1	17.305	14.423	24.474	13.211	93.511	13.310	60.055	62.657
2	15.023	12.583	18.955	7.282	65.116	17.827	35.279	40.569
3	5.633	29.020	14.343	18.498	52.826	22.344	37.810	54.015
4	7.316	17.755	13.854	8.584	64.543	20.015	3.791	31.782
5	15.243	7.991	12.660	7.011	70.383	22.304	29.906	17.715
6	4.521	6.536	13.647	13.796	84.268	9.858	25.922	18.283
7	9.650	11.112	12.340	17.017	87.787	8.339	20.816	30.143
8	4.374	20.713	7.324	11.246	48.091	4.247	7.442	41.227
9	19.562	15.034	26.198	26.576	90.777	15.584	30.172	59.112
10	13.139	23.827	6.481	22.644	71.480	12.595	19.096	42.042
11	8.961	9.310	21.736	16.201	99.947	10.171	27.546	37.029
12	13.269	25.057	28.613	12.551	97.713	3.808	35.404	30.662

A MULTIVARIATE APPLICATION

A simple illustration is given here to show the use of the model in a multivariate case. Suppose that one wants to produce the new samples of precipitation data for the station of Columbia, Kansas City, and Springfield simultaneously by preserving the areal dependence among them. Assume further that only the most rainy month for the region is of interest. This is June, roughly corresponding to the 6th season. First one must find the values of μ, σ, ρ, and α for each station. Next step is to find each of the three lag-zero cross correlation coefficients between station series. This can be accomplished solving equation(12) which is due to Rosenbaum (1961).

$$-\left[\frac{\hat{\mu}(j)}{\hat{\sigma}(j)}+\frac{\hat{\mu}(k)}{\hat{\sigma}(k)}\right]\hat{\rho}^2(j,k)$$

$$+\left\{\left[\frac{\hat{\mu}(j)}{\hat{\sigma}(j)}+\frac{\hat{\mu}(k)}{\hat{\sigma}(k)}\right]\hat{m}(j,k)+\frac{\hat{\mu}(j)\hat{\mu}(k)}{\hat{\sigma}(j)\hat{\sigma}(k)}\left[\hat{m}_1(j)+\hat{m}_1(k)\right]\right\}\hat{\rho}(j,k)$$

$$+\left[\frac{\hat{\mu}(j)}{\hat{\sigma}(j)}+\frac{\hat{\mu}(k)}{\hat{\sigma}(k)}\right]-\frac{\hat{\mu}(j)\hat{\mu}(k)}{\hat{\sigma}(j)\hat{\sigma}(k)}\left[\hat{m}_1(j)+\hat{m}_1(k)\right]-\frac{\hat{\mu}(j)}{\hat{\sigma}(j)}\hat{m}_2(j)-\frac{\hat{\mu}(k)}{\hat{\sigma}(k)}\hat{m}_2(k)=0 \qquad (12)$$

where

$$\hat{m}_1(j)=\frac{\sum^{n} x^{\hat{\alpha}(j)}-\hat{\mu}(j)n}{n\ \hat{\sigma}(j)},\quad \hat{m}_2(j)=\sum^{n}\left[\frac{\left[x^{\hat{\alpha}(j)}-\hat{\mu}(j)\right]^2}{n\ \hat{\sigma}^2(j)}\right],$$

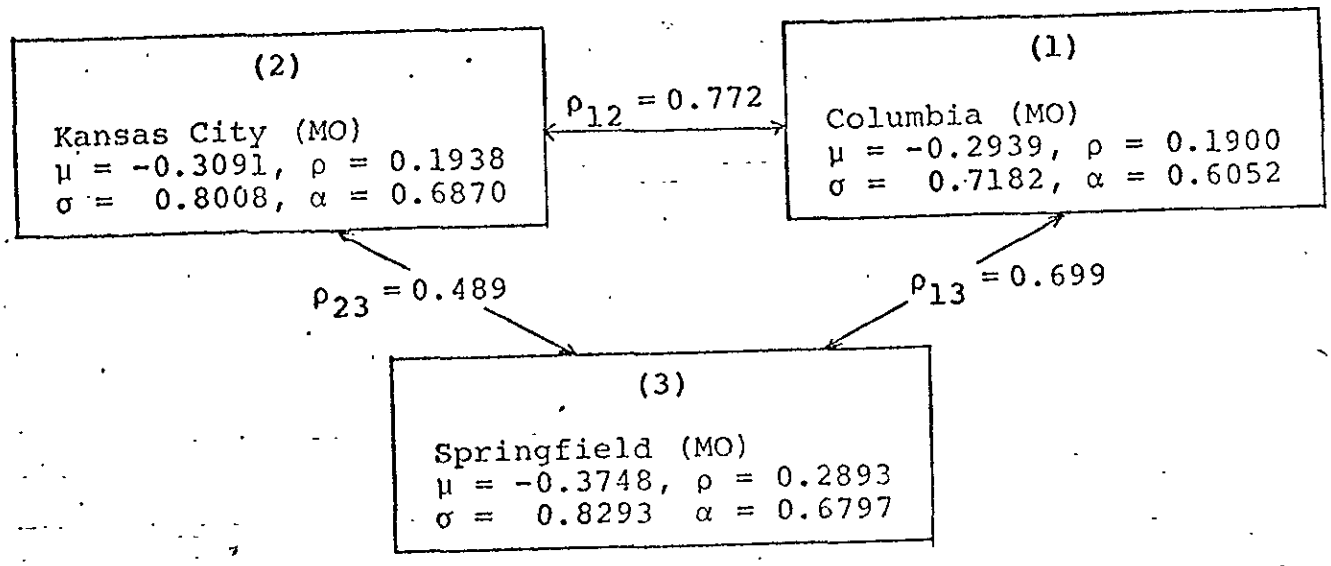
$$\hat{m}_1(k)=\frac{\sum^{n} y^{\hat{\alpha}(k)}-\hat{\mu}(k)n_2}{n\ \hat{\sigma}(k)},\quad \hat{m}_2(k)=\sum^{n}\left[\frac{\left[y^{\hat{\alpha}(k)}-\hat{\mu}(k)\right]^2}{n\ \hat{\sigma}^2(k)}\right]$$

$$\hat{m}(j,k) = \sum^{n} \left[\frac{(x^{\hat{\alpha}(j)} - \hat{\mu}(j))(y^{\hat{\alpha}(k)} - \hat{\mu}(k))}{n_2 \hat{\sigma}(j)\hat{\sigma}(k)} \right]$$

and $\hat{\rho}(j,k)$ is the only unknown. It is emphasized that the expression of Eq.(12) is to be used for the data of days with non-zero observations occur in both stations under consideration. All the remaining information is neglected. Because of the sample variation Eq.(12) may not have the real roots.

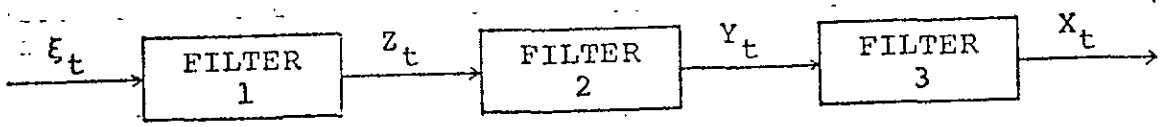
The results are summarized in Figure 3.

Figure 3. Representation of Parameters Needed for Generation of Daily Precipitation Series for The Month of June.



Once the parameters are estimed, the generation of new samples can be accomplished by following the stepwise procedure illustrated in Figure 4.

Figure 4. Representation of the Intermittent Model



In the multivariate case some care must be paid in generating $\xi_{t,j}$; $j = 1,2....,\ell$ because the variables may not be independent, with j as the station subscript. A way of doing this is by the use of:

$$\underline{\xi}_t = \pi \underline{\eta}_t, \qquad (13)$$

in which π is a $\ell x \ell$ matrix and $\underline{\eta}_t$ is a $\ell x 1$ vector of independent standard normal deviations. Then

$$\text{cov}(\underline{\xi}_t) = \text{cov}(\pi \underline{\eta}_t) = \pi \, \text{cov}(\underline{\eta}_t) \, \pi' \qquad (14)$$

where cov (·) means the covariance matrix of the argument vector. But

$$\mathrm{cov}(\eta_t) = I_\ell, \tag{15}$$

where I_ℓ is the $\ell x \ell$ identity matrix . Then from Eqs (14) and (15)

$$\mathrm{Cov}\ (\xi_t) = \pi\pi' \tag{16}$$

On the other hand, the linear autoregressive equations for stations j and k are:

$$z_{t,j} = \mu(j) + \rho(j)(z_{t-1,j} - \mu(j)) + \sigma(j)\sqrt{1-\rho^2(j)}\ \xi_{t,j} \tag{17}$$

and

$$z_{t,j} = \mu(k) + \rho(k)(z_{t-1,k} - \mu(k)) + \sigma(k)\sqrt{1-\rho^2(k)}\ \xi_{t,k} \tag{18}$$

Multiplying Eqs. (17) and (18) and finding the expected values, then

$$\mathrm{cov}(\xi_{t,j};\xi_{t,k}) = \frac{\rho(j,k)(1-\rho(j)\rho(k))}{\sqrt{(1-\rho^2(j))(1-\rho^2(k))}} \tag{19}$$

where $\rho(j)$ and $\rho(k)$ are the serial correlation coefficients respectively for stations j and k, and $\rho(j,k)$ is the lag-zero-cross correlation between the two station series. From Eqs. (16) and (19) one may conclude that the (j,k)-element of matrix $\pi\pi'$, $j \neq k$, is given by Eq. (19). The diagonal elements are unities. Several methods are available for finding the matrix π when $\pi\pi'$ is known; Young (1968) gives a straightforward one.

One hundred trivariate samples each for the month of June, were generated simultaneously according to the procedure explained above. Out of many ways of comparing the historic and the generated series, it was decided to focus attention on the joint positive runs. A joint positive run is defined as a succession of days for which the precipitation is observed at all three stations, preceded and followed by days for which at least at one of stations no precipitation occurred. For each joint positive run, the two variables of interest are: (i) the length, defined by $L_2 - L_1 + 1$, and (ii) the joint run - sum, defined as

$$\sum_{i=1}^{3} \sum_{j=L_1}^{L_2} X_{i,j}$$

where X_{ij} is the amount of precipitation at the ith station in the jth day, L_1 = the first day of the joint positive run, and L_2 = the last day of the joint positive run. These two variables were selected with the solution of flood problems in mind. Table 5 gives the absolute frequencies of run-lengths for the historic and generated series.

Table 5. Absolute Frequency of the Joint Positive Run-Lengths

Sample	Run-length						Total
	1	2	3	4	5	6	
Historic	31	8	5	1	0	0	45
Generated	190	31	19	4	0	1	245

Whether the two samples of Table 5 can be considered as drawn from the same population is of crucial importance in the evaluation of the model. A way of answering this question is by using the test of equality of two multinominal distributions. The reader is referred to Mood et al. (1974) where a description of the test is given (pages 448-452). It is sufficient to state here that the sample space is divided in $k+1$ subsets and the null hypothesis states that Ho: $P_{ij} = P_{2j}$, $j = 1,2,\ldots,k+1$ where P_{1j} = the probability that an outcome of the first population belongs to the jth subset, and P_{2j} = the same as P_{1j} but in regard to the second population. For the above data the division in three subsets ($k = 2$) seems convenient, namely: (i) run of length 1; (ii) run of length 2; and (iii) run of length > 2.

It can be shown that

$$z = \sum_{i=1}^{2} \sum_{j=1}^{k+1} \frac{(G_{ij} - g_i(G_{ij} + G_{2j})/(g_1 + g_2))^2}{g_i(G_{1j} + G_{2j})/(g_1 + g_2)} \qquad (20)$$

has a limiting chi-square distribution with k degrees of freedom, where g_1 = the total number of observations for the first population (in the present case, 45); g_2 = the same as g_1, but for the second population (245); G_{1j} = number of outcomes in class j, from the first population; and G_{2j} = the same as G_{1j}, but from the second population.

The use of Eq. (20) with the data of Table 5 yields a value of $z = 1.58$. Since the 95 percent quantile of the chi-square distribution with two degrees of freedom is 5.99, the null hypothesis cannot be rejected at the 5 percent significance level.

With regard to the joint run-sums, again the test whether the two samples (not given in tables) were drawn from the same population if performed. Since this variable is continuous, the two-sample-Smirnov test seems more suitable than the multinomial one. For a description of that test see Bradley (1968). Here it is sufficient to state that under the null hypothesis of equality of the two distributions, the random variable

$$W = \max_x |S_1(x) - S_2(x)| \qquad (20')$$

has some distribution which the 95 percent quantile is given approximately by

$$1.358 \sqrt{\frac{g_1 + g_2}{g_1 g_2}} \qquad (21)$$

where $S_1(x)$ is the sample c.d.f. of the historic sample and $S_2(x)$ is its counterpart for the generated sample.

The application of Eqs (20) and (21) to data gives the values of 0.1868 and 0.2202, respectively. Therefore, the hypothesis stating that the two samples can be considered as drawn from the same population should be accepted at the 5 percent significance level.

It can be said that, the application of the model to the multivariate case is satisfactory for the example used. This is a positive indication about the feasibility of using the model.

ACKNOWLEDGMENTS

This study was developed while the writer was a Ph.D. candidate in Colorado State University, USA, with the financial support of the National Research Counsel of Brazil (CNPq). Thanks go to Dr. Vujica Yevjevich, Professor of Civil Engineering, and Dr. D.C. Boes, Associate Professor of Statistics, for their suggestions and guidance.

REFERENCES

Bradley, J., "Distribution Free Statistical Tests", Prentice-Hall Inc., 1968.

Cohen, A.C., Jr., "Simplified Estimators for the Normal Distribution When Samples are Singly Censored or Truncated", Technometrics, Vol 1, N0. 3, pp. 217-237, 1959.

Kelman, J., "A Simulation Model for Intermittent Processes" Proceedings of the Second International Symposium on Stochastic Hydraulics, Lund, Sweden, 1976.

Mood, A. M., Graybill, F. A., and Boes, D. C., "Introduction to the Theory of Statistics", McGraw-Hill, 1974.

Rosenbaum, S., "Moments of a Truncated Bivariate Normal Distribution", Journal of The Royal Statistical Society, Series B, 23, pp. 405-408, 1961.

WIC, "Climate of the States", Water Information Center Inc., Port Washington, New York, 1974.

Young, C.K., discussion on "Mathematical Assessment of Synthetic Hydrology", by N. C. Matalas, Water Resources Research, Vol. 4, No. 3, pp. 681-682, 1968.