Precipitation as a Chain-Dependent Process

RICHARD W. KATZ

National Center for Atmospheric Research,[1] *Boulder, Colo. 80307*

(Manuscript received 24 November 1976, in revised form 9 May 1977)

ABSTRACT

A probabilistic model for the sequence of daily amounts of precipitation is proposed. This model is a generalization of the commonly used Markov chain model for the occurrence of precipitation. Methods are given for computing the distribution of the maximum amount of daily precipitation and the distribution of the total amount of precipitation. The application of this model is illustrated by an example, using State College, Pennsylvania, precipitation data.

1. Introduction

The Markov chain model for the daily occurrence of precipitation (Gabriel and Neumann, 1962) has achieved widespread use. However, in the literature there are virtually no models for the daily amount of precipitation which employ this Markov chain model as part of the process. One exception is the model recently introduced by Todorovic and Woolhiser (1975), which is a special case of the model to be presented in this paper.

The probabilistic model for the daily amounts of precipitation which will be considered is itself a special case of a so-called *chain-dependent process* or sequence of random variables defined on a Markov chain [see Katz (1977) for a formal definition of this general process]. While the model does introduce dependence into the precipitation process through the dependence of the Markov chain, it is still simple enough to apply for computation purposes. For example, the distribution of the total amount of precipitation in n days ($n=1, 2, \ldots$) and the distribution of the maximum amount of daily precipitation in n days can be readily calculated. Also, limit theorems for general chain-dependent processes can be applied to obtain asymptotic distributions for the total and maximum amounts of precipitation. The theoretical results used in this paper are derived in a related article (Katz, 1977).

The application of this model is illustrated by an example, using State College, Pennsylvania, precipitation data. Methods are given for verifying that the assumptions of the model hold and for estimating the parameters of the model.

[1] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

2. Definition of process

The chain-dependent process for precipitation consists of a bivariate stochastic process $\{(J_{n-1},X_n): n=1, 2, \ldots\}$. Here

$$J_n=\begin{cases}1 & \text{if precipitation occurs on } n\text{th day}\\ 0 & \text{otherwise}\end{cases}$$

and it is assumed that $\{J_n: n=0, 1, \ldots\}$ constitutes a first-order two-state Markov chain. This assumption means that the J_n process satisfies the condition

$$\Pr[J_n=j\,|\,J_0, J_1, \ldots, J_{n-2}, J_{n-1}=i] = \Pr[J_n=j\,|\,J_{n-1}=i],$$

$i, j=0, 1, n=1, 2, \ldots$. The nth day is said to be a dry day if $J_n=0$, and a wet day if $J_n=1$. We define transition probabilities $P_{ij}=\Pr[J_n=j\,|\,J_{n-1}=i]$, $i, j=0, 1$, and initial probabilities $\lambda_i=\Pr[J_0=i]$, $i=0, 1$. The stationary probabilities

$$\pi_i=\lim_{n\to\infty}\Pr[J_n=i]$$

are given by

$$\pi_0=\frac{P_{10}}{P_{01}+P_{10}},\quad \pi_1=\frac{P_{01}}{P_{01}+P_{10}}.$$

Finally, observe that the Markov chain is completely characterized by the parameters λ_1, P_{01} and P_{11}.

The amount of precipitation which occurs on the nth day is denoted by X_n. We make the following assumptions about the process $\{X_n: n=1, 2, \ldots\}$:

(i) The distribution of X_n depends on J_{n-1}.

(ii) The X_n's are conditionally independent, given the J_n process.

These two conditions are understood in the sense that

$$\Pr[X_n \leqslant x | J_0, X_1, \ldots, J_{n-2}, X_{n-1}, J_{n-1}=i] = \Pr[X_n \leqslant x | J_{n-1}=i], \quad x \geqslant 0.$$

In particular, we define conditional distribution functions $F_i(x) = \Pr[X_n \leqslant x | J_{n-1}=i, J_n=1]$, $i=0, 1$. Assumption (i) means that, given a wet day, the amount of precipitation which occurs on that day is chosen from either distribution F_0 or F_1, depending on whether the previous day was dry or wet. Further, assumption (ii) implies that the amounts of precipitation on consecutive wet days are conditionally independent.

The process considered by Todorovic and Woolhiser (1975) can be obtained as a special case of this chain-dependent process by assuming that $F_0 = F_1$. They derive results by further assuming that this common distribution is exponential.

3. Maximum amount of precipitation

Consider the stochastic process $\{M_n: n=1, 2, \ldots\}$ denoting the maximum amount of daily precipitation in n days, i.e.,

$$M_n = \max_{1 \leqslant i \leqslant n} X_i.$$

Set

$$G_n(x) = \Pr[M_n \leqslant x]$$

and

$$G_n(x; i) = \Pr[M_n \leqslant x | J_0 = i],\ i=0, 1.$$

a. Exact distribution

Generalizing the approach given in Katz (1974a) for the Markov chain model for the occurrence of precipitation, a recurrence relation for computing the distribution of M_n can be derived. Conditioning on J_0 yields

$$G_n(x) = \lambda_0 G_n(x; 0) + \lambda_1 G_n(x; 1), \tag{1}$$

and further conditioning on J_1 yields

$$G_n(x; 0) = P_{00} G_{n-1}(x; 0) + P_{01} F_0(x) G_{n-1}(x; 1), \tag{2}$$

$$G_n(x; 1) = P_{10} G_{n-1}(x; 0) + P_{11} F_1(x) G_{n-1}(x; 1), \tag{3}$$

$n=1, 2, \ldots$, with initial conditions $G_0(x; 0) = G_0(x; 1) = 1$. The recurrence relations (2) and (3), along with (1), can be used to compute the distribution of M_n explicitly.

b. Asymptotic distribution

The Type I extreme value distribution is commonly used to approximate the distribution of the maximum amount of daily precipitation over long periods of time (see, for example, Gumbel, 1958). Consistent with this practice, it has been shown that the maximum of a chain-dependent process has one of the so-called extreme value distributions as its limiting distribution (Denzel and O'Brien, 1975; Katz, 1974b).

The gamma distribution is often used to fit amounts of daily precipitation (see, e.g., Neyman and Scott, 1967). Assume that F_0 and F_1 are both gamma distribution functions; i.e.,

$$F_i'(x) = \frac{1}{\beta_i \Gamma(\alpha_i)} \left(\frac{x}{\beta_i}\right)^{\alpha_i - 1} \exp\left(-\frac{x}{\beta_i}\right), \tag{4}$$

$x > 0$, α_i, $\beta_i > 0$, $i=0, 1$, where Γ denotes the gamma function. Then the maximum amount of daily precipitation M_n, suitably normalized, has the Type I extreme value distribution as its limiting distribution (Katz, 1977, Theorem 2).

Specifically, we define

$$F(x) = \pi_0 + \pi_1 P_{01} F_0(x) + \pi_1 P_{11} F_1(x), \quad x \geqslant 0. \tag{5}$$

The result can be stated as follows: if F_0 and F_1 are given by (4) and F by (5), then

$$\lim_{n \to \infty} G_n\left[u_n + \frac{x}{nF'(u_n)}\right] = \exp[-\exp(-x)], \tag{6}$$

where $1 - F(u_n) = 1/n$, $n \geqslant 2$.

4. Total amount of precipitation

Consider the stochastic process $\{S_n: n=1, 2, \ldots\}$ denoting the total amount of precipitation in n days, i.e.,

$$S_n = \sum_{i=1}^{n} X_i.$$

Set

$$H_n(x) = \Pr[S_n \leqslant x]$$

and

$$H_n(x; i) = \Pr[S_n \leqslant x | J_0 = i],\ i=0, 1.$$

a. Exact distribution

Again a recurrence relation for computing the distribution of S_n can be derived. Conditioning on J_0 yields

$$H_n(x) = \lambda_0 H_n(x; 0) + \lambda_1 H_n(x; 1), \tag{7}$$

and further conditioning on J_1 yields

$$H_n(x; 0) = P_{00} H_{n-1}(x; 0) + P_{01} F_0(x) * H_{n-1}(x; 1), \tag{8}$$

$$H_n(x; 1) = P_{10} H_{n-1}(x; 0) + P_{11} F_1(x) * H_{n-1}(x; 1), \tag{9}$$

$n=1, 2, \ldots$, with initial conditions $H_0(x; 0) = H_0(x; 1) = 1$, where the asterisk denotes the convolution operator. If F_0 and F_1 are distributions such that the convolutions in (8) and (9) can be readily evaluated, then the recurrence relations (8) and (9), along with (7), can be used to compute the distribution of S_n explicitly.

b. Asymptotic distribution

The normal distribution is commonly used to approximate the distribution of the total amount of precipitation over long periods of time. Consistent with this practice, it can be shown that the sum of a chain-dependent process is asymptotically normally distributed (O'Brien, 1974; Katz, 1974b).

Let μ_i and σ_i^2 denote the mean and variance, respectively, for the distribution function F_i $(i=0,1)$. Expressions for the asymptotic mean and variance (see Katz, 1977) are given by

$$\mu \equiv \lim_{n\to\infty} E(X_n) = \pi_0 P_{01}\mu_0 + \pi_1 P_{11}\mu_1, \tag{10}$$

$$\rho \equiv \lim_{n\to\infty} \mathrm{var}(X_n) = \pi_0 P_{01}(\mu_0^2+\sigma_0^2) + \pi_1 P_{11}(\mu_1^2+\sigma_1^2) - \mu^2.$$

Set

$$\sigma^2 \equiv \rho + \frac{2}{1-d}\mu\pi_0(P_{11}\mu_1 - P_{01}\mu_0), \tag{11}$$

where $d=P_{11}-P_{01}$. Then $(S_n-n\mu)/(n^{\frac{1}{2}}\sigma)$ is asymptotically normally distributed with zero mean and unit variance (Katz, 1977, Theorem 3). Note that this result is identical to the central limit theorem for the case of a sum of independent random variables, except for the modification of the expression for the variance normalizing factor.

5. Numerical example

State College, Pennsylvania daily precipitation data for 1–28 February 1930–69 will be used.

a. Markov chain

A wet day is defined to be one on which at least 0.01 inch of precipitation occurs. The usual relative frequency estimates for the Markov chain parameters are given in Table 1.

Several statistical tests have been performed to insure that the sequence of daily occurrences of precipitation does constitute a first-order Markov chain as defined in Section 2 (see Anderson and Goodman, 1957). The results of these tests are summarized in Table 1. The test for independence indicates overwhelming evidence that sequences of dry and wet days are dependent (observed significance level = 0.001). On the other hand, the test of first-order versus second-order Markov chain does not indicate that a higher than first order model is necessary (observed significance level = 0.134).

Table 1. Parameter estimates and statistical tests for Markov chain model. $\hat{P}_{01}=0.342$, $\hat{P}_{11}=0.441$, $\hat{\pi}_1=0.379$.

Test	Statistic	Observed significance level
1. Independence vs first-order Markov chain	$\chi^2=10.55$ (1 df)*	0.001
2. First-order vs second-order Markov chain	$\chi^2=4.02$ (2 df)*	0.134
3. Stationarity of transition probabilities (day to day)	$\chi^2=63.00$ (52 df)*	0.141
4. Stationarity of transition probabilities (year to year)	$\chi^2=88.84$ (78 df)*	0.189

* df = degrees of freedom.

It has been assumed that the transition probabilities do not vary from day to day. The test for day-to-day stationarity supports this assumption (observed significance level = 0.141). It has also been assumed that the transition probabilities do not vary from year to year. The test for year-to-year stationarity supports this assumption (observed significance level = 0.189).

b. Conditional distributions

Fig. 1 gives the sample histograms for the conditional distribution functions F_0 and F_1. For this example, there appears to be little difference between these two distributions, although there is some evidence of a

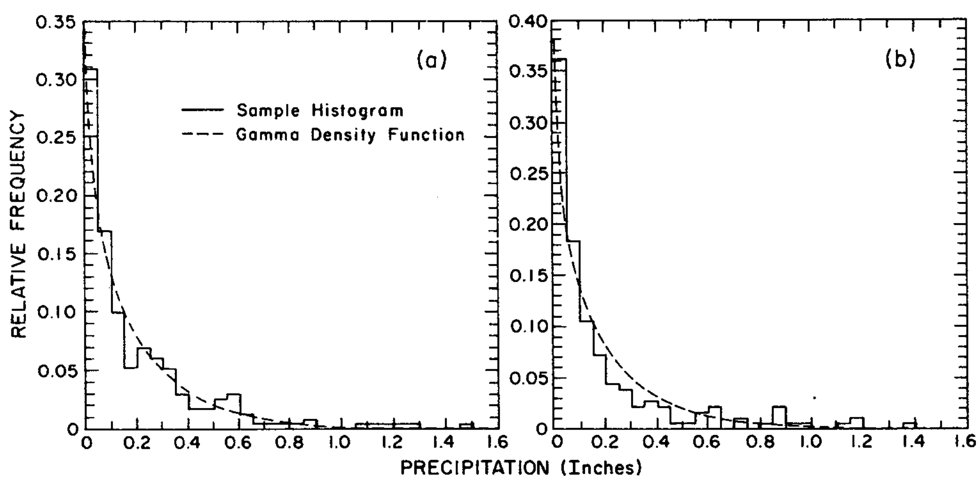

Fig. 1. Sample histograms and gamma density functions for the conditional distribution of daily precipitation given (a) dry preceding day and (b) wet preceding day.

TABLE 2. Approximate maximum likelihood estimates for parameters of gamma distributions used to fit conditional distributions of daily precipitation.

Distribution	Sample size	$\hat{\alpha}_i$	$\hat{\beta}_i$
Dry preceding day (F_0)	230	0.811	0.262 inch
Wet preceding day (F_1)	180	0.794	0.251 inch

higher frequency of small amounts of precipitation on wet days following wet days than on wet days following dry days.

The conditional distribution functions F_0 and F_1 have been fitted using the gamma distributions specified by (4). The method of Greenwood and Durand (Johnson and Kotz, 1970, p. 189) is used to obtain approximate maximum likelihood estimates for the shape parameters α_i and the scale parameters β_i $(i=0, 1)$. The parameter estimates are given in Table 2, and the two gamma density functions are shown in Fig. 1.

c. Conditional independence

It has been assumed that the amounts of precipitation on consecutive wet days are conditionally independent. As a test of this assumption, the bivariate log-normal distribution has been used to fit the amounts of precipitation on consecutive wet days for the State College data. The bivariate log-normal distribution is used, instead of the bivariate gamma distribution, because of the relative ease in obtaining the maximum likelihood estimates (Flueck and Mielke, 1975).

The bivariate log-normal density function is given by

$$f(x_1,x_2)=\frac{1}{2xy\alpha_1\alpha_2\pi(1-\phi^2)^{\frac{1}{2}}}\exp\left(-\frac{1}{2(1-\phi^2)}\times\left\{\left[\frac{\ln(x/\beta_1)}{\alpha_1}\right]^2-2\phi\left[\frac{\ln(x/\beta_1)}{\alpha_1}\right]\left[\frac{\ln(y/\beta_2)}{\alpha_2}\right]+\left[\frac{\ln(y/\beta_2)}{\alpha_2}\right]^2\right\}\right),$$

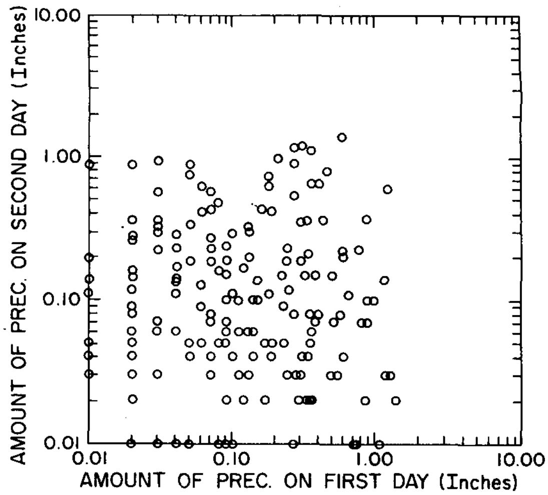

FIG. 2. Amounts of precipitation on consecutive wet days plotted on log-log scale.

where $x,y>0$, α_1, α_2, β_1, $\beta_2>0$ and $|\phi|<1$. Here α_1 and α_2 are the shape parameters and β_1 and β_2 are the scale parameters for the marginal log-normal distributions, and ϕ is the dependence parameter. Under the assumption of a bivariate log-normal distribution, the logarithmically transformed data have a bivariate normal distribution with the parameter ϕ being the correlation coefficient. Thus a value of ϕ near zero is indicative of a lack of dependence.

For the State College data, Table 3 gives the maximum likelihood estimates of the bivariate log-normal distribution parameters. Besides considering the amounts of precipitation on all consecutive wet days, the data have been stratified according to the days of the wet period (a run of consecutive wet days) on which the amounts occurred. Estimates are given for the first and second days of the wet period and the second and third days of the wet period, the only cases with adequate sample size. The dependence parameter estimates are all quite close to zero (observed significance level=0.770 for all consecutive wet days), indicating that there is a lack of significant evidence of dependence between amounts of precipitation on

TABLE 3. Maximum likelihood estimates for parameters of bivariate log-normal distribution fit to amounts of precipitation on consecutive wet days and significance test for dependence parameter.

	Sample size	$\hat{\alpha}_1$, $\hat{\alpha}_2$	$\hat{\beta}_1$, $\hat{\beta}_2$ (inch)	$\hat{\phi}$	Test statistic*	Observed significance level
All consecutive wet days	180	1.284, 1.263	0.109, 0.094	0.0219	t=0.2923 (178 df)**	0.770
First and second days of wet period	106	1.332, 1.272	0.120, 0.094	−0.0014	t=−0.0143 (104 df)**	0.989
Second and third days of wet period	39	1.249, 1.214	0.097, 0.101	0.0197	t=0.1199 (37 df)**	0.905

* Test for nonzero dependence parameter.
** df=degrees of freedom.

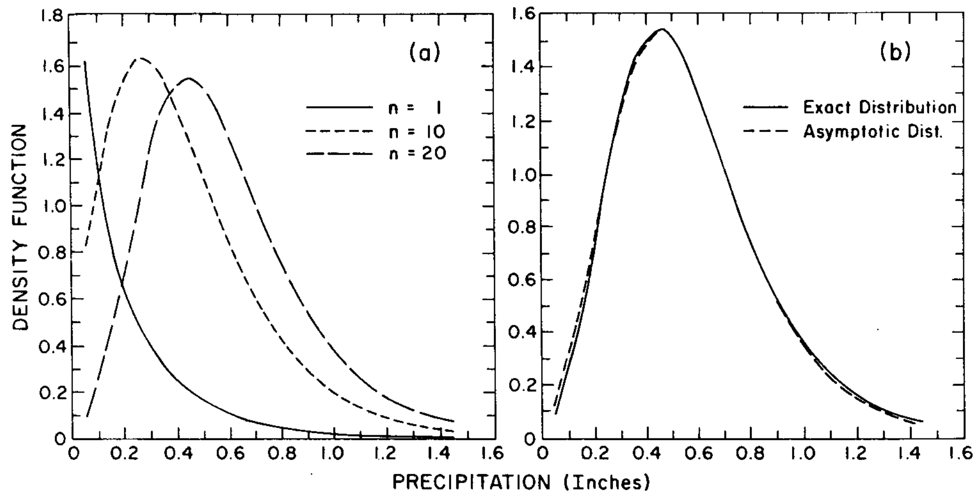

FIG. 3. Distribution of the maximum amount of daily precipitation in n days: (a) exact distribution for $n=1$, 10, 20 and (b) exact and asymptotic distribution for $n=20$.

consecutive wet days. In addition, the essentially random scatter of the data for all consecutive wet days plotted on a log-log scale (Fig. 2) is further evidence of a lack of dependence.

d. Maximum amount of precipitation

For the State College parameters (Tables 1 and 2), setting $\alpha_i=\pi_i$ $(i=0, 1)$, the exact distribution of the maximum amount of daily precipitation in n days has been computed using the recurrence relations (2) and (3), along with (1). Fig. 3a illustrates the resulting density functions for $M_n>0$ and selected values of n.

Applying (6), the asymptotic Type I extreme value distribution for $n=20$ has location parameter $u_{20}=0.434$ and scale parameter $1/[20F'(u_{20})]=0.235$. Fig. 3b gives both the exact and asymptotic density functions for $n=20$. It is evident that the extreme value approximation is quite accurate for this sample size.

c. Total amount of precipitation

The distribution of the total amount of precipitation has also been computed, again setting $\alpha_i=\pi_i$ $(i=0, 1)$, using the recurrence relations (8) and (9) along with (7). In this case, for ease in evaluating the convolutions, it has been assumed that F_0 and F_1 have a common scale parameter of 0.257 inch (estimated by pooling the data). Fig. 4a illustrates the resulting density functions for $S_n>0$ and selected values of n.

Using (10) and (11), the asymptotic normal distribution has parameters $\mu=0.078$ inch and $\sigma^2=0.032$ inch^2. Fig. 4b gives both the exact and asymptotic density functions for $n=20$. While the convergence of the exact

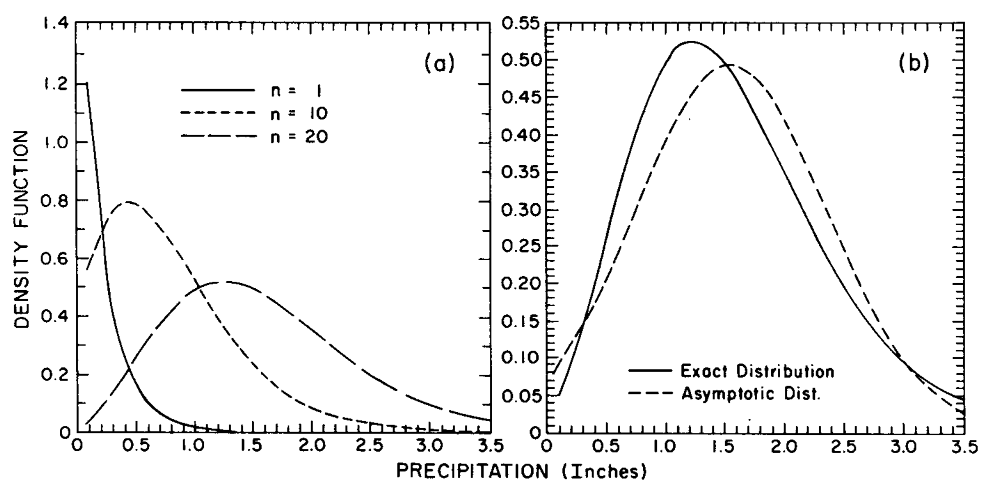

FIG. 4. Distribution of the total amount of precipitation in n days: (a) exact distribution for $n=1$, 10, 20 and (b) exact and asymptotic distribution for $n=20$.

distribution to the normal distribution as n increases is evident from Fig. 4a, Fig. 4b indicates that a sample size of 20 is not large enough for an accurate approximation.

6. Summary

A probabilistic model, representing the sequence of daily amounts of precipitation, has been considered. The key assumptions required for the application of this process are that 1) the sequence of daily amounts of precipitation constitutes a Markov chain, and 2) the amounts of precipitation on consecutive wet days are conditionally independent. This process might be used to characterize, in a climatological sense, a precipitation sequence in terms of a few parameters which take into account the day-to-day dependence of precipitation. Methods have been given which will enable a potential user of the model to check whether these assumptions hold and to estimate the parameters for a given set of precipitation data. These methods have been applied to State College, Pennsylvania, data, and the results indicate that the requirements of the model are satisfied for this particular example.

Acknowledgments. This research includes a portion of the author's Ph.D. dissertation and was conducted while he was supported by a National Science Foundation Traineeship at Pennsylvania State University. The author has also received support from the National Science Foundation as a postdoctoral fellow in the Advanced Study Program at the National Center for Atmospheric Research. Finally, he acknowledges the aid of his thesis adviser Frank Haight.

REFERENCES

Anderson, T. W., and L. A. Goodman, 1957: Statistical inference about Markov chains. *Ann. Math. Stat.*, **28**, 89–110.

Denzel, G. E., and G. L. O'Brien, 1975: Limit theorems for extreme values of chain-dependent processes. *Ann. Prob.*, **3**, 773–779.

Flueck, J. A., and P. W. Mielke, 1975: Some bivariate distributions having meteorological applications. *Preprints Fourth Conf. Probability and Statistics*, Tallahassee, Amer. Meteor. Soc., 65–69.

Gabriel, K. R., and J. Neumann, 1962: A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quart. J. Roy. Meteor. Soc.*, **88**, 90–95.

Gumbel, E. J., 1958: *Statistics of Extremes*. Columbia University Press, 375 pp.

Johnson, N. L. and S. Kotz, 1970: *Distributions in Statistics: Continuous Univariate Distributions*—1. Wiley, 300 pp.

Katz, R. W., 1974a: Computing probabilities associated with the Markov chain model for precipitation. *J. Appl. Meteor.*, **13**, 953–954.

——, 1974b: A stochastic process defined on a Markov chain: properties and an application to meteorology. Ph.D. dissertation, Pennsylvania State University, 113 pp.

——, 1977: An application of chain-dependent processes to meteorology. *J. Appl. Prob.* (in press).

Neyman, J., and E. L. Scott, 1967: Some outstanding problems relating to rain modification. *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 5, University of California Press, 293–326.

O'Brien, G. L., 1974: Limit theorems for sums of chain-dependent processes. *J. Appl. Prob.*, **11**, 582–587.

Todorovic, P., and D. A. Woolhiser, 1975: A stochastic model of n-day precipitation. *J. Appl. Meteor.*, **14**, 17–24.