Overdispersion Phenomenon in Stochastic Modeling of Precipitation

RICHARD W. KATZ

Environmental and Societal Impacts Group, National Center for Atmospheric Research, Boulder, Colorado*

MARC B. PARLANGE

Department of Geography and Environmental Engineering, The Johns Hopkins University, Baltimore, Maryland

(Manuscript received 18 October 1996, in final form 16 May 1997)

ABSTRACT

Simple stochastic models fit to time series of daily precipitation amount have a marked tendency to underestimate the observed (or interannual) variance of monthly (or seasonal) total precipitation. By considering extensions of one particular class of stochastic model known as a chain-dependent process, the extent to which this ''overdispersion'' phenomenon is attributable to an inadequate model for high-frequency variation of precipitation is examined. For daily precipitation amount in January at Chico, California, fitting more complex stochastic models greatly reduces the underestimation of the variance of monthly total precipitation. One source of overdispersion, the number of wet days, can be completely eliminated through the use of a higher-order Markov chain for daily precipitation occurrence. Nevertheless, some of the observed variance remains unexplained and could possibly be attributed to low-frequency variation (sometimes termed ''potential predictability''). Of special interest is the fact that these more complex stochastic models still underestimate the monthly variance, more so than does an alternative approach, in which the simplest form of chain-dependent process is conditioned on an index of large-scale atmospheric circulation.

1. Introduction

It is well known that when simple stochastic models are fitted to time series of daily precipitation amount, there is a marked tendency to underestimate the observed (or interannual) variance of monthly (or seasonal) total precipitation (Buishand 1978; Wilks 1989). In the statistics literature, this situation in which the observed variance exceeds that for the fitted model is termed ''overdispersion'' (e.g., Cox 1983). The explanation for the overdispersion phenomenon of precipitation, however, is not agreed upon. Some researchers view this discrepancy as evidence of an inadequate model for high-frequency variation of precipitation (Gregory et al. 1993). Others regard it as attributable to low-frequency variation that these models do not account for and, as such, constitutes a measure of the ''potential predictability'' of monthly total precipitation on an interannual timescale (Shea and Madden 1990; Shea et al. 1995; Singh and Kripalani 1986). But, if the first explanation were valid, then this approach would overestimate the degree of potential predictability.

In the present paper, we examine the first explanation, identifying and eliminating the various sources of variance underestimation for one particular class of stochastic model for high-frequency variation of precipitation, known as a chain-dependent process (Katz 1977a). This model involves dividing the precipitation process into two component models, one for its occurrence and another for its intensity (i.e., amount of precipitation conditional on its occurrence). Limited extensions of such models, including higher-order Markov chains for the daily occurrence process and autocorrelation of intensities on consecutive wet days, will be considered. In the literature on stochastic modeling of precipitation, evidence exists in support of making these adjustments, but the specific focus has not been on how well the variance of monthly total precipitation is approximated. So the present approach should provide additional insight concerning the relationship between daily and monthly variation of precipitation.

In section 2, some properties of stochastic models for time series of daily precipitation amount are reviewed. Essential to the present study is the representation of monthly total precipitation as a ''random sum,'' enabling its variance to be decomposed into two terms, one of which involves the variance of the number of

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Dr. Richard W. Katz, Environmental and Societal Impacts Group, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: rwk@ucar.edu

wet days. In section 3, these theoretical results are applied to a time series of daily precipitation amount in January at Chico, California. The same dataset was previously analyzed by Katz and Parlange (1993, 1996), in conjunction with a statistical downscaling study of the effect of large-scale atmospheric circulation on local precipitation. These studies raised the possibility that a more complex model for high-frequency variation of precipitation at Chico might reduce the extent of the overdispersion phenomenon. Through a combination of analytical expressions, computational algorithms, and simulation techniques, the variance of January total precipitation and related statistics are estimated for various extensions of a chain-dependent process. Some technical details are relegated to appendices. Finally, section 4 consists of a discussion.

2. Stochastic models

In this section, the probabilistic representation of total precipitation as a random sum is utilized. Loosely speaking, the basic idea is that monthly total precipitation consists of a sum of precipitation amounts contributed by individual storms. This representation enables the variance of total precipitation to be related to the various components of particular stochastic models for daily precipitation amount. For this reason, the approach has the advantage of applying more generally than just for the simplest form of chain-dependent process for precipitation (Katz 1977a).

a. Random sum representation

The total precipitation over some time period of length T days (e.g., a month), denoted by $S(T)$ say, can be expressed as

$$S(T) = Z_1 + Z_2 + \ldots + Z_{N(T)}. \tag{1}$$

Here the number of occurrences of precipitation within the time period (i.e., the number of ''wet days''), denoted by $N(T)$, is itself a *random variable.* The ''intensities'' $Z_k > 0$ corresponding to the kth occurrence (i.e., amount of precipitation on kth wet day), $k = 1, 2, \ldots, N(T)$, are taken independent and identically distributed (i.i.d.) with mean $\mu = E(Z_k)$ and variance $\sigma^2 = \text{Var}(Z_k)$. The counting process $N(T)$ is assumed independent of the Z_k's, with π denoting the unconditional probability of occurrence of precipitation on a given day (i.e., the rate of occurrence of the ''counts''). For simplicity and because attention will be restricted to time periods of a single month in duration, any annual cycles in these parameters are ignored.

Through conditioning on the number of wet days, the variance of monthly total precipitation is given by

$$\begin{aligned}\text{Var}[S(T)] &= E\{\text{Var}[S(T)\,|\,N(T)]\} + \text{Var}\{E[S(T)\,|\,N(T)]\} \\ &= E[N(T)]\,\text{Var}(Z_k) + \text{Var}[N(T)][E(Z_k)]^2 \\ &= T\pi\sigma^2 + \text{Var}[N(T)]\mu^2. \end{aligned} \tag{2}$$

The first expression actually holds for any two random variables (e.g., Lindgren 1968, 118), the second is specialized to a random sum (e.g., Feller 1968, chap. XII), and the third utilizes some simplifying notation. Thus, (2) provides a decomposition of the variance of total precipitation into two terms, one that corresponds to the variance of an ordinary (i.e., nonrandom) sum of $T\pi$ intensities (note that the expected number of wet days is $E[N(T)] = T\pi$) and another that involves the variance of the number of wet days, $\text{Var}[N(T)]$.

b. Chain-dependent process

A chain-dependent process (Katz 1977a; Todorovic and Woolhiser 1975) has the desirable feature of requiring only a relatively small number of parameters, while still accounting for the most important statistical features of precipitation time series. Its simple structure enables the analytical determination of many of its properties, including the variance of monthly total precipitation (Katz 1977b). In particular, the random sum representation for total precipitation, introduced in section 2a, applies to this class of stochastic model.

The tendency of wet spells (i.e., runs of consecutive days on which precipitation occurs) or of dry spells to persist is represented by a two-state, first-order Markov chain model for daily precipitation occurrence. Let $\{J_t: t = 1, 2, \ldots\}$ denote the sequence of daily precipitation occurrence (i.e., $J_t = 1$ indicates a wet day, $J_t = 0$ a dry day). This model is characterized by the transition probabilities

$$P_{ij} = \Pr\{J_{t+1} = j\,|\,J_t = i\}, \qquad i, j, = 0, 1, \tag{3}$$

with the constraint that $P_{i0} + P_{i1} = 1$, $i = 0, 1$.

It is convenient to reparameterize the Markov chain in terms of the probability of a wet day, denoted as in section 2a by $\pi = \Pr\{J_t = 1\}$, and the first-order autocorrelation coefficient (or persistence parameter), $d = \text{Corr}(J_t, J_{t+1})$. These two parameters, π and d, are related to the transition probabilities by

$$\pi = P_{01}/[1 - (P_{11} - P_{01})], \qquad d = P_{11} - P_{01}. \tag{4}$$

Note that $d > 0$ for time series of daily precipitation occurrence.

The number of wet days can be represented as a sum, $N(T) = J_1 + J_2 + \ldots + J_T$. Its variance, as appears in the general expression (2) for the variance of total precipitation, can be approximated (for large number of days T) by

$$\text{Var}[N(T)] \approx T\pi(1 - \pi)[(1 + d)/(1 - d)], \tag{5}$$

under the assumption of a first-order Markov chain (3)

(Gabriel 1959). It is evident that the Markovian dependence inflates this variance (provided $d > 0$) relative to independence [i.e., the variance of a binomial distribution is $T\pi(1 - \pi)$], as the factor $(1 + d)/(1 - d)$ appears on the right-hand side of (5). The exact variance for the number of wet days can be determined via a recursive calculation of the exact distribution for $N(T)$ (Katz 1974).

As already defined in section 2a in conjunction with a general random sum, the daily precipitation intensities Z_k for a chain-dependent process are assumed i.i.d. Nevertheless, it is convenient to introduce alternative notation, letting $\{X_t: t = 1, 2, \ldots\}$ denote the time series of daily precipitation amount (i.e., X_t assumes both zero and positive values). The equivalent assumption in terms of the intensities $X_t > 0$ (i.e., on days t for which $J_t = 1$) is that they are conditionally i.i.d., given the states of the Markov chain model for daily precipitation occurrence. In particular, the intensity mean and variance can be defined in terms of X_t as $\mu = E(X_t | J_t = 1)$ and $\sigma^2 = \text{Var}(X_t | J_t = 1)$. The daily intensity has a positively skewed distribution, taken to be exponential by Todorovic and Woolhiser (1975), gamma by Katz (1977a), and based on a power transformation to normality by Katz and Parlange (1993).

The unconditional mean, variance, and autocorrelation function of the X_t process are related to the intensity mean and variance, as well as the parameters of the Markov chain, by

$$\begin{aligned} E(X_t) &= \pi\mu, \\ \text{Var}(X_t) &= \pi\sigma^2 + \pi(1 - \pi)\mu^2, \\ \text{Corr}(X_t, X_{t+l}) &= \pi(1 - \pi)\mu^2 d^l/\text{Var}(X_t), \\ l &= 1, 2, \ldots \end{aligned} \tag{6}$$

(Katz and Parlange 1995). It is evident that the autocorrelations for the X_t process are induced through the autocorrelations of the Markov chain [i.e., $\text{Corr}(J_t, J_{t+l}) = d^l$, $l = 1, 2, \ldots$, appears in (6)].

Of course, the monthly total precipitation also has the more conventional representation of an ordinary, *nonrandom* sum, $S(T) = X_1 + X_2 + \ldots + X_T$. So its variance involves a sum of the autocorrelations of the X_t process [see (A1) and (A2) in appendix A]. Either substitution of the approximate expression (5) for the variance of the number of wet days into (2) or substitution of the expression (6) for the autocorrelation function into (A2) yields

$$\begin{aligned} &\text{Var}[S(T)] \\ &\quad \approx T(\pi\sigma^2 + \pi(1 - \pi)[(1 + d)/(1 - d)]\mu^2) \end{aligned} \tag{7}$$

for the variance of total precipitation for large T (Katz and Parlange 1993). By comparing (7) to the corresponding observed variance, an estimate of the extent of overdispersion can be obtained. Because a chain-dependent process fit to daily precipitation amount will effectively reproduce the daily variance, it is only through extensions affecting the autocorrelation function that the estimated variance of monthly total precipitation can be increased.

c. Extensions

1) Higher-order Markov chains

It is natural to permit the stochastic model for the daily occurrence of precipitation to be a Markov chain whose order is higher than first (Chin 1977; Gates and Tong 1976). For example, a second-order Markov chain is characterized by transition probabilities that depend on whether or not precipitation has occurred on the previous two days:

$$\begin{aligned} P_{ijk} &= \Pr\{J_{t+1} = k | J_t = j, J_{t-1} = i\}, \\ &\quad i, j, k = 0, 1. \end{aligned} \tag{8}$$

Note that a first-order chain is a special case of (8), with the constraint that $P_{ijk} = P_{jk}$, $i, j, k = 0, 1$. The probability of a wet day π and the persistence parameter d (still interpreted as the first-order autocorrelation coefficient) can be derived from the second-order transition probabilities (8) (see appendix B).

Because of the state-space representation of a higher-order Markov chain as a first-order chain with vector states, this model is equivalent to a first-order chain with more than two states. Thus, if the stochastic model for daily precipitation amount described in section 2b is extended to incorporate a higher-order Markov chain, the more general theory of chain-dependent processes can be employed (Katz 1977b). Specifically, the expression (7) for the variance of a sum of a chain-dependent process is a special case of a formula that involves the inverse of a matrix, an approach taken by Klugman and Klugman (1981).

Alternatively, the random sum representation (1) can be exploited, again applying the decomposition (2) for the variance of monthly total precipitation. The approximate expression (5) for the variance of the number of wet days, $\text{Var}[N(T)]$, is based on a first-order Markov chain and requires modification. Although no simple analog to (5) exists, the exact variance can be calculated via recursive methods. Appendix C gives an algorithm for determining the exact distribution of the number of wet days for a second-order Markov chain (8), a generalization of the method given in Katz (1974). It is anticipated that allowing for higher-order Markov chains would increase the estimated variance of the number of wet days, thus increasing the estimated variance of monthly total precipitation.

2) Autocorrelated intensities

Another, less common, extension of a chain-dependent process involves allowing the intensities within a given wet spell to be autocorrelated. Formally, the intensities are assumed to follow a first-order autoregres-

sive [AR(1)] process with autocorrelation coefficient $\phi > 0$. This process ''randomly'' terminates when the end of a wet spell is reached. In Katz and Parlange (1995), the AR(1) process is actually fitted to power transformed intensities to allow for skewness as well. They used this approach in modeling hourly precipitation amount, a situation in which the autocorrelation of intensities is more apparent. Previous attempts to allow for dependence among intensities have relied on a multistate Markov chain, an approach that requires the estimation of a large number of transition probabilities (Gregory et al. 1993; Haan et al. 1976).

If the first-order Markov chain model for the occurrence process is retained as in section 2b, then the autocorrelation function of the X_t process is given by

$$\mathrm{Corr}(X_t, X_{t+l}) = [\pi\sigma^2(P_{11}\phi)^l + \pi(1-\pi)\mu^2 d^l]/\mathrm{Var}(X_t), \quad l = 1, 2, \ldots \tag{9}$$

(Katz and Parlange 1995). Note that the transition probability that appears in (9) can be expressed as $P_{11} = \pi + (1 - \pi)d$ from (4). Here $\mathrm{Var}(X_t)$ is still as given in (6), but now the autocorrelation function (9) is a weighted average of two terms, one related to the autocorrelation function of the intensities (i.e., ϕ^l) and the other to the autocorrelation function of the occurrences (i.e., d^l).

It follows from (9) and (A2) that the variance of monthly total precipitation is approximately

$$\mathrm{Var}[S(T)] \approx T\langle \pi[(1 + P_{11}\phi)/(1 - P_{11}\phi)]\sigma^2 + \pi(1-\pi)[(1+d)/(1-d)]\mu^2\rangle, \tag{10}$$

for large T. We note that (10) generalizes (7), increasing the variance (provided $\phi > 0$) because the right-hand side now includes the factor $(1 + P_{11}\phi)/(1 - P_{11}\phi)$. Although no longer strictly speaking involving a random sum because the intensities that appear in (1) are assumed independent, (10) still constitutes a variance decomposition somewhat analogous to (7). The first term corresponds to the variance of a nonrandom sum of an AR(1) process with autocorrelation coefficient $P_{11}\phi$, the factor P_{11} arising because it governs the random termination of wet spells.

3) Nonidentical distributions

The original formulation of a chain-dependent process for daily precipitation actually allowed for another complication, permitting the daily intensity distribution to depend on whether or not precipitation occurred on the previous day (Katz 1977a). Specifically, define conditional means and variances

$$\mu_i = E(X_t \mid J_{t-1} = i, J_t = 1),$$
$$\sigma_i^2 = \mathrm{Var}(X_t \mid J_{t-1} = i, J_t = 1), \qquad i = 0, 1. \tag{11}$$

Besides Katz (1977a), Chin and Miller (1980) and Klugman and Klugman (1981) have fitted this model to daily precipitation data (see also Chapter 2 of Guttorp 1995).

For this form of chain-dependent process, the mean, variance, and autocorrelation function of daily precipitation amount are given by (Katz 1977a,b):

$$E(X_t) = (1 - \pi)P_{01}\mu_0 + \pi P_{11}\mu_1,$$
$$\mathrm{Var}(X_t) = (1 - \pi)P_{01}(\sigma_0^2 + \mu_0^2) + \pi P_{11}(\sigma_1^2 + \mu_1^2) - [E(X_t)]^2,$$
$$\mathrm{Corr}(X_t, X_{t+l}) = [\pi(1 - \pi)\mu(P_{11}\mu_1 - P_{01}\mu_0)d^{l-1}] \div \mathrm{Var}(X_t) \qquad l = 1, 2, \ldots. \tag{12}$$

Note that the unconditional intensity mean μ [that appears in (12)] and variance σ^2 can be related to the conditional intensity means and variances, μ_i's and σ_i^2's [see (B3) in appendix B]. These expressions in (12) reduce to those in (6) when $\mu_0 = \mu_1$ and $\sigma_0^2 = \sigma_1^2$. Finally, the approximate variance of monthly total precipitation can be derived from the autocorrelation function in (12) and (A2) as

$$\mathrm{Var}[S(T)] \approx T\langle \mathrm{Var}(X_t) + [2\pi(1 - \pi)\mu(P_{11}\mu_1 - P_{01}\mu_0)] \div (1 - d)\rangle, \tag{13}$$

for large T. To the extent that the two conditional intensity means and variances differ, this variance (13) will exceed that for an ordinary chain-dependent process (7).

4) Combinations of extensions 1), 2), and 3)

Naturally, various combinations of these three types of extensions of a chain-dependent process could be applied simultaneously. As noted previously, the general theory for chain-dependent processes (Katz 1977b) already encompasses both extensions 1) and 3), and consequently their combination (i.e., higher-order Markov chain for occurrences and nonidentically distributed intensities). Although a closed-form expression exists for the variance of monthly total precipitation, it involves large state spaces and matrix inversion (Klugman and Klugman 1981). A theory that encompasses all three extensions has yet to be devised. Nevertheless, in appendix A it is shown how to derive the first- and second-order autocorrelation coefficients of daily precipitation amount for this general situation.

3. Results

A time series of daily precipitation amount in January for 78 yr (during the period 1907–88, with 4 yr eliminated because of missing observations) at Chico, California, is analyzed. Because of its long record, a reasonably reliable estimate of the interannual variance of January total precipitation can be obtained. Chico is

TABLE 1. Transition probability estimates for Markov chains of various orders fit to 78 yr of time series of daily precipitation occurrence in January at Chico, California.

States on previous days				$\Pr\{J_{t+1} = 1 \mid J_t, J_{t-1}, J_{t-2}, J_{t-3}\}$			
J_{t-3}	J_{t-2}	J_{t-1}	J_t	Order 1	Order 2	Order 3	Order 4
0	0	0	0	0.2109	0.1838	0.1691	0.1584
0	0	0	1	0.5705	0.5882	0.5767	0.5541
0	0	1	0	0.2109	0.3105	0.2541	0.2857
0	0	1	1	0.5705	0.5576	0.5806	0.6083
0	1	0	0	0.2109	0.1838	0.2488	0.2184
0	1	0	1	0.5705	0.5882	0.6344	0.5161
0	1	1	0	0.2109	0.3105	0.3523	0.3188
0	1	1	1	0.5705	0.5576	0.5415	0.4857
1	0	0	0	0.2109	0.1838	0.1691	0.2039
1	0	0	1	0.5705	0.5882	0.5767	0.5918
1	0	1	0	0.2109	0.3105	0.2541	0.2000
1	0	1	1	0.5705	0.5576	0.5806	0.5254
1	1	0	0	0.2109	0.1838	0.2488	0.2679
1	1	0	1	0.5705	0.5882	0.6344	0.7167
1	1	1	0	0.2109	0.3105	0.3523	0.3922
1	1	1	1	0.5705	0.5576	0.5415	0.5882

situated near the west coast of the United States, a region where large-scale atmospheric circulation patterns have a dominant influence on local weather during the winter season. Thus, the precipitation process is expected to be relatively persistent, making this a stringent test for simple stochastic models. For these data, Katz and Parlange (1993) have already established that an ordinary chain-dependent process has a substantial degree of overdispersion.

a. Fitted models

The fitted models for the daily occurrences and intensities are presented separately. We reiterate that our focus is not on whether more complex models necessarily provide an improved fit (i.e., in terms of parameter estimates that are deemed ''statistically significant''), but rather on the ability of these models to estimate the variance of January total precipitation and related statistics.

1) OCCURRENCE PROCESS

Table 1 gives the estimated transition probabilities (based on the criterion of approximate maximum likelihood) for two-state Markov chains of orders 1–4 fit to the time series of daily precipitation occurrence in January at Chico. Because of the constraints on the transition probabilities (as noted in section 2b), only the conditional probability of a wet day is listed. To facilitate comparisons and because any lower-order chain can be viewed as a special case of a higher-order chain (as explained in section 2c), the estimates for all orders are presented in a form corresponding to a fourth-order chain. For each row in Table 1, the transition probability estimates would be constant (except for sampling errors) if the time series of precipitation occurrence were actually generated by a first-order chain. Some evidence is present in the table that dry spells exhibit a form of persistence that cannot be modeled by a first-order chain. For instance, the estimated conditional probability of a wet day decreases from about 0.211, given only that the previous day is dry, to about 0.158, given that the last 4 days are dry (these two probabilities would be identical for a first-order chain). The pattern is less clear for wet spells.

For a first-order Markov chain, the transition probability estimates can be converted [via (4)] into the corresponding estimates for the unconditional probability of a wet day and the persistence parameter, $\hat{\pi} = 0.3293$ and $\hat{d} = 0.3596$ (where ''$\hat{}$'' denotes the estimator of a parameter). As the order is increased, only slight differences in the numerical values of π and d arise (e.g., for a second-order chain, $\hat{\pi} = 0.3290$ and $\hat{d} = 0.3603$; determined by method in appendix B). The situation is somewhat analogous to that for autoregressive processes, in which a higher-order model contributes to an improved fit only through adjustments in the autocorrelations at higher lags.

2) INTENSITY PROCESS

Table 2 gives the estimated parameters for the conditional means, conditional standard deviations, and first-order autocorrelation coefficient of daily precipitation intensity in January at Chico. For convenience, these estimates are presented in a general form for nonidentical intensity distributions, recalling that the identically distributed case corresponds to the constraint that $\mu_0 = \mu_1$ and $\sigma_0^2 = \sigma_1^2$ (see section 2c). It is evident

TABLE 2. Parameter estimates for various forms of model fit to daily precipitation intensities in January at Chico, California (787 wet days).

Form of model		Means		Standard deviations		Autocorrelation
Identical distribution?	Autocorrelation?	μ_0 (mm)	μ_1 (mm)	σ_0 (mm)	σ_1 (mm)	ϕ
Yes	No	13.36	13.36	14.68	14.68	0
Yes	Yes	13.36	13.36	14.68	14.68	0.161
Yes	Inflated	13.36	13.36	14.68	14.68	0.411
No	No	11.62	14.84	12.15	16.28	0
No	Yes	11.62	14.84	12.15	16.28	0.145
No	Inflated	11.62	14.84	12.15	16.28	0.298

from Table 2 that the estimated conditional intensity mean and standard deviation are smaller given that the previous day was dry as opposed to wet. To facilitate comparison of estimates of the variance of monthly total precipitation, these parameters were estimated by the method of moments. An alternative approach would be to base the parameter estimates on a power transformation to normality, allowing for the positively skewed distribution of daily intensity (Katz and Parlange 1993). But the use of a nonlinear transformation introduces some degree of bias, implying that the sample mean and variance of the original, untransformed daily intensities are not exactly reproduced by this technique. Taking into account this skewness would only matter if the shape of the distribution of monthly total precipitation, not just its variance, were being studied.

The estimates of the first-order autocorrelation coefficient ϕ for daily intensity in Table 2 are relatively small positive values, 0.161 or 0.145, depending on whether or not the intensity distribution is assumed identical. More generally, the estimated autocorrelation is quite sensitive to any differences in the intensity distribution depending on the position within the wet spell (first vs second day of wet spell, etc.), a refinement of the model for nonidentical distributions described in section 2c. To circumvent this problem, an alternative method of estimating ϕ is also included in Table 2, predicated upon reproducing the sample first-order autocorrelation coefficient of daily precipitation amount. In the case of identically distributed intensities, this method involves substituting the estimates for μ and σ (along with π, d, and P_{11} for a first-order Markov chain) into the right-hand side of (9) with lag $l = 1$, equating this expression with the sample first-order autocorrelation coefficient for daily precipitation amount of 0.279 and then solving for ϕ. In the case of nonidentically distributed intensities, the same approach is taken, but now μ_i and σ_i^2, $i = 0, 1$ (along with π, P_{01}, and $P_{011} = P_{111} = P_{11}$ for a first-order Markov chain) are substituted into (A4). Quite a bit larger estimates of ϕ are obtained, termed ''inflated'' in Table 2, 0.411 versus 0.161 for identical distributions, with the degree of inflation being somewhat less, 0.298 versus 0.145, for nonidentical distributions. A more complex model for nonidentical distributions (i.e., explicitly taking into account the position within a wet spell) could eliminate these discrepancies entirely.

b. Overdispersion estimates

Through use of these fitted stochastic models for time series of daily precipitation amount, the variance of monthly total precipitation can be estimated. Because of their diagnostic capability, the variance of the number of wet days and the autocorrelation function of daily precipitation amount are also examined. To obtain estimates of these statistics, either an explicit formula, a computational algorithm, or stochastic simulation (based on the generation of 10 000 yr of January daily precipitation amount) is employed. All of these approaches require numerical values for the parameters of the fitted models. We simply substitute the corresponding parameter estimates (given in section 3a), ignoring the sampling errors associated with those estimates. The large number of daily observations of precipitation (i.e., $2418 = 78 \times 31$) suggests that such uncertainties are relatively small.

1) NUMBER OF WET DAYS

In view of the variance decomposition (2), we now focus on how well the observed standard deviation of the number of wet days in January at Chico is matched by the Markov chain model for the daily occurrence of precipitation. Table 3 includes the estimated standard deviation of the number of wet days for Markov chains of order 1–4 whose parameter estimates are given in Table 1. These estimated standard deviations were obtained through the computational algorithm outlined in appendix C.

For a first-order Markov chain, the standard deviation of the number of wet days is estimated as 3.76 days [the approximate expression (5) yields nearly the same value, 3.81], well below the observed value of 4.33 days, or an overdispersion of about 25% in terms of variance (Table 3). As the order of the chain is increased, this estimated standard deviation increases as well, being only slightly below the observed value for a third-order chain and slightly above for a fourth-order chain. This overestimate for a fourth-order chain most likely does not reflect real ''underdispersion,'' but rather just the sampling error in both the observed and model-estimated standard deviations.

Most importantly, the overdispersion phenomenon with respect to the number of wet days has been completely eliminated through the use of a higher-order chain. One drawback is that the number of transition probabilities required to be estimated increases at a rapid rate as the order is increased (e.g., 16 parameters for a fourth-order chain—see Table 1). So the possibility of overfitting is present. An alternative, not explored here, would be to fit a more parsimonious model for a higher-order chain in which certain constraints are placed among the parameters (Raftery 1985).

2) AUTOCORRELATION

In view of the representations (A1) and (A2) of the variance of monthly total precipitation, we now focus on how well the autocorrelation function of daily precipitation amount is reproduced by the various forms of a stochastic model. The lag $l = 1$ and 2 day autocorrelation coefficients, calculated through use of (A4) and (A5), are included in Table 3. The ordinary form of the chain-dependent process has a marked tendency to underestimate the first-order autocorrelation coefficient of

TABLE 3. Overdispersion estimates and related statistics for January precipitation at Chico, California, based on daily stochastic models (parameter estimates in Tables 1 and 2), along with corresponding observed values.

Form of model			Derived statistics			
Markov chain order	Identical intensity distribution?	Intensity autocorrelation?	Std dev no. wet days	Amount autocorrelation (lag 1)	Amount autocorrelation (lag 2)	Std dev total precipitation (mm)
1	Yes	No	3.76	0.128	0.046	68.7
1	Yes	Yes	3.76	0.187	0.052	71.8
1	Yes	Inflated	3.76	0.279	0.082	77.6
1	No	No	3.76	0.160	0.058	71.4
1	No	Yes	3.76	0.218	0.062	74.5
1	No	Inflated	3.76	0.279	0.078	77.3
2	Yes	No	4.00	0.129	0.065	71.1
2	Yes	Yes	4.00	0.188	0.071	74.5
2	Yes	Inflated	4.00	0.279	0.100	79.5
2	No	No	4.00	0.160	0.073	73.2
2	No	Yes	4.00	0.217	0.078	76.4
2	No	Inflated	4.00	0.278	0.093	78.4
3	Yes	No	4.23	0.129	0.065	73.5
3	Yes	Yes	4.23	0.189	0.071	76.9
3	Yes	Inflated	4.23	0.281	0.100	81.0
3	No	No	4.23	0.160	0.073	76.5
3	No	Yes	4.23	0.218	0.078	79.5
3	No	Inflated	4.23	0.279	0.093	82.5
4	Yes	No	4.42	0.129	0.068	75.4
4	Yes	Yes	4.42	0.189	0.073	79.4
4	Yes	Inflated	4.42	0.281	0.103	83.7
4	No	No	4.42	0.160	0.076	79.0
4	No	Yes	4.42	0.218	0.080	80.6
4	No	Inflated	4.42	0.278	0.095	83.9
Observed			4.33	0.279	0.113	88.6

daily precipitation amount (i.e., 0.128 vs an observed value of 0.279 in Table 3). Permitting either nonidentically distributed or autocorrelated intensities increases this estimate somewhat, with their combination producing a value of 0.218, still well below the observed value. Necessarily, the ''inflated'' intensity autocorrelations do reproduce the desired value. Increasing the order of the Markov chain has no effect on the first-order autocorrelation coefficient, with the very slight numerical differences being attributable to the manner in which the lower-order probabilities are derived (appendix B).

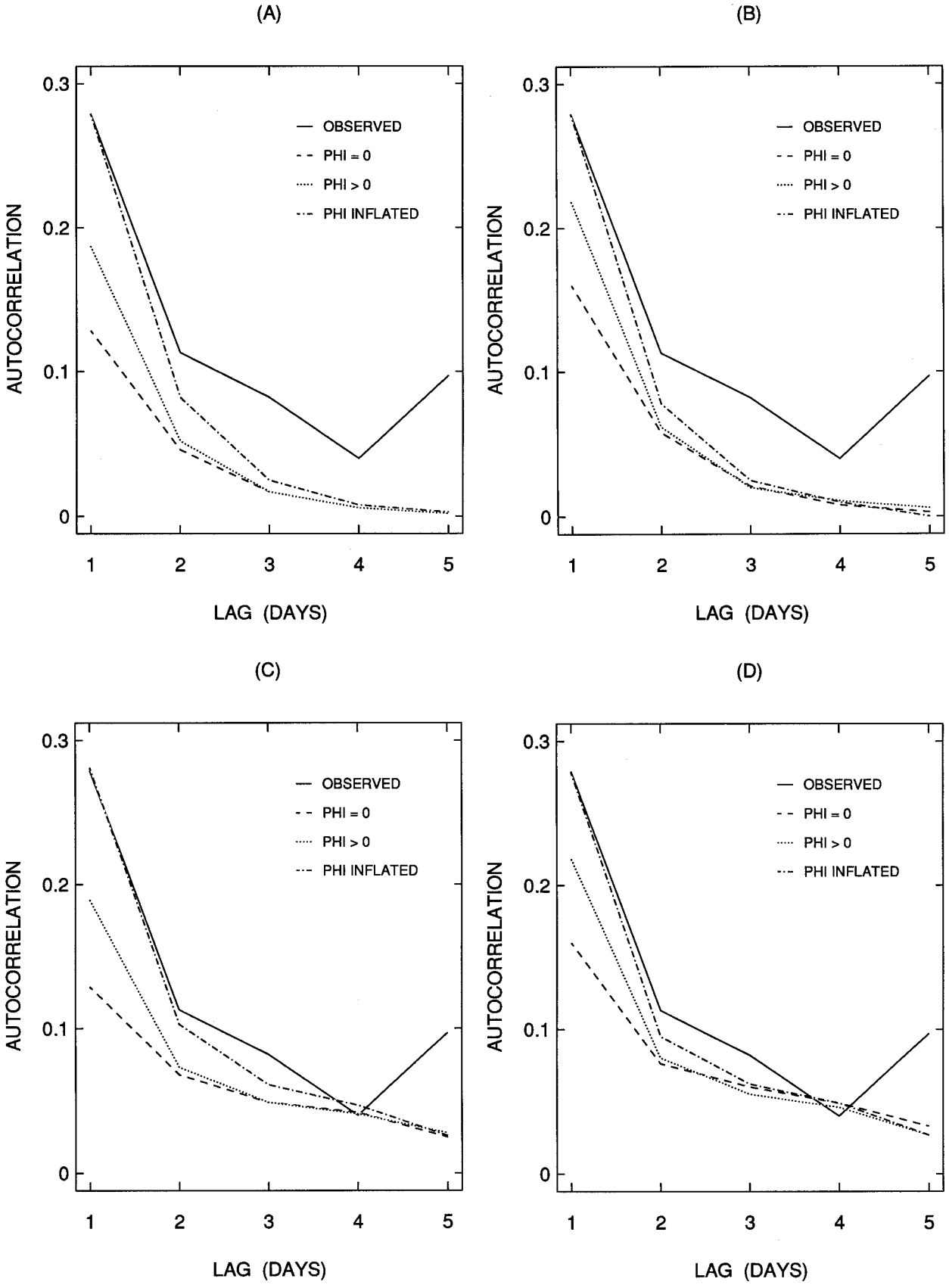
Likewise, the second-order autocorrelation coefficient is underestimated by the ordinary form of the chain-dependent process (i.e., 0.046 vs 0.113 in Table 3). In this case, increasing the order of the Markov chain from one to two, as well as allowing for nonidentically distributed and autocorrelated intensities, all contribute to increases in this estimate, with their combination producing a value of 0.078. When the intensity autocorrelation is inflated as well, the largest value produced is 0.103 (for a fourth-order chain with identical distributions), still slightly below that observed. It is important to recognize that this approach of inflating the parameter ϕ is not constrained to reproduce the autocorrelation at lags $l \geq 2$ days. This deficiency in estimating the autocorrelations was also found by Gregory et al. (1993) for precipitation in the United Kingdom.

Figure 1 shows the autocorrelation function up to lag $l = 5$ days for a subset of these stochastic models (i.e., Markov chain order restricted to first or fourth), with the lags $l = 3$, 4, and 5 days being estimated by stochastic simulation. For a first-order chain, it is evident that the higher-order autocorrelations (i.e., lags $l \geq 2$) are substantially underestimated, no matter whether the intensities are identically distributed or not or autocorrelated or not (Figs. 1a,b). On the other hand, not much underestimation is evident for a fourth-order chain, no matter what the other model assumptions, provided the apparent increase of the observed fifth-order autocorrelation over the fourth is not regarded as real (Figs. 1c,d).

3) TOTAL PRECIPITATION

Table 3 also includes the estimated standard deviation of January total precipitation at Chico for the various forms of stochastic model, and the same numbers are displayed in Fig. 2. For identically distributed intensities without any autocorrelation, the estimates are obtained from (2) for any order Markov chain (using the calculation of the variance of the number of wet days). For a first-order chain with identically distributed, autocorrelated intensities, the estimates are obtained from (A1) and (9) [for nonidentically distributed, uncorrelated intensities from (A1) and (12)]. Otherwise, the estimates are based on stochastic simulation.

As anticipated from Katz and Parlange (1993), the



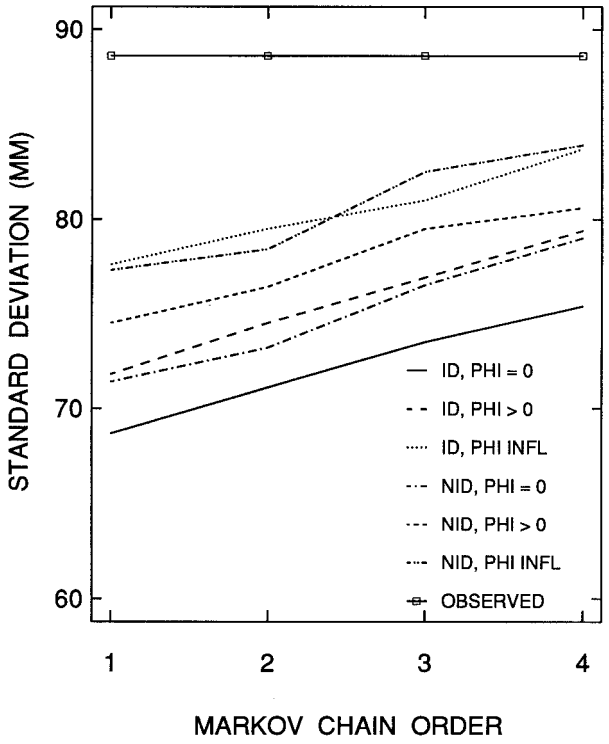


FIG. 2. Estimated standard deviation of January total precipitation at Chico, California, for various forms of stochastic model for daily precipitation amount, shown as function of order of Markov chain model for occurrence (curves correspond to six versions for intensity distribution and autocorrelation, and horizontal line depicts observed value).

estimated standard deviation for the ordinary form of chain-dependent process is well below that observed, 68.7 mm [the approximate expression (7) yields nearly the same value, 69.2 mm] as compared to 88.6 or an overdispersion of about 40% in terms of variance (Table 3). The highest estimate is 80.6 mm (17% overdispersion) for a fourth-order chain with nonidentically distributed, autocorrelated intensities, or 83.9 (10% overdispersion) if inflated autocorrelation is permitted as well. Figure 2 illustrates that increasing the Markov chain order has the greatest effect on the estimated standard deviation, with the intensity autocorrelation having a lesser effect (roughly comparable if autocorrelations are allowed to be inflated), and with nonidentical distributions making the smallest contribution (but recall that inflated autocorrelations may well be a surrogate for a more complex form of nonidentically distributed intensities).

In any event, it is evident that the extent of overdispersion can be greatly reduced, if not eliminated, through use of a more complex form of stochastic model for high-frequency variation of precipitation. We note that Klugman and Klugman (1981) also found that the estimated standard deviation of seasonal total precipitation at a site in Oregon is sensitive to the assumed form of model. The question remains whether even more complex models than those considered here could completely eliminate this overdispersion.

4. Discussion

It has been established that much of the overdispersion for January total precipitation at Chico, California, could be attributable to an inadequate stochastic model for high-frequency variation of precipitation. A higher-order Markov chain model for daily precipitation occurrence completely eliminates one source of overdispersion, the number of wet days. The allowance for autocorrelated and nonidentically distributed intensities also contributes to this reduction in overdispersion. Although the appropriate form of stochastic model for daily precipitation at other locations might well differ from that for Chico, it is anticipated that the estimated variance of monthly total precipitation would likewise be sensitive to the assumed form of model.

Could the overdispersion for monthly total precipitation be further reduced? Results obtained through relating daily precipitation statistics to large-scale atmospheric circulation shed some light on this question. Katz and Parlange (1993) fit the simplest version of chain-dependent processes conditionally, given an index of large-scale atmospheric circulation, to the same daily precipitation data for Chico in January. When these conditional models are combined into a single overall ''induced'' model, the overdispersion of January total precipitation is reduced to about 4%, smaller yet than the reductions obtained in the present paper (i.e., 10% or 17% for best models). Of interest is the fact that the induced model completely eliminates the overdispersion in the number of wet days, in agreement with the result obtained here (Katz and Parlange 1996).

The implications of the present work for estimating potential predictability remain to be explored. Although the two approaches are not equivalent, the induced model resembles in some respects a single, more complex stochastic model that could have been directly fitted to the data (Katz and Parlange 1996). Future work will seek ways in which these two approaches could be unified. One possibility would involve so-called hidden Markov models and their generalizations (Guttorp 1995,

←

FIG. 1. Autocorrelation functions for daily precipitation amount in January at Chico, California, derived from various forms of stochastic model (curves indicate three versions for intensity autocorrelation), along with sample autocorrelation function (solid curve): (a) first-order Markov chain for occurrences and identical distributions for intensities, (b) first-order and nonidentical, (c) fourth-order and identical, and (d) fourth-order and nonidentical.

chap. 2). These models involve a hidden state, like an index of atmospheric circulation but unobserved. The variance in monthly total precipitation associated with the hidden states could perhaps be construed as an estimate of potential predictability.

Acknowledgments. Research was partially supported by NSF Grant DMS-9312686 to the NCAR Geophysical Statistics Project. M. B. Parlange received support from NCAR's Environmental and Societal Impacts Group and performed a portion of this research at the University of California, Davis.

APPENDIX A

Autocorrelation Function of Generalized Chain-Dependent Process

For any stationary stochastic process $\{X_t: t = 1, 2, \ldots\}$, its sum, $S(T) = X_1 + X_2 + \ldots + X_T$, has variance that is related to its autocorrelation function by

$$\text{Var}[S(T)] = T\,\text{Var}(X_t)\left[1 + 2\sum_{l=1}^{T-1}(1 - l/T)\,\text{Corr}(X_t, X_{t+l})\right]. \tag{A1}$$

For large T, (A1) can be approximated as

$$\text{Var}[S(T)] \approx T\,\text{Var}(X_t)\left[1 + 2\sum_{l=1}^{\infty}\text{Corr}(X_t, X_{t+l})\right], \tag{A2}$$

provided the autocorrelation function is absolutely summable (e.g., Brockwell and Davis 1991, chap. 7).

We outline a general approach to deriving expressions for the autocorrelations, as appear in (A1) and (A2), that applies to any of the extensions of a chain-dependent process considered. It is always the case that

$$E(X_t, X_{t+l}) = \Pr\{J_t = 1, J_{t+l} = 1\}E(X_tX_{t+l}\,|\,J_t = 1, J_{t+l} = 1), \quad l = 1, 2, \ldots. \tag{A3}$$

For lag $l = 1$, (A3) can be expanded in terms of the model parameters as

$$E(X_tX_{t+1}) = (1 - \pi)P_{01}P_{011}(\mu_0\mu_1 + \sigma_0\sigma_1\phi) + \pi P_{11}P_{111}(\mu_1^2 + \sigma_1^2\phi). \tag{A4}$$

This expression (A4) is written in a form that involves transition probabilities of order two (i.e., P_{ijk}'s) and order one (i.e., P_{ij}'s).

For lag $l = 2$, a somewhat more complex expression can be obtained:

$$E(X_t, X_{t+2}) = (1 - \pi)P_{01}P_{010}P_{0101}(\mu_0^2) + (1 - \pi)P_{01}P_{011}P_{0111}(\mu_0\mu_1 + \sigma_0\sigma_1\phi^2) + \pi P_{11}P_{110}P_{1101}\mu_0\mu_1 + \pi P_{11}P_{111}P_{1111}(\mu_1^2 + \sigma_1^2\phi^2). \tag{A5}$$

This expression (A5) is written in a form that involves transition probabilities of order 3 [denoted by P_{ijkl}'s in (A5) analogous to (3) and (8)] or lower. Derivations of higher-order autocorrelation coefficients by this approach are more tedious.

APPENDIX B

Relationships among Parameters

a. Second- versus first-order Markov chain

For a second-order Markov chain, define the joint probabilities

$$q_{ij} = \Pr\{J_t = i, J_{t+1} = j\} = \pi_iP_{ij}, \quad i, j, = 0, 1, \tag{B1}$$

where $\pi_1 = \pi$, $\pi_0 = 1 - \pi$, and P_{ij} is still defined as in (3). These q_{ij}'s can be determined from the transition probabilities for the second-order chain (8) through the following system of equations:

$$q_{ij} = q_{0i}P_{0ij} + q_{1i}P_{1ij}, \qquad i, j = 0, 1. \tag{B2}$$

Once (B2) is solved for the q_{ij}'s, the lower-order probabilities can be derived by first obtaining $\pi_i = q_{0i} + q_{1i}$, $i = 0, 1$, and then solving (B1) for P_{ij}, $i, j = 0, 1$. It is straightforward but requires more complex notation to obtain analogous relationships for higher than second-order chains. Although (B2) is simple enough that closed-form expressions for the q_{ij}'s can be obtained, numerical techniques would be required for higher- than second-order chains.

b. Nonidentically versus identically distributed intensities

For nonidentically distributed intensities, the conditional means and variances are related to the unconditional mean and variance, μ and σ^2, as defined in section 2b, by

$$\mu = (1 - P_{11})\mu_0 + P_{11}\mu_1,$$
$$\sigma^2 = (1 - P_{11})\sigma_0^2 + P_{11}\sigma_1^2 + P_{11}(1 - P_{11})(\mu_1 - \mu_0)^2. \tag{B3}$$

The formula for the mean μ in (B3) is obtained by equating the two expressions for $E(X_t)$, (6) and (12), and using the fact that $(1 - \pi)P_{01} = \pi(1 - P_{11})$. The formula for the variance σ^2 in (B3) can be derived by a similar approach.

APPENDIX C

Distribution of Number of Wet Days for Second-Order Markov Chain

Define the following conditional probability distributions for the number of wet days (here it is convenient to condition on the two days prior to time $t = 1$):

$$p_T(k; i, j) = \Pr\{N(T) = k \mid J_0 = j, J_{-1} = i\},$$
$$i, j = 0, 1; \qquad k = 0, 1, \ldots, T. \tag{C1}$$

Now the unconditional probability distribution of the number of wet days can be determined from (C1) by

$$\Pr\{N(T) = k\} = \sum_{i,j} q_{ij} p_T(k; i, j), \qquad k = 0, 1, \ldots, T, \tag{C2}$$

where q_{ij} is as defined in (B1).

The conditional probability distributions (C1) satisfy the following recursion for $T = 2, 3, \ldots$:

$$p_T(k; i, j) = P_{ij0} p_{T-1}(k; j, 0) + P_{ij1} p_{T-1}(k - 1; j, 1), \tag{C3}$$

$i, j = 0, 1$; $k = 0, 1, \ldots, T$. The initial conditions for this recursion are

$$p_1(k; i, j) = P_{ijk}, \qquad i, j, k = 0, 1. \tag{C4}$$

The following boundary conditions also need to be imposed on (C3):

$$p_T(T + 1; j, 0) = p_T(-1; j, 1) = 0,$$
$$j = 0, 1; \qquad T = 1, 2, \ldots. \tag{C5}$$

These expressions are straightforward extensions of the approach for a first-order chain (Katz 1974), with the extension to a higher than second-order chain being likewise straightforward.

REFERENCES

Brockwell, P. J., and R. A. Davis, 1991: *Time Series: Theory and Methods.* 2d ed. Springer-Verlag, 577 pp.

Buishand, T. A., 1978: Some remarks on the use of daily rainfall models. *J. Hydrol.,* **36,** 295–308.

Chin, E. H., 1977: Modeling daily precipitation occurrence process with Markov chain. *Water Resour. Res.,* **13,** 949–956.

——, and J. F. Miller, 1980: On the conditional distribution of daily precipitation amounts. *Mon. Wea. Rev.,* **108,** 1462–1464.

Cox, D. R., 1983: Some remarks on overdispersion. *Biometrika,* **70,** 269–274.

Feller, W., 1968: *An Introduction to Probability Theory and Its Applications.* Vol. I. 3d ed. Wiley, 509 pp.

Gabriel, K. R., 1959: The distribution of the number of successes in a sequence of dependent trials. *Biometrika,* **46,** 454–460.

Gates, P., and H. Tong, 1976: On Markov chain modeling to some weather data. *J. Appl. Meteor.,* **15,** 1145–1151.

Gregory, J. M., T. M. L. Wigley, and P. D. Jones, 1993: Application of Markov models to area-average daily precipitation series and interannual variability in seasonal totals. *Climate Dyn.,* **8,** 299–310.

Guttorp, P., 1995: *Stochastic Modeling of Scientific Data.* Chapman and Hall, 372 pp.

Haan, C. T., D. M. Allen, and J. O. Street, 1976: A Markov chain model of daily rainfall. *Water Resour. Res.,* **12,** 443–449.

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

——, 1977a: Precipitation as a chain-dependent process. *J. Appl. Meteor.,* **16,** 671–676.

——, 1977b: An application of chain-dependent processes to meteorology. *J. Appl. Probability,* **14,** 598–603.

——, and M. B. Parlange, 1993: Effects of an index of atmospheric circulation on stochastic properties of precipitation. *Water Resour. Res.,* **29,** 2335–2344.

——, and ——, 1995: Generalizations of chain-dependent processes: Application to hourly precipitation. *Water Resour. Res.,* **31,** 1331–1341.

——, and ——, 1996: Mixtures of stochastic processes: Application to statistical downscaling. *Climate Res.,* **7,** 185–193.

Klugman, M. R., and S. A. Klugman, 1981: A method for determining change in precipitation data. *J. Appl. Meteor.,* **20,** 1506–1509.

Lindgren, B. W., 1968: *Statistical Theory.* 2d ed. Macmillan, 521 pp.

Raftery, A. E., 1985: A model for higher-order Markov chains. *J. Roy. Stat. Soc., Ser. B,* **47,** 528–539.

Shea, D. J., and R. A. Madden, 1990: Potential for long-range prediction of monthly mean surface temperatures over North America. *J. Climate,* **3,** 1444–1451.

——, N. A. Sontakke, R. A. Madden, and R. W. Katz, 1995: The potential for long-range prediction over India for the southwest monsoon season: An analysis of variance approach. Preprints, *Sixth Int. Meeting on Statistical Climatology,* Galway, Ireland, University College, 475–477.

Singh, S. V., and R. H. Kripalani, 1986: Potential predictability of lower-tropospheric monsoon circulation and rainfall over India. *Mon. Wea. Rev.,* **114,** 758–763.

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

Wilks, D. S., 1989: Conditioning stochastic daily precipitation models on total monthly precipitation. *Water Resour. Res.,* **25,** 1429–1439.