Mixture Model For Overdispersion of Precipitation

RICHARD W. KATZ

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

XIAOGU ZHENG

National Institute of Water and Atmospheric Research, Wellington, New Zealand

(Manuscript received 8 July 1998, in final form 30 October 1998)

ABSTRACT

Stochastic models fit to time series of daily precipitation amount generally ignore any year-to-year (i.e., low frequency) source of random variation, and such models are known to underestimate the interannual variance of monthly or seasonal total precipitation. To explicitly account for this ‘‘overdispersion’’ phenomenon, a mixture model is proposed. A hidden index, taking on one of two possible states, is assumed to exist (perhaps representing different modes of atmospheric circulation). To represent the intermittency of precipitation and the tendency of wet or dry spells to persist, a stochastic model known as a chain-dependent process is applied. The parameters of this stochastic model are permitted to vary conditionally on the hidden index.

Data for one location in California (whose previous study motivated the present approach), as well as for another location in New Zealand, are analyzed. To estimate the parameters of a mixture of two conditional chain-dependent processes by maximum likelihood, the ‘‘expectation-maximization algorithm’’ is employed. It is demonstrated that this approach can either eliminate or greatly reduce the extent of the overdispersion phenomenon. Moreover, an attempt is made to relate the hidden indexes to observed features of atmospheric circulation. This approach to dealing with overdispersion is contrasted with the more prevalent alternative of fitting more complex stochastic models for high-frequency variations to time series of daily precipitation.

1. Introduction

Simple stochastic models fit to time series of daily precipitation amount commonly do not include a component that explicitly accounts for interannual (i.e., low frequency) variation. Of course, such models are still capable of producing a substantial interannual variance in monthly (or seasonal) total precipitation, in this case being solely attributable to their representation of high-frequency, day-to-day variations. Nevertheless, these models typically underestimate the observed interannual variance of monthly total precipitation by a nonnegligible fraction (Buishand 1978; Wilks 1989). This phenomenon is, more generally, termed ‘‘overdispersion’’ in the statistics literature (e.g., Cox 1983).

Two conflicting explanations for the overdispersion phenomenon have been proposed. The first one involves viewing the discrepancy in variance as evidence of an inadequate model for the high-frequency variations of daily precipitation (Gregory et al. 1993). In this regard, Katz and Parlange (1998) obtained results suggesting that the extent of the overdispersion could be reduced, but not necessarily eliminated, by fitting more complex stochastic models that better reflect the nature of the temporal dependence of daily precipitation.

The second explanation involves attributing the overdispersion phenomenon to low-frequency variations ignored by these stochastic models. In fact, some researchers have treated this difference in variance as a measure of the ‘‘potential predictability’’ of precipitation on an interannual timescale (Madden et al. 1999; Singh and Kripalani 1986). In the present paper, the extent to which the overdispersion of precipitation can be reduced by explicitly accounting for low-frequency variations is studied in greater detail. Recent research in statistics indicates that these two explanations can be difficult to distinguish in practice without systematically considering each possibility (Fitzmaurice 1997).

The approach to be taken is motivated by the work of Katz and Parlange (1993), in which stochastic models for daily precipitation were fitted conditionally on an index of large-scale atmospheric circulation. Despite the index being only imperfectly related to local precipi-

* 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

tation, this approach did greatly reduce, if not eliminate, the overdispersion. Essentially the same approach to the modeling of daily precipitation is taken, except that now the index is treated as ''hidden'' (i.e., unobserved). The premise of the present approach is that, a priori, the relationship between large-scale atmospheric circulation patterns and precipitation at a particular location may not necessarily be well understood. As such, it could play a diagnostic role in climate research, detecting hidden sources of low-frequency variation whose origin might be the focus of subsequent study.

Specifically, a mixture model is proposed with a hidden index that takes on one of two possible states in a given year. Conditional on this index, the parameters of a stochastic model for daily time series of precipitation amount, known as a chain-dependent process (Katz and Parlange 1993), are permitted to vary. It should be noted that both Jones et al. (1995) and Zheng (1996) have formulated statistical models for times series of daily temperature in which the mean is permitted to vary interannually, in effect, taking on infinitely many states (termed a ''random effect'' in the statistics literature). Because of the complex nature of the precipitation process (especially its intermittency), it is not feasible to directly apply this analysis of variance approach. The assumption of two (or a small number of) hidden states, while highly restrictive, is consistent with the belief that only a few dominant modes of large-scale atmospheric circulation exist (e.g., Hansen and Sutera 1995).

In section 2, a stochastic model for time series of daily precipitation amount, consisting of a mixture of two conditional chain-dependent processes, is defined and some of its properties are outlined. Because the index is hidden, section 3 describes a specialized statistical technique, known as the expectation-maximization (EM) algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997), needed to estimate the model parameters by maximum likelihood. This method has only rarely been applied in the climate literature (Sansom 1995; Sansom 1998; Sansom and Thomson 1992). Some technical details concerning the implementation of the EM algorithm are relegated to an appendix. In section 4, results are presented for a location in California previously analyzed by Katz and Parlange (1993, 1996, 1998), as well as for another site in New Zealand. Finally, section 5 consists of a discussion of the interpretation of the results and of other potential applications of the methodology, as well as of some possible extensions.

2. Stochastic model

a. Chain-dependent process

First the definition and properties of a chain-dependent process, a relatively simple stochastic model for daily precipitation, are briefly reviewed. This model represents the most important features of precipitation, including its intermittency and the tendency of wet or dry spells to persist (for further details, see Katz and Parlange 1993).

1) DEFINITION

Let $\{J_t : t = 1, 2, \ldots, T\}$ denote the sequence of daily precipitation occurrence (i.e., $J_t = 1$ indicates a ''wet day'' and $J_t = 0$ a ''dry day''). It is assumed that this process is a first-order Markov chain, a model completely characterized by the transition probabilities

$$P_{jk} = \Pr\{J_{t+1} = k | J_t = j\}, \qquad j, k = 0, 1. \quad (1)$$

Note that $P_{j0} = 1 - P_{j1}$, $j = 0, 1$. For some purposes, it is convenient to reparameterize the Markov chain in terms of the unconditional probability of a wet day, $\pi = \Pr\{J_t = 1\}$, and the first-order autocorrelation coefficient (or ''persistence parameter''), $d = \mathrm{Corr}(J_t, J_{t+1})$. Now it is well known that

$$\pi = P_{01}/[1 - (P_{11} - P_{01})], \qquad d = P_{11} - P_{01}. \quad (2)$$

Let $\{X_t : t = 1, 2, \ldots, T\}$ denote the time series of daily precipitation amount. The ''intensities'' $X_t > 0$ (i.e., days for which $J_t = 1$) are taken to be conditionally independent and identically distributed with mean $\mu = E(X_t | J_t = 1)$ and $\sigma^2 = \mathrm{Var}(X_t | J_t = 1)$. It is further assumed that daily precipitation intensity has a power transform distribution. That is, a transformation,

$$X_t^* = X_t^p, \qquad 0 < p < 1, \quad (3)$$

exists such that the transformed variable X_t^* has a normal distribution, say, with mean μ^* and variance $(\sigma^*)^2$ [note that μ and σ^2 each are functions of both μ^* and $(\sigma^*)^2$; e.g., Katz (1999).] For example, a value of $p = ½$, ⅓, or ¼ is commonly employed to account for the high degree of positive skewness in the distribution of daily intensity. Alternatively, a positively skewed distribution, such as the gamma, may be fitted directly to the untransformed intensities.

2) PROPERTIES

The overdispersion phenomenon concerns the variance of precipitation totaled over a period of length T days (e.g., a month or season), $S(T) = X_1 + X_2 + \cdots + X_T$. The mean and variance of total precipitation are related to the parameters of a chain-dependent process by

$$E[S(T)] = T\pi\mu,$$
$$\mathrm{Var}[S(T)] \approx T\{\pi\sigma^2 + \pi(1 - \pi)[(1 + d)/(1 - d)]\mu^2\} \quad (4)$$

(e.g., Katz and Parlange 1993). The expression for the variance is an approximation valid for a large number of days T. Estimates based on (4) can be substantially lower than the observed interannual variance of monthly or seasonal total precipitation (Katz and Parlange 1998).

b. Mixture model

1) Definition

For a given year, the time series of daily precipitation amount $\{X_t : t = 1, 2, \ldots, T\}$ is assumed to be a conditional chain-dependent process, with parameters π_i, d_i, μ_i, and σ_i^2 [or, equivalently, $P_{01}(i)$, $P_{11}(i)$, μ_i^*, $(\sigma_i^*)^2$] that depend on a two-state index I ($I = i$, $i = 0, 1$). The power transform parameters p_i, $i = 0, 1$, for the intensity distributions of the conditional models [see (3)], are taken the same as for the single, unconditional chain-dependent process (i.e., $p = p_0 = p_1$). It is further assumed that the index remains constant over the time period of T days (in our application, $T \approx 30$) and that the annual sequence of index states is independent and identically distributed, with common distribution

$$w = \Pr\{I = 1\} = 1 - \Pr\{I = 0\}. \tag{5}$$

2) Properties

Expressions for the unconditional statistics of the $\{X_t\}$ process were derived by Katz and Parlange (1996) in their study of the situation in which the index is actually observed, but still hold when it is hidden. In particular, the variance of total precipitation can be expressed as

$$\begin{aligned}\mathrm{Var}[S(T)] \approx T\{(1-w)\mathrm{Var}[S(T)|I=0] + w\mathrm{Var}[S(T)|I=1] \\ + Tw(1-w)(E[S(T)|I=1] \\ - E[S(T)|I=0])^2\}.\end{aligned} \tag{6}$$

Here the conditional mean and variance of monthly total precipitation, $E[S(T)|I = i]$ and $\mathrm{Var}[S(T)|I = i]$, are obtained through the substitution of the parameters for the conditional chain-dependent process, given index state $I = i$, into (4).

It is evident in (6) that the monthly variance for the mixture of two conditional chain-dependent processes is not simply a weighted average of the two conditional monthly variances, but includes the variation in the conditional monthly means as well [i.e., second term on right-hand side of (6)]. In this way, the mixture model is capable of increasing the monthly variance and, consequently, reducing overdispersion. As pointed out by Katz and Parlange (1998), this overdispersion can be attributed to several possible factors, including overdispersion in the monthly number of wet days. The present approach allows for this particular possibility by permitting the two transition probabilities of the Markov chain model for the occurrence process to vary between the two hidden states. Similar comments apply to the intensity component of the chain-dependent process for precipitation.

It might be natural to presume that such a mixture model would require that the unconditional distribution of monthly total precipitation be bimodal. But this feature would not be present unless the differences in the parameters of the two conditional chain-dependent processes were sufficiently large. For instance, in the simpler situation of a mixture of two conditional normal distributions, the resultant unconditional distribution would still be unimodal unless the two conditional means were far enough apart relative to the two conditional standard deviations (Johnson and Kotz 1970, 87–92). Nevertheless, in the atmospheric sciences literature, most searches for evidence of multiple regimes have focused on multimodality (e.g., Hansen and Sutera 1995; Nitsche et al. 1994).

3. Parameter estimation method

Parameter estimation for a mixture of two conditional chain-dependent processes would be straightforward if the index were actually observed. In this case, the estimation problem can be separated in two subsets (i.e., by classifying the daily precipitation time series according to which index state occurs in a given year). Then conventional maximum likelihood techniques can be applied to fit a chain-dependent process to each subset individually (e.g., as in Katz and Parlange 1993). However, when the index is hidden, the likelihood function is sufficiently complex that direct maximization is infeasible. In particular, even for just a mixture of two conditional normal distributions (effectively, the mixture model for the intensity component of the precipitation process), iterative numerical techniques are required to obtain maximum likelihood estimates (e.g., McLachlan and Krishnan 1997).

The EM algorithm (Dempster et al. 1977) is an iterative numerical technique to obtain maximum likelihood estimates, with the basic idea being to exploit the relative simplicity of likelihood maximization in the ‘‘complete-data’’ situation (i.e., if the index were observed). The situation actually faced is termed ‘‘incomplete data,’’ because the index state is regarded as ‘‘missing.’’ The ‘‘E’’ (for expectation) step of the EM algorithm involves replacing the unobservable complete-data likelihood function with its conditional expectation. Then updated parameter estimates can be obtained (the ‘‘M’’ or maximization step of the EM algorithm) in essentially as simple a manner as for the complete-data situation. Navidi (1997) provides a heuristic explanation of how the EM algorithm works, whereas McLachlan and Krishnan (1997) give an in-depth treatment. Making use of their conditional independence, it might have been anticipated that the occurrence and intensity components could be treated separately. But this simplification is not possible, because both components provide evidence about the likelihood of a particular hidden state having occurred during a given year. In other words, although the two components can be treated separately in the M step of the algorithm, they must be treated simultaneously in the E step.

a. Parameter estimation for chain-dependent process

First parameter estimation is reviewed for a single chain-dependent process, as these results will be utilized in formulating the EM algorithm. Suppose that an observed time series of daily precipitation for a single year $\{x_t : t = 1, 2, \ldots, T\}$ is generated from a chain-dependent process with parameters P_{01}, P_{11}, μ^*, and $(\sigma^*)^2$. Treating the first observation x_1 as if it were fixed, the exact maximum likelihood estimates of these parameters are given by

$$\hat{P}_{j1} = n_{j1}/n_{j\cdot}, \qquad j = 0, 1, \qquad \hat{\mu}^* = s_1/n_{\cdot 1},$$
$$(\hat{\sigma}^*)^2 = s_2/n_{\cdot 1} - (\hat{\mu}^*)^2, \tag{7}$$

where

$$s_1 = \sum_t x_t^*, \qquad s_2 = \sum_t (x_t^*)^2, \qquad n_{j\cdot} = n_{j0} + n_{j1},$$
$$j = 0, 1 \tag{8}$$

(e.g., Katz and Parlange 1993). Here the transition count n_{jk} denotes the number of times that the Markov chain for precipitation occurrence makes a transition from state j to state k in the sample (e.g., n_{11} denotes the number of times that a wet day is followed by a wet day). Note that the total number of transitions is $n_{00} + n_{01} + n_{10} + n_{11} = T - 1$, and that the two summations appearing in (8) are only taken over those days $t = 2, 3, \ldots, T$ that are wet (i.e., $n_{\cdot 1} = n_{01} + n_{11}$ terms). An expression for the corresponding likelihood function of a single chain-dependent process is included in appendix A [see (A1)].

b. Complete-data likelihood function

Now the complete-data likelihood function for a mixture of two conditional chain-dependent processes is considered. The daily precipitation data consist of M time series, each of length T days. Let $x_t(m)$, $t = 1, 2, \ldots, T$; $m = 1, 2, \ldots, M$, denote the observed precipitation amount on the tth day of the mth year, with $i(m)$ denoting the known index state for the mth year. Also let the vector of parameters be denoted by $\boldsymbol{\theta} = (w, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$, where

$$\boldsymbol{\theta}_i = [P_{01}(i), P_{11}(i), \mu_i^*, (\sigma_i^*)^2], \qquad i = 0, 1 \tag{9}$$

are the parameter vectors for the individual conditional chain-dependent processes. It is convenient to work with log likelihood, and the logarithm of the complete-data likelihood function is related to those for the two individual processes by

$$\ln L_C(\boldsymbol{\theta}) = \sum_m \{[1 - i(m)] \ln[(1 - w)L_m(\boldsymbol{\theta}_0)]$$
$$+ i(m) \ln[wL_m(\boldsymbol{\theta}_1)]\}. \tag{10}$$

Here $L_m(\boldsymbol{\theta}_i)$ denotes the likelihood function (A1), with parameter vector corresponding to index state i, evaluated for the daily precipitation time series in the mth year. Note that the first term (second term) on the right-hand side of (10) only appears if $i(m) = 0$ [$i(m) = 1$]. The simple form of the likelihood function is evident, as (10) reduces to a separate sum for each chain-dependent process.

The maximum likelihood estimates for the model parameters are given by

$$\hat{w} = \Big[\sum_m i(m)\Big]\Big/M,$$
$$\hat{P}_{j1}(1) = \Big[\sum_m i(m)n_{j1}(m)\Big]\Big/\Big[\sum_m i(m)n_{j\cdot}(m)\Big], \qquad j = 0, 1,$$
$$\hat{\mu}_1^* = \Big[\sum_m i(m)s_1(m)\Big]\Big/\Big[\sum_m i(m)n_{\cdot 1}(m)\Big],$$
$$(\hat{\sigma}_1^*)^2 = \Big[\sum_m i(m)s_2(m)\Big]\Big/\Big[\sum_m i(m)n_{\cdot 1}(m)\Big] - (\hat{\mu}_1^*)^2. \tag{11}$$

Here the "(m)" appended to a statistic, defined in section 3a for a single chain-dependent process, indicates that it is now calculated for the time series in the mth year, and the summations in (11) all range over $m = 1, 2, \ldots, M$. Expressions for the corresponding estimators of $P_{j1}(0)$, μ_0^*, and $(\sigma_0^*)^2$ can be obtained by replacing $i(m)$ with $1 - i(m)$ in (11). These estimates in (11) correspond to those for a single chain-dependent process fit to a subset of years, as categorized by the index state (see Katz and Parlange 1993). With a slight modification, these same expressions will be utilized in the EM algorithm.

c. EM algorithm

When the state of the index is actually hidden, the incomplete-data log likelihood function can be expressed as

$$\ln L(\boldsymbol{\theta}) = \sum_m \ln[(1 - w)L_m(\boldsymbol{\theta}_0) + wL_m(\boldsymbol{\theta}_1)]. \tag{12}$$

Here the likelihood function $L_m(\boldsymbol{\theta}_i)$ is the same as the one that appears on the right-hand side of (10). Contrasting (12) with (10), the complex structure of the actual likelihood function is evident, as each term in the sum involves both conditional chain-dependent processes. Recalling that the terms $L_m(\boldsymbol{\theta}_i)$, $i = 0, 1$, in (12) are given by (A1), it is clear that direct maximization is not practicable.

1) E STEP

At any stage of the EM algorithm, the E step involves the determination of the conditional expectation of the complete-data likelihood function (10) given the observed data. Because (10) is linear in the unobservable index states, this conditional expectation simply requires the calculation of the posterior probability of the index state for each year given the present parameter estimates. By Bayes's theorem,

$$\Pr\{I(m) = 1\} = wL_m(\boldsymbol{\theta}_1)/[(1 - w)L_m(\boldsymbol{\theta}_0) + wL_m(\boldsymbol{\theta}_1)], \tag{13}$$

$m = 1, 2, \ldots, M$. In practice, the likelihood functions that appear in (13) are evaluated at the parameter estimates produced by the most recent M step of the EM algorithm (see description of M step below). The posterior probabilities, corresponding to the parameter estimates obtained at the final stage of the algorithm, can also be used to relate the hidden index to known measures of atmospheric circulation (see section 4). As remarked earlier, these posterior probabilities of the hid-

TABLE 1. Model selection for mixture of two conditional chain-dependent processes fit to time series of January daily precipitation at Chico, CA (78 yr).

Model	Number parameters	Log likelihood	AIC	BIC
Completely constrained	4	−1924.246	3856.49	3865.92
$P_{01}(0) = P_{01}(1)$, $\sigma_0^* = \sigma_1^*$	7	−1912.393	3838.79*	3855.28*
Completely unconstrained	9	−1912.386	3842.77	3863.98

* Denotes minimum.

den states reflect contributions from both the occurrence and intensity components of the chain-dependent processes.

2) M STEP

The M step of the EM algorithm involves replacing the hidden index state $i(m)$ with the posterior probability $\Pr\{I(m) = 1\}$ [determined in the E step by (13)] in the expressions (11) for the estimates if the index were observed. To start the algorithm, initial values are required for the model parameters. Then the E and M steps are repeated alternately until convergence [i.e., maximizing the incomplete-data log likelihood function (12)]. In other words, the parameter estimates are just weighted analogs of the observed index case, with these weights being revised at each stage of the algorithm. Details about the implementation of the EM algorithm are provided in appendix B.

4. Results

Two time series of daily precipitation amounts are considered, one for January at Chico, California, and another for July at Napier, New Zealand. Both of these sites are roughly 40° away from the equator, and in each case the time period is midwinter. Because the large-scale atmospheric circulation is known to exert a major influence on local precipitation patterns in winter in California (e.g., Cayan and Peterson 1989), the Chico example can be viewed as somewhat confirmatory in nature. So this example is somewhat unrealistic, in the sense that it fails to take into account known information about circulation influences. In contrast, New Zealand precipitation appears to be only weakly related to those indexes of atmospheric circulation that have been constructed so far (Sallinger 1980; Tait and Fitzharris 1998; Trenberth 1976). So the Napier example can be regarded as somewhat exploratory in nature, a more realistic application.

The signal of the hidden index may be relatively weak, suggesting that various constraints on the model parameters should be considered to make the modeling approach more parsimonious. In this context, the two competing stochastic models defined in section 2 could be termed the ‘‘completely unconstrained’’ model (i.e., a mixture of two conditional chain-dependent processes with all parameters differing) and the ‘‘completely constrained’’ model (i.e., a single unconditional chain-dependent process). As additional candidate models, constraints will be imposed on either of the two transition probabilities [i.e., either $P_{01}(0) = P_{01}(1)$ or $P_{11}(0) = P_{11}(1)$] and on the transformed intensity variance (i.e., $\sigma_0^* = \sigma_1^*$), with how these constraints are implemented within the EM algorithm being described in appendix B.

To compare the fit of these candidate models, Akaike's information criterion (AIC) (Akaike 1974) and the Bayesian information criterion (BIC) (Schwarz 1978) are employed. Both of these criteria involve penalizing the maximized likelihood function for the number of parameters estimated:

$$\mathrm{AIC}(l) = -2\ln\hat{L}_l + 2K_l,$$
$$\mathrm{BIC}(l) = -2\ln\hat{L}_l + K_l\ln M. \qquad (14)$$

Here $\hat{L}_l$ denotes the maximized likelihood function for the lth model, with this model requiring the estimation of K_l parameters. The model with minimum AIC or BIC is ordinarily selected, although we use these criteria as guidelines rather than rigid rules. For the particular form of stochastic model being fit in our application, the appropriate sample size in the penalty term for the BIC is not completely clear (Kass and Wasserman 1995). In (14), the relevant sample size has been treated as M years, as opposed to MT days.

a. Chico

The January dataset at Chico has a length of 78 yr (i.e., $M = 78$ and $T = 31$ days) during the period 1907–88, with several years having been eliminated because of missing observations. These data have been previously analyzed by Katz and Parlange (1993, 1996, 1998). The month of January is in the midst of a marked wet season, with a substantial fraction of the variance of January (or winter) total precipitation being associated with the contemporaneous mean sea level pressure (SLP) over the adjacent Pacific Ocean (Cayan and Peterson 1989). Katz and Parlange (1993) found that some of the parameters of a chain-dependent process ought to be varied, depending on whether the mean January SLP at 40°N, 130°W is above or below normal (i.e., an observed index with two states).

Tables 1 and 2 summarize the results of fitting a mixture of two conditional chain-dependent processes to the January daily precipitation data at Chico. As in Katz

TABLE 2. Parameter estimates for mixture of two conditional chain-dependent processes fit to time series of January daily precipitation at Chico, CA. Model estimated and observed interannual standard deviation of monthly total precipitation also included.

Model	$\hat{w}$	$\hat{P}_{01}(i)$, $i = 0, 1$	$\hat{P}_{11}(i)$, $i = 0, 1$	$\hat{\mu}_i^*$, $i = 0, 1$ (mm$^{1/4}$)	$\hat{\sigma}_i^*$, $i = 0, 1$ (mm$^{1/4}$)	Std dev of total (mm)
Completely constrained	—	0.2109	0.5705	1.7020	0.5212	70.41
		0.2109	0.5705	1.7020	0.5212	
$P_{01}(0) = P_{01}(1)$, $\sigma_0^* = \sigma_1^*$	0.371	0.2109	0.5048	1.5848	0.5032	89.84
		0.2109	0.6595	1.8590	0.5032	
Completely unconstrained	0.368	0.2142	0.5054	1.5872	0.5043	88.86
		0.2047	0.6620	1.8614	0.5020	
Observed	—	—	—	—	—	88.63

and Parlange (1993), a power transformation of $p = ¼$ is used to account for skewness in the distribution of intensity for both the unconditional and conditional chain-dependent processes [see (3)]. The various models mentioned earlier were fitted, with the model selection statistics only being included in Table 1 for the completely constrained model, the completely unconstrained model, and the optimal model [i.e., with constraints $P_{01}(0) = P_{01}(1)$ and $\sigma_0^* = \sigma_1^*$, according to both the AIC and BIC]. It may be difficult to distinguish among the constrained models, but the Chico data do provide clear support for the presence of a hidden mixture.

Table 2 includes the maximum likelihood estimates of the parameters for the same three models as listed in Table 1. For the optimal model, the hidden state $I = 1$ is associated with wetter weather in all respects. By (2), the estimated conditional probabilities of a wet day are $\hat{\pi}_0 = 0.299$ and $\hat{\pi}_1 = 0.383$, with the estimated persistence parameters being $\hat{d}_0 = 0.294$ and $\hat{d}_1 = 0.449$. The estimated conditional means and standard deviation of the transformed intensity can be converted into the corresponding untransformed statistics (Katz 1999), $\hat{\mu}_0 = 10.3$, $\hat{\mu}_1 = 17.4$, $\hat{\sigma}_0 = 12.2$, and $\hat{\sigma}_1 = 17.8$ mm (note that these two conditional standard deviations differ, despite the equality of the transformed conditional standard deviations). These effects on the daily precipitation parameters can be translated into corresponding shifts in the conditional distribution of monthly total precipitation between the two hidden states, with the conditional monthly means being 95.5 and 206.2 mm and the standard deviations 51.4 and 97.9 mm [using (4)]. They resemble the effects found for an observed SLP index (Katz and Parlange 1993).

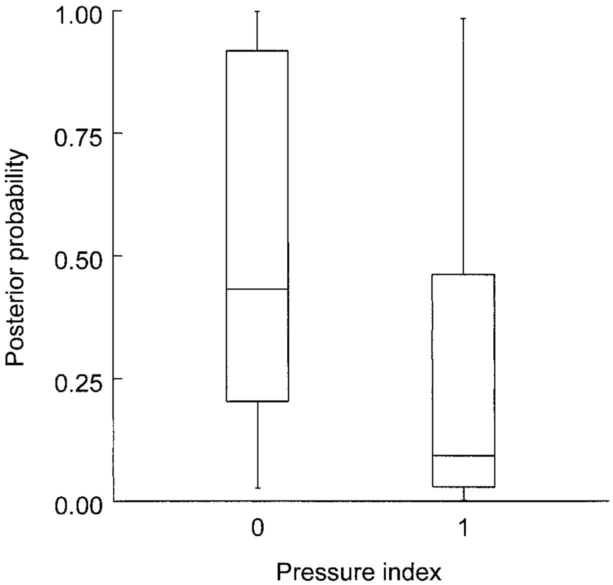

FIG. 1. Box plots (showing minimum, lower quartile, median, upper quartile, and maximum) of conditional distribution of posterior probability of hidden state $I = 1$ (for optimal mixture model of Chico daily precipitation) given whether January mean SLP at 40°N, 130°W is below (indicated by 0) or above (indicated by 1) average.

Regarding the overdispersion phenomenon, Table 2 also includes the estimated standard deviations of January total precipitation at Chico for the three models. The single unconditional chain-dependent process substantially underestimates the observed interannual standard deviation (by roughly 37% in terms of variance; Katz and Parlange 1993), in part because of substantial overdispersion in the monthly number of wet days (Katz and Parlange 1996, 1998). In contrast, both the optimal and completely unconstrained models essentially reproduce the observed value. This effect on overdispersion is comparable to that obtained for Chico when daily precipitation is conditioned on an observed SLP index instead (Katz and Parlange 1993).

As a partial confirmation of the validity of this approach, the time series of January posterior probabilities of the hidden index state (produced by the EM algorithm) for the optimal mixture model is compared to the corresponding time series of January mean SLP from which the observed index of Katz and Parlange (1993) is derived. Figure 1 shows box plots of the conditional distribution of the posterior probability of hidden state $I = 1$, given whether SLP is below or above average. A marked shift in the median probability is evident, along with a corresponding change in variability (as measured by the interquartile range). This result is indicative of some connection between the hidden index and the observed SLP. Given that other features of large-scale atmospheric circulation surely affect precipitation at Chico, a closer linkage with SLP should not necessarily have been anticipated.

TABLE 3. Model selection for mixture of two conditional chain-dependent processes fit to time series of July daily precipitation at Napier, New Zealand (89 yr).

Model	Number parameters	Log likelihood	AIC	BIC
Completely constrained	4	−2449.419	4906.84	4916.79
$P_{01}(0) = P_{01}(1)$	8	−2427.090	4870.18	4890.09
$P_{11}(0) = P_{11}(1)$	8	−2425.588	4867.18*	4887.09*
Completely unconstrained	9	−2425.181	4868.36	4890.76

* Denotes minimum.

b. Napier

The July dataset at Napier has a length of 89 yr (i.e., $M = 89$ and $T = 31$ days) during the period 1896–1994, likewise with several years having been discarded due to missing observations. In part, the Napier dataset was selected because of this relatively long record. One of the indices of large-scale atmospheric circulation in the New Zealand region is the so-called Z1 index (Trenberth 1976). A measure of the strength of the zonal circulation over New Zealand, this index is the difference in monthly mean SLP anomalies between Auckland and Christchurch, New Zealand. Here the pressure anomalies are calculated by subtracting the corresponding 1951–80 means. Precipitation over New Zealand has at best a relatively weak relationship to indexes such as Z1 (Sallinger 1980).

Tables 3 and 4 summarize the results of fitting a mixture of two conditional chain-dependent processes to the July daily precipitation data at Napier. As for Chico, a power transformation of $p = 1/4$ is applied to all of the daily intensity distributions. Besides the completely constrained model, the completely unconstrained model, and the optimal model [i.e., with the constraint $P_{11}(0) = P_{11}(1)$, according to both the AIC and BIC], another model [with the constraint $P_{01}(0) = P_{01}(1)$, termed the ‘‘preferred’’ model] is also included in Table 3. The model selection statistics in Table 3 indicate support for the existence of a hidden mixture, with the preferred model (although suboptimal according to both the AIC and BIC) still being superior to the completely constrained model. In the subsequent discussion, the results are presented primarily for the preferred model, but do not differ substantially for the optimal one.

Table 4 gives the maximum likelihood estimates of the parameters for the same four models as listed in Table 3. For the optimal and completely unconstrained models, the interpretation of the hidden state $I = 1$ is somewhat complex, in some respects corresponding to wetter conditions, in other respects drier. In contrast, the preferred model has the simpler property that the hidden state $I = 1$ is associated with wetter weather in all respects. The estimated effects on daily precipitation occurrence, according to the preferred model, are relatively small: $\hat{\pi}_0 = 0.384$ versus $\hat{\pi}_1 = 0.406$ and $\hat{d}_0 = 0.265$ versus $\hat{d}_1 = 0.305$. On the other hand, the estimated effects on daily intensity are substantial: $\hat{\mu}_0 = 3.1$ versus $\hat{\mu}_1 = 8.2$ mm and $\hat{\sigma}_0 = 3.6$ versus $\hat{\sigma}_1 = 10.6$ mm. The corresponding shifts in the distribution of monthly total precipitation are relatively larger than for Chico, with the conditional monthly means being 37.3 and 103.3 mm and the standard deviations 16.7 and 48.6 mm.

Turning to the overdispersion phenomenon, the estimated standard deviations of July total precipitation at Napier for the four models are included in Table 4. To an even greater extent than for Chico, the single chain-dependent process underestimates the observed interannual standard deviation (by roughly 46% in terms of variance). Unlike for Chico, the mixture model does not eliminate, only substantially reduces the extent of the overdispersion, with the preferred model being apparently the best in this respect (roughly 11% underestimation in terms of variance).

TABLE 4. Parameter estimates for mixture of two conditional chain-dependent processes fit to time series of July daily precipitation at Napier, New Zealand. Model estimated and observed interannual standard deviation of monthly total precipitation also included.

Model	$\hat{w}$	$\hat{P}_{01}(i)$ $i = 0, 1$	$\hat{P}_{11}(i)$, $i = 0, 1$	$\hat{\mu}_i^*$, $i = 0, 1$ (mm$^{1/4}$)	$\hat{\sigma}_i^*$, $i = 0, 1$ (mm$^{1/4}$)	Std dev of total (mm)
Completely constrained	—	0.2821	0.5771	1.3916	0.4976	40.17
		0.2821	0.5771	1.3916	0.4976	
$P_{01}(0) = P_{01}(1)$	0.750	0.2821	0.5469	1.1836	0.3655	51.51
		0.2821	0.5873	1.4622	0.5163	
$P_{11}(0) = P_{11}(1)$	0.782	0.3411	0.5771	1.1766	0.3632	48.14
		0.2667	0.5771	1.4580	0.5144	
Completely unconstrained	0.769	0.3383	0.5497	1.1811	0.3665	49.22
		0.2662	0.5860	1.4608	0.5152	
Observed		—	—	—	—	54.50

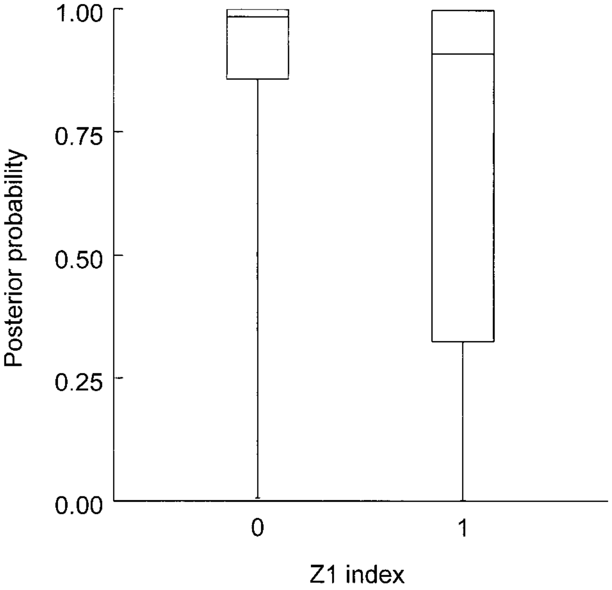

FIG. 2. Box plots of conditional distribution of posterior probability of hidden state $I = 1$ (for preferred mixture model of Napier daily precipitation) given whether July Z1 pressure index is negative (indicated by 0) or positive (indicated by 1).

To explore whether the hidden index has any physical interpretation, the time series of July posterior probabilities of the index state for the preferred mixture model is compared to the corresponding time series of the July Z1 pressure index. Figure 2 shows box plots of the conditional distribution of the posterior probability of the hidden state $I = 1$, given whether the Z1 index is negative or positive (because of the high degree of skewness, the upper quartile and maximum are virtually indistinguishable). A slight shift in the median probability is evident, along with a considerably larger change in variability. The corresponding conditional distribution of the Z1 index given the posterior probability (not shown) also suggests at least a weak relationship. Given that the link between the Z1 index and July total precipitation at Napier is itself relatively weak (a correlation of about -0.4), any more than a weak relationship between Z1 and the hidden index would be unexpected.

5. Discussion

The method introduced for fitting a hidden mixture of two conditional chain-dependent processes to time series of daily precipitation amount is a direct means of providing evidence of the presence of low-frequency modes of variation. The question remains of how to reconcile these results with those obtained from the more traditional approach of fitting increasingly complex stochastic models for high-frequency variations of daily precipitation (e.g., Katz and Parlange 1998). If a source of low-frequency variation were actually present, then the resultant unconditional stochastic model for daily precipitation would actually resemble a more complex form of chain-dependent process [e.g., higher than first-order Markov chain model for precipitation occurrence; Katz and Parlange (1996)]. In this situation, the traditional approach might well result in the erroneous conclusion that a more complex stochastic model is appropriate. Even model selection criteria, such as AIC and BIC, specifically designed to deal with model complexity would not necessarily be robust against the possible presence of such a source of overdispersion (Fitzmaurice 1997).

The proposed method could provide corroboration of the hypothesized existence of potential predictability for precipitation on an interannual timescale. Estimates of such predictability have been based on a generalized analysis of variance approach (Madden et al. 1999). The present approach can be viewed as complementary in the sense that, through a hidden index, it specifies the most likely state of a source of low-frequency variation. Nevertheless, the results of Zheng (1996) suggest that it should not necessarily be relied on to give an alternative estimate of potential predictability. Through the elimination or reduction of overdispersion, this research also has potential application as a technique for improving the performance of stochastic weather generators used to produce scenarios of climate variability and change (Katz 1996).

Several extensions of the methodology developed in the present paper could be considered in future work. One of the limitations of the EM algorithm is that it does not automatically provide standard errors of the parameter estimates as a byproduct (McLachlan and Krishnan 1997). Bayesian methods would be a natural way to generate such information, as well as to quantify the uncertainty in derived statistics, such as estimates of overdispersion. Conditioning simultaneously on both observed and hidden variables would constitute a more realistic treatment of the present situation in climate research (Hughes et al. 1999). Although limited by the length of climate records, it would be natural to allow the hidden index to assume more than two possible states.

Some recent research has dealt with the somewhat analogous problem of detecting hidden sources of high-frequency variation in daily time series of climate variables such as precipitation. One technique relies on ‘‘hidden Markov models,’’ essentially the same approach as treated here, except that the hidden index is permitted to change its state on a daily basis and that the sequence of index states is modeled as a Markov chain (Guttorp 1995). A systematic comparison of these two approaches would be worthwhile.

Acknowledgments. We thank John Sansom for providing New Zealand precipitation data. This research was performed while R. W. Katz was a NIWA Visiting Scientist supported by the Foundation for Research, Science and Technology, and was also partially supported by NSF Grant DMS-9312686 to the NCAR Geophysical Statistics Project. The comments of two anonymous referees are gratefully acknowledged.

APPENDIX A

Likelihood Function for Chain-Dependent Process

The likelihood function (treating the first observation x_1 as if it were fixed) for a single chain-dependent process can be expressed as

$$\begin{aligned} L[P_{01}, P_{11}, \mu^*, (\sigma^*)^2] &= \Big[\prod_j (1 - P_{j1})^{n_{j0}} P_{j1}^{n_{j1}}\Big][2\pi(\sigma^*)^2]^{-(n_{\cdot 1}/2)} \\ &\quad \times \Big\langle \exp\Big\{-\Big[\sum_t (x_t^* - \mu^*)^2\Big]/2(\sigma^*)^2\Big\}\Big\rangle. \end{aligned} \tag{A1}$$

In (A1), the product is over $j = 0, 1$, and the summation is over the same terms t as for the two summations that appear in (8). The first term (in large square brackets) on the right-hand side of (A1) is the likelihood function of the Markov chain model for observed daily precipitation occurrence (e.g., Guttorp 1995, chapter 2), whereas the remaining terms constitute the likelihood function of the normal distribution for observed power transformed daily precipitation intensity.

APPENDIX B

EM Algorithm

The EM algorithm has the desirable property that the value of the likelihood function increases at each stage of the iteration (McLachlan and Krishnan 1997, chapter 3). As with virtually all numerical algorithms for nonlinear optimization, the starting values for the model parameters need to be carefully selected. Otherwise, convergence to a global maximum is not guaranteed. It would be natural to set the parameters equal to the maximum likelihood estimates for the completely constrained model [i.e., a single chain-dependent process fit to the entire dataset; Katz and Parlange (1993)]. However, these parameter values need to be perturbed slightly, differing between the two index states to obtain a nontrivial mixture model.

The EM algorithm was iterated until accuracy of at least four decimal places was obtained for the log likelihood function. This convergence criterion corresponded to at least three-decimal-place accuracy for the parameter estimates $\hat{w}$, $\hat{\mu}_i^*$, and $\hat{\sigma}_i^*$, and at least four for $\hat{P}_{01}(i)$ and $\hat{P}_{11}(i)$. To achieve this degree of accuracy, the number of iterations required ranged from about 40 to 50 for the best fitting constrained models at Chico and Napier to nearly 200 for the completely unconstrained model at Chico.

It is relatively straightforward to modify the EM algorithm to impose constraints on the model parameters. For instance, an equality constraint on individual transition probabilities just requires holding them fixed at each stage of the algorithm (i.e., setting both equal to the corresponding maximum likelihood estimate for a single chain-dependent process). Constraining the transformed intensity variances to be equal is slightly more involved. For each iteration of the algorithm, the unconstrained estimates of the two variances still need to be produced. Then these estimates are combined into a single, ‘‘pooled’’ variance estimate; that is, of the form

$$(1 - \hat{w})(\hat{\sigma}_0^*)^2 + \hat{w}(\hat{\sigma}_1^*)^2.$$

As a check on the performance of the EM algorithm, a limited simulation study was conducted. Synthetic precipitation time series were generated from a mixture of two conditional chain-dependent process, and then the EM algorithm was applied. These simulations were based on actual parameter values that mimicked the estimates obtained for Chico in the case of an observed circulation index (Katz and Parlange 1993). Among other things, it was verified that approximately unbiased estimates of the interannual variance of monthly total precipitation would be obtained.

REFERENCES

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Control,* **19,** 716–723.

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

Cayan, D. R., and D. H. Peterson, 1989: The influence of North Pacific atmospheric circulation on streamflow in the west. *Aspects of Climate Variability in the Pacific and the Western Americas, Geophysical Monogr.,* No. 55, Amer. Geophys. Union, 375–397.

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

Dempster, A. P., N. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc., Series B,* **39,** 1–38.

Fitzmaurice, G. M., 1997: Model selection with overdispersed data. *Statistician,* **46,** 81–91.

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

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

Hansen, A. R., and A. Sutera, 1995: The probability density distribution of the planetary-scale atmospheric wave amplitude revisited. *J. Atmos. Sci.,* **52,** 2463–2472.

Hughes, J. P., P. Guttorp, and S. P. Charles, 1999: A non-homogeneous hidden Markov model for precipitation occurrence. *Appl. Stat.,* **48,** 15–30.

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

Jones, R. H., R. A. Madden, and D. J. Shea, 1995: A new methodology for investigating long range predictability. Preprints, *Sixth Int. Meeting on Statistical Climatology,* Galway, Ireland, University College, 531–534.

Kass, R. E., and L. Wasserman, 1995: A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Stat. Assoc.,* **90,** 928–934.

Katz, R. W., 1996: Use of conditional stochastic models to generate climate change scenarios. *Climate Change,* **32,** 237–255.

——, 1999: Moments of power transformed time series. *Environmetrics,* in press.

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

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

——, and ——, 1998: Overdispersion phenomenon in stochastic modeling of precipitation. *J. Climate,* **11,** 591–601.

Madden, R. A., D. J. Shea, R. W. Katz, and J. W. Kidson, 1999: The potential long-range predictability of precipitation over New Zealand. *Int. J. Climatol.,* **19,** 405–421.

McLachlan, G. J., and T. Krishnan, 1997: *The EM Algorithm and Extensions.* Wiley, 274 pp.

Navidi, W., 1997: A graphical illustration of the EM algorithm. *Amer. Stat.,* **51,** 29–31.

Nitsche, G., J. M. Wallace, and C. Kooperberg, 1994: Is there evidence of multiple equilibria in planetary wave amplitude statistics? *J. Atmos. Sci.,* **51,** 314–322.

Sallinger, M. J., 1980: New Zealand climate: I. Precipitation patterns. *Mon. Wea. Rev.,* **108,** 1892–1904.

Sansom, J., 1995: Rainfall discrimination and spatial variation using breakpoint data. *J. Climate,* **8,** 624–636.

——, 1998: A hidden Markov model for rainfall using breakpoint data. *J. Climate,* **11,** 42–53.

——, and P. J. Thomson, 1992: Rainfall classification using breakpoint pluviograph data. *J. Climate,* **5,** 755–764.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.,* **6,** 461–464.

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.

Tait, A. B., and B. B. Fitzharris, 1998: Relationships between New Zealand rainfall and south–west Pacific pressure patterns. *Int. J. Climatol.,* **18,** 407–424.

Trenberth, K. E., 1976: Fluctuations and trends in indices of the southern hemispheric circulation. *Quart. J. Roy. Meteor. Soc.,* **102,** 65–75.

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

Zheng, X., 1996: Unbiased estimation of autocorrelations of daily meteorological variables. *J. Climate,* **9,** 2197–2203.