NONHOMOGENEOUS MARKOV MODEL FOR DAILY PRECIPITATION

By Balaji Rajagopalan,[1] Upmanu Lall,[2] Member, ASCE, and David G. Tarboton,[3] Member, ASCE

ABSTRACT: This paper presents a one-step nonhomogeneous Markov model for describing daily precipitation at a site. Daily transitions between wet and dry states are considered. The one-step, 2 × 2 transition-probability matrix is presumed to vary smoothly day by day over the year. The daily transition-probability matrices are estimated nonparametrically. A kernel estimator is used to estimate the transition probabilities through a weighted average of transition counts over a symmetric time interval centered at the day of interest. The precipitation amounts on each wet day are simulated from the kernel probability density estimated from all wet days that fall within a time interval centered on the calendar day of interest over all the years of available historical observations. The model is completely data-driven. An application to data from Utah is presented. Wet- and dry-spell attributes [specifically the historical and simulated probability-mass functions (PMFs) of wet- and dry-spell length] appear to be reproduced in our Monte Carlo simulations. Precipitation amount statistics are also well reproduced.

INTRODUCTION

Markov chains (Gabriel and Neumann 1962; Todorovic and Woolhiser 1975; Smith and Schreiber 1974) have been a popular method for modeling daily precipitation occurrence. Typically a two-state (wet or dry), one-step model is used, and the state transition probabilities (e.g., transition from a wet day to a wet day, a wet day to a dry day) are estimated from the data. One problem with such a description is that the transition probabilities may vary over the year, i.e., the process of precipitation occurrence is nonstationary.

Two approaches are commonly used to address this problem. In the first approach, the year is divided into periods (or seasons) and the transition probabilities are estimated separately for each period. There is an implicit assumption that the occurrence process is stationary over the period. This assumption may not be tenable. The second approach is to consider essentially a nonhomogeneous Markov process by allowing the transition probabilities to vary systematically over the year, and to model such a variation through a Fourier-series expansion (Feyerherm and Bark 1965; Woolhiser et al. 1973; Woolhiser and Pegram 1979). This can be an effective approach where adequate data is available, and the seasonality in the precipitation process can be captured by a few Fourier-series terms. Our nonparametric analyses (Rajagopalan and Lall 1995) of the seasonality of precipitation for stations along a meridional transect in the western United States, suggests that sometimes the number of Fourier-series terms needed may be large relative to the amount of data available.

In this paper, a nonhomogeneous Markov (NM) model is presented that uses kernel methods to estimate a nonhomogeneous transition-probability matrix, and to estimate a corresponding nonstationary probability-density function (PDF) of daily precipitation amount. Kernel methods are local, weighted averages of the target function (relative frequency of occurrence in this case). Since they are capable of approximating a wide variety of target functions with asymptotically vanishing error, and use only data from a "small" neighborhood of the point of estimate, they are considered nonparametric. Fourier-series methods are shown to be a subset of kernel methods by Eubank (1988, secs. 3.4 and 4.1). A review of hydrologic applications of nonparametric function estimation methods is provided by Lall (1995).

A brief description of the Markov chain and its terminology is first presented as a background to motivate our formulation. The general structure of the NM model proposed is next outlined with the nonparametric estimators for the transition probabilities. The simulation procedure is then outlined. Results from an application of the model to a precipitation data from Utah follow. Musings on the results and discussion of some limitations of the approach conclude the paper.

BACKGROUND

The basic assumption in a two state Markov-chain model is that the present state (wet or dry) depends only on the immediate past. The transition probabilities for transitions [i.e., wet-wet (WW), wet-dry (WD), dry-wet (DW), dry-dry (DD)] between the two states [wet (W) or dry (D)] are estimated directly from the data through a counting process. Two elements of the transition-probability matrix are the probability of a dry day following a wet day, $P_{WD} = a_1$, and the probability of a wet day following a dry day, $P_{DW} = a_2$. The other probabilities, the probability of a wet day following a wet day, P_{WW}, and the probability of a dry day following a dry day, P_{DD}, are $(1 - a_1)$ and $(1 - a_2)$, respectively.

Seasonal variations in the transition probabilities can be accounted for by expressing the changing transition probabilities through a Fourier series (Woolhiser and Pegram 1979; Roldan and Woolhiser 1982). As an illustration, the transition probability P_{WD} can be expressed as follows:

$$P_{WD}(t) = \bar{P}_{WD} + \sum_{k=1}^{m} c_k \sin(2\pi tk/365 + \theta_k), \quad t = 1, 2, \ldots, 365 \tag{1}$$

where m = maximum number of harmonics required to describe the seasonal variability of the transition probability; $\bar{P}_{WD}$ = annual mean value of the parameter; c_k = amplitude; and θ_k = phase angle in radians for the kth harmonic.

The means, amplitudes, and phase angles are estimated by numerical optimization of the log-likelihood function, as described by Woolhiser and Pegram (1979) and Roldan and Woolhiser (1982). Fourier-series representations of parameters of a first-order Markov chain for precipitation have been used (among others) by Feyerherm and Bark (1965), who

[1]Post Doctoral Res. Sci., Lamont-Doherty Earth Observatory of Columbia Univ., P.O. Box 1000, RT 9W, Palisades, N.Y. 10964-8000.

[2]Prof., Utah State Univ., Utah Water Res. Lab., Logan, UT 84322-8200.

[3]Asst. Prof., Utah State Univ., Utah Water Res. Lab., Logan, UT.

Note. Discussion open until June 1, 1996. To extend the closing date one month, a written request must be filed with the ASCE Manager of Journals. The manuscript for this paper was submitted for review and possible publication on February 10, 1995. This paper is part of the *Journal of Hydrologic Engineering*, Vol. 1, No. 1, January, 1996. ©ASCE, ISSN 1084-0699/96/0001/0033–0040/$4.00 + $.50 per page. Paper No. 10080.

used least-squares techniques for parameter estimation, and by Stern and Coe (1984), who formulated the estimation problem as a generalized linear model to obtain maximum likelihood estimators.

The degree of dependence in time is limited by the order (i.e., the number of past days the present state is presumed to depend on) of the Markov chain (MC). Feyerharm and Bark (1967) and Chin (1977) suggest that the order may need to be seasonally variable as well. Lack of parsimony is a drawback of MC models as the order is increased. A number of researchers (Hopkins and Robillard 1964; Haan et al. 1976; Srikanthan and McMahon 1983; Guzman and Torrez 1985) have also stressed the need for multistate MC models that consider the dependence between transition probabilities and rainfall amount. In this paper, we shall consider only a two-state, first-order Markov chain. Extensions to other situations follow in the same spirit.

MODEL FORMULATION

The NM model that we present allows the one-step transition probability matrix to change over each day, thus capturing the day-to-day variation in the occurrence process in a natural manner. The daily transition-probability matrices are estimated using a discrete kernel estimator, which we describe in the following section. Daily-precipitation-occurrence sequences are then simulated using the transition-probability matrices. To complete the model, precipitation amounts on each wet day are simulated from the nonparametric probability density estimated from all wet days that fall within a time interval or bandwidth centered on the calendar day of interest over all the years of available historical record. The model is data driven, i.e., all parameters are estimated directly from available data.

Transition Probabilities and their Estimation

The precipitation occurrence process is shown in Fig. 1. From the daily precipitation record we can obtain four types of data (for illustration refer to Fig. 1), which are (1) the day indices $t_{w_1}, t_{w_2}, \ldots, t_{w_{nw}}$ of nw wet days; (2) the day indices $t_{d_1}, t_{d_2}, \ldots, t_{d_{nd}}$ of nd dry days; (3) the day indices $t_{wd_1}, t_{wd_2}, \ldots, t_{wd_{nwd}}$ of the nwd days on which a transition occurs from wet to dry, meaning days t_{wd_1}, t_{wd_2}, etc., are wet and days $t_{wd_1} + 1, t_{wd_2} + 1$, etc., are dry; and (4) the day indices $t_{dw_1}, t_{dw_2}, \ldots, t_{dw_{ndw}}$ of the ndw days on which a transition occurs from dry to wet, meaning days t_{dw_1}, t_{dw_2}, etc., are dry and days $t_{dw_1} + 1, t_{dw_2} + 1$, etc., are wet. A day index refers to a number between 1 and 366, representing the calendar day of the observation. From these we estimate the transition probabilities $P_{wd}(t)$ (probability of transition from a wet day on calendar day t to a dry day on calendar day $t + 1$) and $P_{dw}(t)$ (probability of transition from a dry day on calendar day t to a wet day on calendar day $t + 1$). The other two transition probabilities [namely $P_{ww}(t)$ and $P_{dd}(t)$] can be estimated directly from the relations $P_{wd}(t) + P_{ww}(t) = 1$ and $P_{dw}(t) + P_{dd}(t) = 1$. The transition probabilities for calendar day t are estimated from the data using discrete nonparametric kernel estimators.

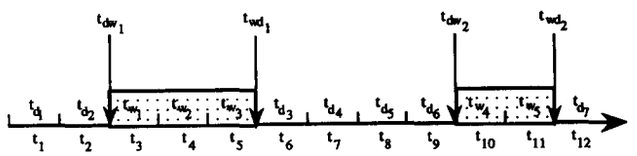

FIG. 1. Precipitation-Occurrence Process ($t_1, t_2, \ldots$ = Day Indices; $t_{w_1}, t_{w_2}, \ldots$ = Wet-Day Indices; $t_{d_1}, t_{d_2}, \ldots$ = Dry-Day Indices; $t_{dw_1}, t_{dw_2}, \ldots$ = Day Indices of Transition from Dry Day to Wet Day; and $t_{wd_1}, t_{wd_2}, \ldots$ = Day Indices of Transition from Wet Day to Dry Day)

For a traditional Markov chain, the transition probabilities are estimated simply as the ratio of the number of transitions in the historical record to the number of wet or dry days in the historical record, as appropriate. Here we try to localize such estimates about the calendar day of interest using kernel estimators. The general idea is that the events (i.e., a wet or dry day, or a state transition) occurring near the calendar day of interest should be given more weightage, and the ones further away should be given a lower weightage. The resulting kernel estimators for the transition probabilities $P_{wd}(t)$ and $P_{dw}(t)$ are given as follows:

$$\hat{P}_{wd}(t) = \frac{\sum_{i=1}^{nwd} K\left(\frac{t - t_{wd_i}}{h_{wd}}\right)}{\sum_{i=1}^{n_w} K\left(\frac{t - t_{w_i}}{h_{wd}}\right)} \tag{2}$$

$$\hat{P}_{dw}(t) = \frac{\sum_{i=1}^{ndw} K\left(\frac{t - t_{dw_i}}{h_{dw}}\right)}{\sum_{i=1}^{nd} K\left(\frac{t - t_{d_i}}{h_{dw}}\right)} \tag{3}$$

where n_{wd} = number of transitions in the historical record from wet day to dry day; n_{dw} = number of transitions in the historical record from dry day to wet day; n_d = number of dry days in the historical record; n_w = number of wet days in the historical record; $K(\,)$ = kernel function (or weight function); $h_{(\,)}$ = a kernel bandwidth; t = calendar day of interest; and the $t_{(\,)}$s have the definitions described earlier. Note that the estimates on any calendar day t are obtained by using the information from days in the range $[t - h_{(\,)}, t + h_{(\,)}]$. Also, the definition of calendar dates is periodic, i.e., day 365 and day 1 are recognized as one day apart for a non-leap year. The contribution to the estimate of an event that lies within this range is determined by the kernel or weight function $K(\,)$, which is described below.

Since we have a discrete situation (i.e., each day being discrete) we use the discrete kernel developed by Rajagopalan and Lall (1995) as

$$K(x) = \frac{3h}{(1 - 4h^2)}(1 - x^2) \quad \text{for} \quad |x| \le 1 \tag{4}$$

where $x = (t - t_{(\,)})/h_{(\,)}$ measures how far an event $t_{(\,)}$, which lies within a bandwidth $h_{(\,)}$ of the day t, is from t; and $h_{(\,)}$ = an integer.

The kernel in (3) was derived from the consideration that the sum of all weights ascribed to events that lie within a bandwidth $h_{(\,)}$ of t, sum to 1, i.e., $\Sigma_{x=-1}^{1} K(x) = 1$; that the weights be symmetric on either side of t, i.e., $\Sigma_{x=-1}^{1} xK(x) = 0$; that each weight be positive; and that the resulting estimates of probability have minimum mean square error.

The estimators in (2) and (3) are fully defined once the respective bandwidths are specified. We choose the bandwidth using a least squared cross validation (LSCV) procedure (Scott 1992, p. 225), where the bandwidth is chosen that minimizes a LSCV function, which is given as follows:

$$\text{LSCV}(h) = \frac{1}{n}\sum_{i=1}^{n}[1 - \hat{P}_{-t_i}(t_i)]^2 \tag{5}$$

where $\hat{P}_{-t_i}(t_i)$ = estimate of the transition probability ($\hat{P}_{wd}$ or $\hat{P}_{dw}$) on day t_i dropping the information on day t_i; and n = number of observations (n_{dw} or n_{wd}). Here we assume a prior probability of transition to be 1 on the days on which transitions have occurred, hence the 1 in (5). The bandwidth

is searched from 1 to 182 (length of half a year). Once the transition probabilities are estimated for each day in the historical record, the simulation of the precipitation occurrence for each day using the transition-probability matrix of the previous day is possible.

Precipitation-Amount Generation

Precipitation amounts for the wet days are generated from a kernel probability density estimated from all wet days that fall within a time interval or bandwidth centered on the calendar day of interest over all the years of historical record. This amounts to two steps: (1) choosing the time interval or bandwidth; and (2) generating from the kernel estimated PDF.

An appropriate bandwidth for localizing the estimate of the probability density of precipitation amount may be obtained by determining the bandwidth appropriate for estimating the probability that a day is wet. If the probability of daily precipitation is low, the precipitation data will be sparse, and the bandwidth needed for stabilizing the variance of the estimated probability distribution of precipitation will be large. Conversely, as the probability of daily precipitation is high, a large number of days with precipitation will occur and the bandwidth needed to localize the estimate can be smaller.

Consequently, we first consider the smoothing of the proportion of wet days ($p_t = n_t/NT$, where n_t = number of times calendar day t was wet; and NT = total number of calendar days t in the historical record) on each calendar day $t = 1, 2, \ldots, 366$. These raw proportions are smoothed using the discrete kernel (DK) estimator of Rajagopalan and Lall (1995) which in this case is

$$\hat{p}_t = \sum_{j=1}^{366} K\left(\frac{t-j}{h_p}\right) p_j \tag{6}$$

where $K(\,)$ = discrete kernel as defined by (3); and h_p = bandwidth in which we are interested. The bandwidth h_p can be obtained using the LSCV procedure similar to (5) as given by Rajagopalan and Lall (1995) as

$$\text{LSCV}(h_p) = \sum_{t=1}^{366} (\hat{p}_t)^2 - 2\sum_{t=1}^{366} \hat{p}_{-t} p_t \tag{7}$$

where $\hat{p}_{-t}$ = estimate of the calendar day t, by dropping the information on that day.

Once we estimate the time interval h_p, the next step is to pick the precipitation amounts on all the wet days that fall within the time interval h_p from the day of interest in all the years of the historical record. Let us say that the precipitation amounts so picked from the historical records are $y_1, y_2, \ldots, y_{np}$ and $t_1, t_2, \ldots, t_{np}$ are the corresponding calendar-day indices. The task now is to generate precipitation amount for the calendar day t, which is a wet day. This can be accomplished by fitting a conditional PDF $f(y|t)$ [see (10)] and then simulating from it. This step is carried out for each wet day that is simulated. Before describing the simulation procedure, we introduce a kernel density estimator for continuous variables, which is given as follows:

$$\hat{f}(y) = \frac{1}{h_y np}\sum_{i=1}^{np} K_c\left(\frac{y-y_i}{h_y}\right) \tag{8}$$

where $K_c(\,)$ = a univariate, continuous kernel; and h_y = bandwidth. Here we use the Epanechnikov kernel given by

$$K_c(x) = 0.75(1.0 - x^2) \quad \text{for} \quad |x| \le 1$$
$$= 0.0 \tag{9}$$

otherwise, where $x = (y - y_i)/h_y$. For a detailed exposition of kernel density estimation for continuous variables and issues relating to bandwidth selection we refer the reader to Silverman (1986) and Scott (1992), and for kernel-density estimation methods with specific application to precipitation modeling we refer to Lall et al. (1995).

A logarithmic transform of the precipitation data prior to density estimation is often considered. Such a transformation is also attractive in the kernel density estimation (KDE) context because it can provide an automatic degree of adaptability of the bandwidth (in real space). This alleviates the need to choose variable bandwidths with heavily skewed data, and also alleviates problems that the KDE has with PDF estimates near the boundary (e.g., the origin) of the sample space. The resulting KDE can be written as

$$\hat{f}(y) = \frac{1}{np}\sum_{i=1}^{np} \frac{1}{h_{LY} y} K_c\left[\frac{\log(y) - \log(y_i)}{h_{LY}}\right] \tag{10}$$

where h_{LY} = bandwidth of the log transformed data. This is chosen using a recursive approach due to Sheather and Jones (1991) (SJ) to minimize the mean integrated square error (MISE) and is recommended by Lall et al. (1995) for precipitation data.

The two-step procedure just discussed can be considered more formally through the conditional PDF $\hat{f}(y|t)$, defined using a product kernel representation as

$$\hat{f}(y|t) = \frac{1}{y h_{LY}}\sum_{i=1}^{np} K_c\left[\frac{\log(y) - \log(y_i)}{h_{LY}}\right] K\left(\frac{t-t_i}{h_p}\right) \Big/ \sum_{i=1}^{np} K\left(\frac{t-t_i}{h_p}\right) \tag{11}$$

Eq. (11) states that the conditional probability density of a rainfall amount y on calendar day t is obtained by considering a window of width h_p centered at t, weighting the precipitation amounts on wet days that fall within this window using the kernel $K(\,)$, and then forming a density estimate by further weighting these amounts with the kernel $K_c(\,)$. Strictly speaking, the bandwidths h_p and h_{LY} should be chosen by optimizing a criteria relevant to the conditional density. The description of our procedure given earlier shows that we are essentially choosing these bandwidths independently. McLachlan (1992, pp. 306–308) discusses the simultaneous selection of bandwidths in each coordinate versus the use of optimal univariate bandwidths in each direction. It is not clear that the additional effort of simultaneous selection of the two bandwidths is justified. Consequently, we choose the bandwidths h_{LY} and h_p by the methods described for the univariate

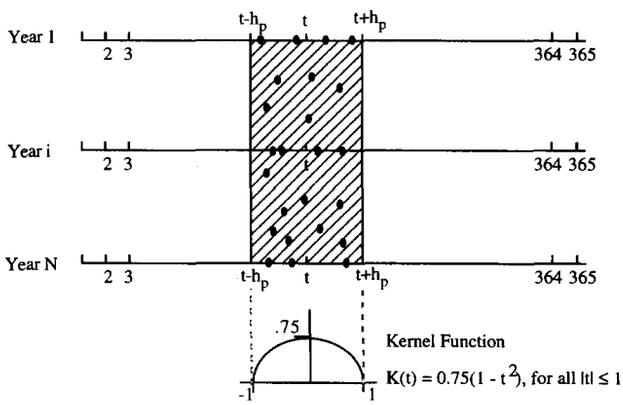

FIG. 2. Precipitation-Amount Generation Process (t = Calendar Day on which Precipitation Is Required; h_p = Time Interval Centered around Calendar Day t; 1, . . . , N = Years in Historical Record; and Thick Dots = Rainy Days in Historical Record. Kernel Function Shown at Bottom Is Used to Weight Rainfall Amounts on each of Rainy Days)

TABLE 1. Statistics of Wet-Day Precipitation for Salt Lake City, Utah; 1961–1991; from Historical Precipitation Record and Averaged over 30 Simulated Precipitation Records

Statistic (1)	Mean wet-day precipitation (cm) (2)	Standard deviation wet-day precipitation (cm) (3)	Fraction of yearly precipitation (4)	Maximum wet-day precipitation (cm) (5)
		(*a*) Season 1		
25% quantile	0.41	0.48	0.23	3.20
Median	0.41	0.51	0.23	3.45
75% quantile	0.43	0.53	0.24	4.04
Historical	0.38	0.43	0.21	2.34
		(*b*) Season 2		
25% quantile	0.48	0.61	0.26	4.42
Median	0.48	0.64	0.27	4.72
75% quantile	0.51	0.66	0.28	5.54
Historical	0.51	0.61	0.28	4.11
		(*c*) Season 3		
25% quantile	0.46	0.69	0.24	4.93
Median	0.46	0.71	0.26	5.84
75% quantile	0.48	0.76	0.26	7.29
Historical	0.46	0.74	0.26	5.79
		(*d*) Season 4		
25% quantile	0.41	0.48	0.24	3.48
Median	0.43	0.53	0.24	4.32
75% quantile	0.46	0.58	0.25	5.49
Historical	0.43	0.48	0.25	3.12
		(*e*) Annual		
25% quantile	0.46	0.61	5.97	—
Median	0.46	0.64	6.48	—
75% quantile	0.48	0.64	8.76	—
Historical	0.43	0.56	5.84	—

case. Rajagopalan et al. (1995) show that bandwidths selected in this way are often satisfactory. For simulation from the kernel estimated PDF [such as in (11)], it is not necessary to explicitly estimate the density $\hat{f}(y|t)$. The estimation of the bandwidths h_{LY} and h_P as well as the subsequent perturbation of the historical data is sufficient.

Simulation Procedure

The simulation procedure from the NM model can be described in the following steps.

1. From the historical precipitation sequence evaluate the transition probabilities [$P_{wd}(t)$, $P_{ww}(t)$, $P_{dw}(t)$, and $P_{dd}(t)$] for each calendar day t using the estimators described in the section on transition probabilities and their estimation. Similarly evaluate the probability density function for precipitation amount on day t using the procedure described in the section on precipitation-amount generation.
2. Start the simulation with a wet or dry day (deciding by generating a uniform random number U in [0, 1], so if $U \leq 0.5$ then the day is wet else it is dry).
3. The precipitation state for the next day is simulated from the transition-probability matrix for the current day (as estimated in step 1).
4. Precipitation amounts on wet days are generated following the process, illustrated in Fig. 2, which is described below:
   a. Pick all the wet day precipitation amounts (e.g., $y_1, y_2, \ldots, y_{np}$) from all the years in the historical record that fall within the window h_p centered on the corresponding calendar day of interest and also the cor-

TABLE 2. Statistics of Wet-Spell Length for Salt Lake City, Utah; 1961–1991; from Historical Precipitation Record and Averaged over 30 Simulated Precipitation Records

Statistic (1)	Mean wet-spell length (days) (2)	Standard deviation wet-spell (days) (3)	Fraction of wet days (4)	Longest wet-spell length (days) (5)
		(*a*) Season 1		
25% quantile	1.89	1.29	0.31	9
Median	1.92	1.37	0.32	10
75% quantile	1.99	1.43	0.33	11.8
Historical	1.86	1.29	0.32	10
		(*b*) Season 2		
25% quantile	1.87	1.27	0.25	8
Median	1.91	1.34	0.25	9
75% quantile	1.95	1.41	0.26	10
Historical	2.12	1.47	0.27	12
		(*c*) Season 3		
25% quantile	1.79	1.23	0.19	8
Median	1.86	1.29	0.20	9
75% quantile	1.91	1.37	0.20	10
Historical	1.60	0.9	0.18	7
		(*d*) Season 4		
25% quantile	1.85	1.27	0.25	8
Median	1.87	1.32	0.26	9
75% quantile	1.92	1.38	0.27	10
Historical	1.97	1.36	0.26	9
		(*e*) Annual		
25% quantile	1.88	1.32	0.26	10
Median	1.91	1.36	0.26	11
75% quantile	1.94	1.39	0.26	13
Historical	1.91	1.31	0.26	12

TABLE 3. Statistics of Dry-Spell Length for Salt Lake City, Utah; 1961–1991; from Historical Precipitation Record and Averaged over 30 Simulated Precipitation Records

Statistic (1)	Mean dry-spell length (days) (2)	Standard deviation dry spell (days) (3)	Fraction of dry days (4)	Longest dry-spell length (days) (5)
		(*a*) Season 1		
25% quantile	3.8	3.5	0.67	23
Median	3.92	3.63	0.68	25
75% quantile	4.0	3.75	0.68	27
Historical	3.91	3.64	0.68	30
		(*b*) Season 2		
25% quantile	5.21	5.64	0.74	39
Median	5.48	5.91	0.75	46
75% quantile	5.59	6.25	0.76	50
Historical	5.5	5.41	0.73	28
		(*c*) Season 3		
25% quantile	6.82	7.12	0.79	44
Median	7.05	7.53	0.80	52
75% quantile	7.26	7.943	0.81	72
Historical	6.87	6.92	0.82	55
		(*d*) Season 4		
25% quantile	4.91	5.47	0.73	38
Median	5.09	5.71	0.74	43
75% quantile	5.28	5.91	0.75	51
Historical	5.21	5.38	0.74	31
		(*e*) Annual		
25% quantile	5.29	6.13	0.74	58
Median	5.41	6.32	0.74	70
75% quantile	5.54	6.67	0.74	86
Historical	5.45	5.99	0.74	61

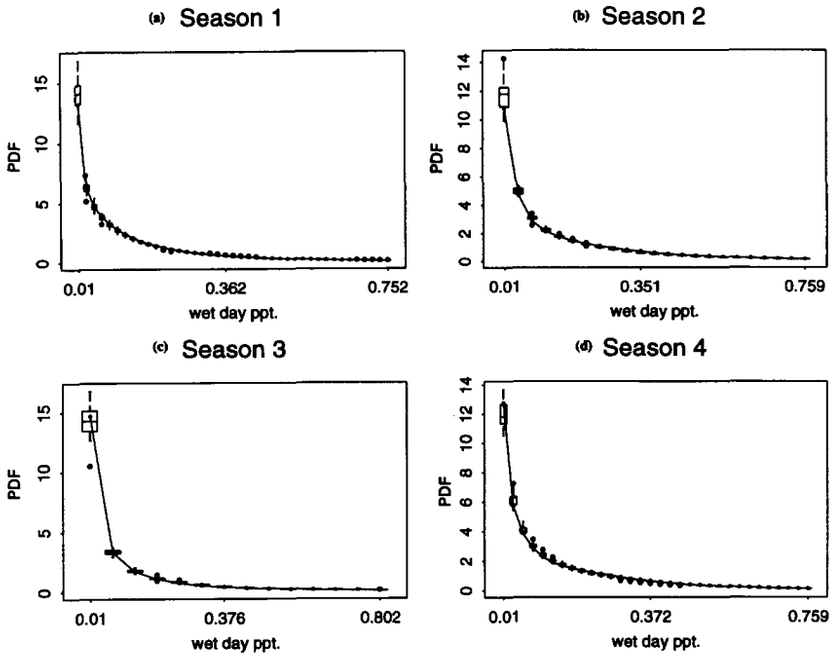

FIG. 3. Boxplots of PDF of Wet-Day Precipitation in Each Season, for Model-Simulated Records along with Historical Values

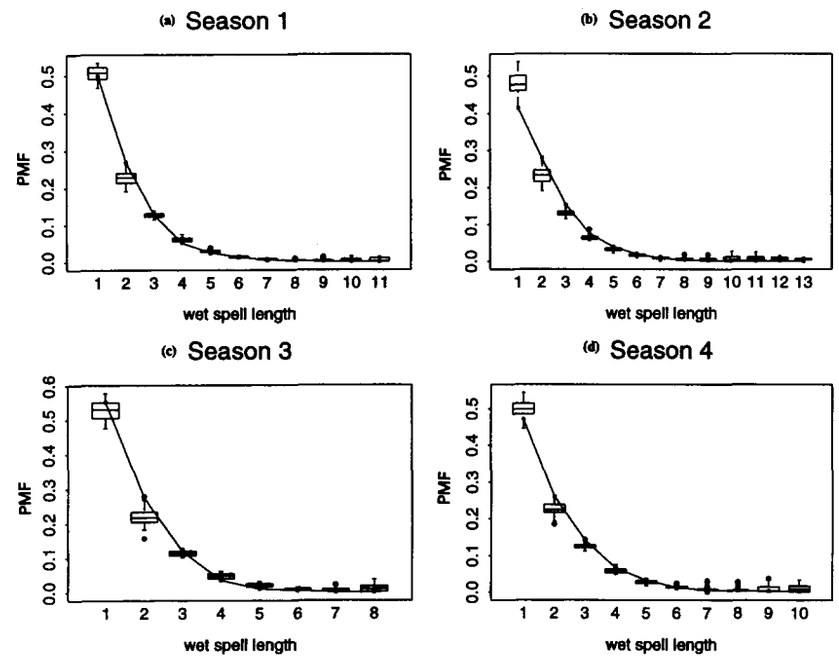

FIG. 4. Boxplots of PMF of Wet-Spell Length in Each Season, for Model-Simulated Records along with Historical Values

responding calendar day indices $t_1, t_2, \ldots, t_{np}$.

b. For the calendar day of interest, pick a historical wet day to perturb using the bandwidth h_p and the kernel $K(x)$ to specify the resampling metric. Recall that the kernel function describes the weight given to each calendar day that lies within h_p of calendar day t, which depend on the "distance" between the two dates relative to the bandwidth h_p, and the kernel function given in (4). Let the weights associated with each of np wet days that are thus identified be $wt_1, wt_2, \ldots, wt_{np}$. Now generate a random integer j between 1 and np from a probability metric given by these weights.

c. The simulated precipitation amount is $y^* = \exp[\log(y_j) + Uh_{LY}]$; where y_j = precipitation on the historical-day point picked to be perturbed. The random variate U is generated from the probability density corresponding to the kernel function $K_c(\ )$. As mentioned earlier, we have used the Epanechnikov kernel in this study, and simulation from this kernel is easily accomplished using the two-step procedure described in Silverman (1986, p. 143)

5. The process (steps 3 and 4) is repeated day by day until the desired length of record is generated.

MODEL APPLICATION

The model described was applied to daily rainfall data from Salt Lake City in Utah. Thirty years of daily weather data

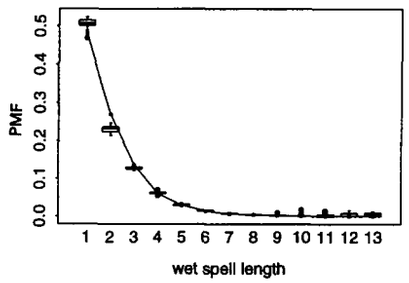

FIG. 5. Boxplots of PMF of Wet-Spell Length over Whole Year, for Model-Simulated Records along with Historical Values

was available from the period 1961–1991. Salt Lake City is at 40°46′ N latitude, 111°58′ W longitude and, at an elevation of 1,288 m (also mean sea level). Most of the precipitation comes in the form of winter snow. Rainfall occurs mainly in spring, with some in fall.

We shall first list some measures of performance that were used to compare the historical record and the model-simulated record, and then outline the experimental design. The aim here is to capture the frequency structure of the events (i.e., the underlying PDF). By "events" we mean the wet-spell length, dry-spell length, and the wet-day precipitation. The wet- and dry-spell lengths are defined as the number of successive wet or dry days. Note that the Markov-chain model considers only transitions from one day to the next, and does not explicitly consider spell statistics. Clearly the wet-spell and dry-spell lengths are defined through the set of integers greater than 1. We look at model performance both at the seasonal scale and the annual scale. For the seasonal-scale comparison we have the year divided into four seasons: winter, or season 1 (January–March); spring or season 2 (April–June); summer or season 3 (July–September); and fall or season 4 (October–December).

Performance Measures

The following statistics are computed on an annual basis and for each seam to judge the performance of the model:

1. Probability-mass function (PMF) of wet-spell length and dry-spell length, and probability-density function of wet-day precipitation.
2. Mean of wet-spell length, dry-spell length, and wet-day precipitation for each.
3. Standard deviation of wet-spell length, dry-spell length, and wet-day precipitation.
4. Length of longest wet spell and dry spell.
5. Maximum wet-day precipitation.
6. Percentage of yearly precipitation in each season.
7. Fraction of wet and dry days.

Experiment Design

Our purpose here is to test the utility of the NM model. The main steps involved in this are:

1. Thirty sets of synthetic records of 30 years each (i.e., the historical-record length) are simulated using the NM model.
2. The statistics of interest are computed for each simulated record, for each season, and are compared to statistics of the historical record using boxplots. The PMFs of wet- and dry-spell length are estimated using the discrete kernel estimator of Rajagopalan and Lall (1995) [same as the estimator in (6)] and the PDFs of the wet-day precipitation is estimated using the estimator in (10). The statistics listed in the section on performance measures are computed for the simulated record and compared with those of the historical record.

RESULTS

In this section, we present comparative results (using the performance measures listed in the section on performance measures) of the NM model for the Salt Lake City data. The PMFs/PDFs of the simulated records are compared with those for the historical record using boxplots, and other statistics are summarized in Tables 1–3. A box in the boxplots (e.g., Fig. 3) indicates the interquartile range of the statistic computed from 30 simulations, the line in the middle of the box indicates the median simulated value. The solid lines corre-

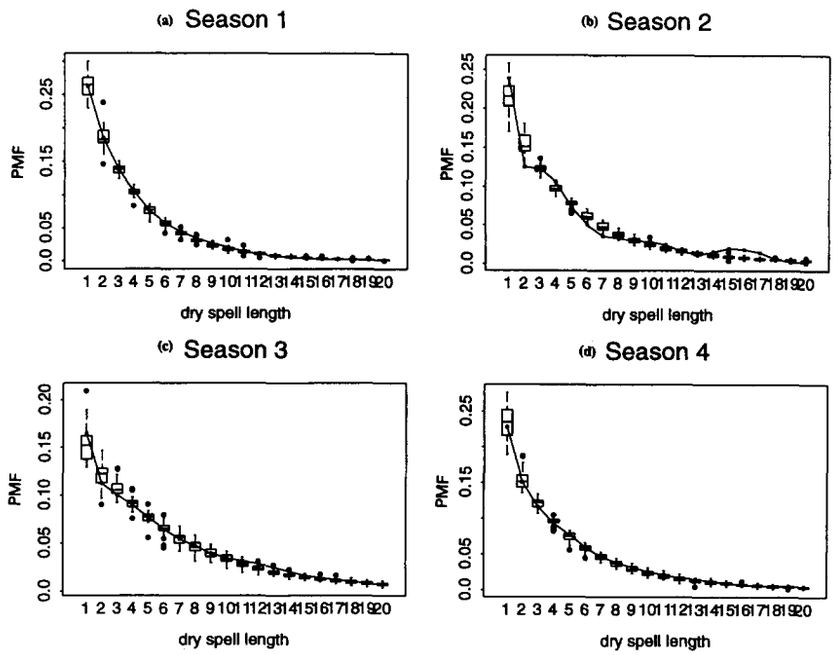

FIG. 6. Boxplots of PMF of Dry-Spell Length in Each Season, for Model-Simulated Records along with Historical Values

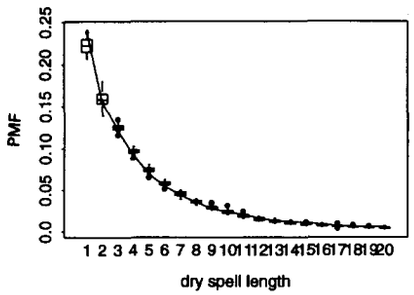

FIG. 7. Boxplots of PMF of Dry-Spell Length over Whole Year, for Model-Simulated Records along with Historical Values

spond to the statistic of the historical record. The boxplots show the range of variation in the statistics from the simulations and also show the capability of the simulations to reproduce historical statistics. The plots of the PDFs are truncated to show a common range across seasons and to highlight differences near the origin (mode).

In each case, the interquartile range across the simulations show the variability in that statistic across simulations. For the 30 simulations considered, one can expect some historical statistics to fall outside the box defined by the interquantile range.

Fig. 3 shows the boxplots of kernel estimated PDFs of simulated data of wet-day precipitation and the historical data. It can be seen that the historical PDFs are very well reproduced by the simulations in all the four seasons. The other statistics are also seen to be well reproduced by the model for all the seasons and also annual, as can be noticed from Table 1.

Boxplots of kernel estimated PMFs of simulated data of wet-spell length are found to enclose the PMF of the historical data of wet-spell length for all the four seasons in Fig. 4 and for the annual in Fig. 5. The other statistics are also preserved quite well by the simulations, as seen from Table 2. Good performance of the model in reproducing the dry-spell statistics can be seen from Figs. 6 and 7 and also from Table 3. The coefficient of skew, the coefficient of variation, the 25% quantile, and the 75% quantile were also preserved for all the three variables, but are not shown here.

The extreme statistics (e.g., longest spell length or maximum wet-day precipitation) exhibit a high degree of variability in the simulations (refer to Tables 1–3) and an asymmetric sampling distribution as one would expect.

Note that most of the statistics that we have listed in the section on performance measures are not explicitly considered in the model. Hence the good reproduction of PDFs/PMFs of the three variables is quite interesting. The suggestion is that consideration of nonstationarity in the Markov chain leads to a first-order model at this site that appears to capture spell statistics.

SUMMARY AND CONCLUSIONS

A nonhomogeneous Markov model for simulating daily precipitation is presented in this paper. The traditional Markov-chain model is extended to consider the smooth variation in the transition probabilities from day to day, thus attempting to capture the nonstationarity in the precipitation-occurrence process. The 2×2 daily transition-probability matrix is estimated nonparametrically. The primary intended use of the model is as a simulator that is faithful to the historical data sequence, obviating the need to divide the year into seasons and subsequently fitting the Markov-chain parameters separately for each season. Simulations from the model are shown to preserve the frequency structure (PDF/PMF) of the wet-spell length, dry-spell length, and wet-day precipitation at both the seasonal and annual time scales.

In many cases, the Fourier-series approach to addressing seasonal variation in Markov-chain parameters may be just as effective. Recall that the Fourier-series approach can be shown to be a subset of the kernel approach with a specific kernel choice. The kernel approach presented here is attractive because it is relatively parsimonious, locally adaptive, and extends quite naturally to localizing the probability-density estimation for precipitation amount as well. Extensions to higher-order chains or those with more states can be made in the same sprint. One needs to define the appropriate events as was done here and go through the solution of the corresponding smoothing problem.

A limitation of the nonparametric density-estimation approach used here is the rather limited extrapolation of daily precipitation values beyond the largest value recorded. If this is a major concern, a suitable parametric density may be fitted to the local windowed precipitation data. We feel that such an approach may not be superior since extrapolation of a parametric density to the tails may suffer from a high degree of uncertainty as well.

ACKNOWLEDGMENTS

Partial support of this work by the U.S. Forest Service under contract notes INT-915550-RJVA and INT-92660-RJVA, Amend #1, is acknowledged. The principal investigator of the project is D. S. Bowles. Thanks are due to Alaa Ibrahim Ali for useful discussions.

APPENDIX. REFERENCES

Chin, E. H. (1977). "Modeling daily precipitation occurrence process with Markov Chain." *Water Resour. Res.*, 13, 949–956.

Eubank, R. L., (1988). *Spline smoothing and nonparametric regression*, Marcel Dekker, Inc., New York, N.Y.

Feyerherm, A. M., and Bark, L. D. (1965). "Statistical methods for persistent precipitation patterns." *J. of Appl. Meteorology*, 4, 320–328.

Feyerherm, A. M., and Bark, L. D. (1967). "Goodness of fit of a Markov chain model for sequences of wet and dry days." *J. Appl. Meteorology*, 6, 770–773.

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

Guzman, A. G., and Torrez, C. W. (1985). "Daily rainfall probabilities: conditional upon prior occurrence and amount of rain." *J. Climate and Appl. Meteorology*, 24(10), 1009–1014.

Haan, C. T., Allen, D. M., and Street, J. O. (1976). "A Markov chain model of daily rainfall." *Water Resour. Res.*, 12(3), 443–449.

Hopkins, J. W., and Robillard, P. (1964). "Some statistics of daily rainfall occurrence for the Canadian prairie provinces." *J. Appl. Meteorology*, 3, 600–602.

Lall, U. (1995). "Nonparametric function estimation: recent hydrologic contributions." *Rev. Geophys. 33 supp., U. S. Nat. Rep., 1991–1994*, Int. Union of Geodesy and Geophys., 1093–1102.

Lall, U., Rajagopalan, B., and Tarboton, D. G. (1995). "A nonparametric wet/dry spell model for resampling daily precipitation." *Water Resour. Res.*

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. John Wiley and Sons, New York. N.Y.

Rajagopalan, B., and Lall, U. (1995). "A kernel estimator for discrete distributions." *J. Nonparametric Statistics*, 4, 409–426.

Rajagopalan, B., and Lall, U. (1995). "Seasonality of precipitation along a meridian in the western U. S." *Geophys. Res. Lett.*, 22(9), 1081–1084.

Roldan, J., and Woolhiser, D. A. (1982). "Stochastic daily precipitation models. 1. A comparison of occurrence processes." *Water Resour. Res.*, 18(5), 1451–1459.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice and visualization*. Wiley Ser. in Probability and Math. Statistics, John Wiley and Sons, New York, N.Y.

Sheather, S. J., and Jones, M. C. (1991). "A reliable data-based bandwidth selection method for kernel density estimation." *J. Roy. Statistical Soc.*; B. 53, 683–690.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Chapman and Hall, New York, N.Y.
Smith, J. A., and Schreiber, H. A. (1974). "Point processes of seasonal thunderstorm rainfall. 1. Distribution of rainfall events." *Water Resour. Res.*, 10(3), 418–423.
Srikanthan, R., and McMahon, T. A. (1983). "Stochastic simulation of daily rainfall for Australian stations." *Trans. ASAE*, 754–766.
Stern, R. D., and Coe, R. G. (1984). "A model fitting analysis of rainfall data (with discussion)." *J. Roy. Statistical Soc. Ser. A.*, 147, 1–34.
Todorovic, P., and Woolhiser, D. A. (1975). "Stochastic model of *n*-day precipitation." *J. Appl. Meteorology*, 14(1), 17–24.
Woolhiser, D. A. Rovey, E. W. and Todorovic, P. (1973). "Temporal and spatial variation of parameters for the distribution of *n*-day precipitation." *Floods and Droughts, Proc. 2nd Int. Symp. in Hydro.*, E. F. Schulz, V. A. Koelzer, and K. Mahmood, eds., Water Resources Publ., Fort Collins, Colo., 605–614.
Woolhiser, D. A., and Pegram, G. G. S. (1979). "Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models." *J. Appl. Meteorology*, 18, 34–42.