ELSEVIER

Journal of Hydrology 210 (1998) 178–191

Journal of Hydrology

Multisite generalization of a daily stochastic precipitation generation model

D.S. Wilks

Atmospheric Science Group, Cornell University, Ithaca, NY 14853, USA

Received 24 December 1997; accepted 25 June 1998

Abstract

The familiar chain-dependent-process stochastic model of daily precipitation, consisting of a two-state, first-order Markov chain for occurrences and a mixed exponential distribution for nonzero amounts, is extended to simultaneous simulation at multiple locations by driving a collection of individual models with serially independent but spatially correlated random numbers. The procedure is illustrated for a network of 25 locations in New York state, with interstation separations ranging approximately from 10 to 500 km. The resulting process reasonably reproduces various aspects of the joint distribution of daily precipitation observations at the modeled locations. The mixed exponential distributions, in addition to providing substantially better fits than the more conventional gamma distributions, are convenient for representing the tendency for smaller amounts at locations near the edges of wet areas. Means, variances, and interstation correlations of monthly precipitation totals are also well reproduced. In addition, the use of mixed exponential rather than gamma distributions yields interannual variability in the synthetic series that is much closer to the observed.

Keywords: Markov chain; Monte Carlo methods; Precipitation; Spatial variations; Stochastic processes; Time series analysis

1. Introduction

Stochastic models for observed weather data sequences are useful in applications such as hydrological design, and agricultural or ecosystem simulation, when the observed data are inadequate with respect either to the length or completeness of the historical record or to the spatial density of observing stations. These stochastic models are sometimes known as 'weather generators' since they can be used to produce synthetic weather sequences, which are time series of random numbers which resemble statistically the observations to which the model has been fit.

A variety of stochastic weather models exist. Probably the most common are parametric empirical–statistical models, which usually represent daily weather sequences, and are based on a relatively simple stochastic process to which the underlying atmospheric physical processes are related only implicitly. Nonparametric analogs to this modeling approach also exist (e.g., Young, 1994; Lall et al., 1996). Reviews describing parametric models of this kind (emphasizing the modeling of the precipitation, which is the more difficult portion of the problem) are provided in Hutchinson (1986), Stern and Coe (1984), and Woolhiser (1992). These models are popular because they are easy to fit; lend themselves to straightforward stochastic simulation; and are oriented primarily toward simulation of daily data, for which many impacts models have been designed because of the availability of meteorological observations at this

0022-1694/98/$ - see front matter
PII S0022-1694(98)00186-3

time scale. However, an outstanding problem associated with the use of these simple empirical–statistical models is that they are single-location, or point-process, models. Therefore using these methods for simultaneous simulation of weather sequences at multiple points, for example to evaluate regional hydrological or agricultural behavior, has not been possible without ignoring the quite strong spatial correlations in daily weather data.

This problem of spatial dependence of weather events has been addressed, for precipitation at least, in a number of space–time stochastic models (e.g., Bras and Rodriguez-Iturbe, 1976; Waymire et al., 1984; Cox and Isham, 1988; Smith and Karr, 1985). Many of these models attempt to represent more or less abstractly, but explicitly, some of the atmospheric physical processes that are responsible for production of precipitation. Recently development of this kind of model has been oriented toward 'downscaling' large-scale climate change simulations, which involves relating local precipitation processes to large-scale flow in the atmosphere (e.g., Hay et al., 1991; Hughes and Guttorp, 1994a; Hughes et al., 1993; Wilson et al., 1992; Hughes and Guttorp, 1994b; Hughes et al., 1998). However, these more physically explicit stochastic models necessarily exhibit trade-offs between simplicity and fidelity to the observations being simulated. As had been noted previously in the review of Georgakakos and Kavvas (1987), the more mechanistic (and more faithful) examples of this kind of model require extensive parameter estimation and statistical verification. Their operational application has been limited.

This paper discusses extension of a particular simple and widely used empirical-statistical stochastic weather model for precipitation (Todorovic and Woolhiser, 1975; Katz, 1977), which also comprises the precipitation portion of a number of larger stochastic weather generators (e.g., Richardson, 1981), to simultaneous simulation at multiple locations. This extension is accomplished by driving each of a collection of individual single-site models with temporally independent but spatially correlated random numbers. Because the marginal distributions of these forcing random numbers are the same as in the conventional single-site case, the individual behaviors of the local models are preserved. However, the spatial correlation in the driving random number streams yields spatial correlations in the synthetic series that mimic those in the observations. The underlying goal has been production of a spatially distributed daily precipitation generator that is practical, and can be readily implemented in conjunction with a variety of operational environmental models.

2. The local precipitation model

2.1. Precipitation occurrence process

The daily precipitation model used here is the familiar chain-dependent process, comprised of a first-order, two-state Markov process governing daily precipitation occurrence, with serially independent precipitation amounts on wet days (e.g., Todorovic and Woolhiser, 1975; Katz, 1977; Waymire and Gupta, 1981; Stern and Coe, 1984; Woolhiser, 1992). It will be convenient to use separate notation for the time series of daily precipitation occurrences and amounts. Let $X_t(k)$ represent the binary event of precipitation or no precipitation occurring at location k on day t:

$$X_t(k) = \begin{cases} 0 & \text{if day } t \text{ is dry at location } k \\ 1 & \text{if day } t \text{ is wet at location } k. \end{cases} \quad (1)$$

Here 'wet' means occurrence of at least the minimum reportable precipitation amount for the day, which is 0.254 mm for the US stations to be used in the following. The time series of precipitation amounts at location k is then

$$Y_t(k) = r_t(k) X_t(k), \quad (2)$$

where $r_t(k)$ represents the non-zero precipitation amounts. Of course, $Y_t(k)$ will equal zero when $X_t(k) = 0$ and will equal $r_t(k)$ when $X_t(k) = 1$.

The first-order Markov chain model for $X_t(k)$ follows from the assumption that the probability of a wet day is defined fully by whether precipitation occurred or did not on the previous day, regardless of the earlier values of the occurrence series $X_{t-2}(k)$, $X_{t-3}(k)$, etc.:

$$Pr\{X_t(k) = 1 | X_{t-1}(k) = 0\} = p_{01}(k) \quad (3a)$$

and

$$Pr\{X_t(k) = 1 | X_{t-1}(k) = 1\} = p_{11}(k). \quad (3b)$$

The parameters $p_{01}(k)$ and $p_{11}(k)$, which are the conditional probability of a wet day following a dry day and the conditional probability of a wet day following a wet day at location k, are sufficient to define the process. Equivalently, the Markov process can be written using the complementary conditional probabilities $p_{00}(k) = 1 - p_{01}(k)$ and $p_{10}(k) = 1 - p_{11}(k)$.

Daily precipitation data for 25 stations in New York State, USA, for the years 1951–1996, are used in the following. These stations are distributed across an area of approximate dimension 500 km (east–west) by 100 km (north–south), centered near 42°N, 76°W. Station separations range approximately from 10 to 500 km. Each of these stations has a morning (07:00–09:00 LST) time of observation. Separate local models were fit for each location and for each calendar month, yielding 300 models in all. Of these, first-order Markov dependence was chosen as adequate in 291 cases by the BIC statistic (Schwarz, 1978), in preference to second- or higher-order Markov models. Since the sample sizes are large (typically 1200–1400 days) the BIC is used here in preference to the AIC statistic (Akaike, 1974), which is known to overspecify model order for large sample sizes (Katz, 1981). Both the BIC and AIC statistics are likelihood-based measures, but impose (different) penalties that increase with the number of free parameters.

Stochastic simulation of the $X_t(k)$ series under first-order Markov dependence is straightforward. The output from a uniform [0,1] random number generator is compared with the appropriate transition probability in Eq. (3a) or Eq. (3b), and a wet day is simulated if the random number is sufficiently small. Defining this threshold as the 'critical' probability

$$p_c(k) = \begin{cases} p_{01}(k), & \text{if } X_{t-1}(k) = 0 \\ p_{11}(k), & \text{if } X_{t-1}(k) = 1 \end{cases}, \tag{4}$$

the next value in the $X_t(k)$ series is then determined as

$$X_t(k) = \begin{cases} 1, & \text{if } u_t(k) \le p_c(k) \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Here $u_t(k)$ denotes uniform (probability density $f[u] = 1$, $0 \le u \le 1$) independent random forcing for the occurrence process at location k. Equivalently, and more conveniently for the multisite generation scheme described below, the occurrence process can be forced with a series of serially independent standard Gaussian random numbers, $w_t(k) \sim N[0,1]$. In this case Eq. (5) is replaced by

$$X_t(k) = \begin{cases} 1, & \text{if } w_t(k) \le \Phi^{-1}[p_c(k)] \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where $\Phi[\cdot]$ indicates the standard normal cumulative distribution function. That is, a wet day is simulated if the next random draw of $w_t(k)$ satisfies $Pr\{W \le w_t(k)\} \le p_c(k)$.

2.2. Precipitation amounts process

Nonzero precipitation amounts $r_t(k)$ are simulated here using the mixed exponential distribution, which has been used previously by Foufoula-Georgiou and Lettenmaier (1987), Hanson et al. (1994), Smith and Schreiber (1974), Wilson et al. (1992) and Woolhiser and Pegram (1979). This is a probability mixture of two one-parameter exponential distributions, with probability density function

$$f[r(k)] = \frac{\alpha(k)}{\beta_1(k)} \exp\left[\frac{-r(k)}{\beta_1(k)}\right] + \frac{1-\alpha(k)}{\beta_2(k)} \exp\left[\frac{-r(k)}{\beta_2(k)}\right]; \quad \beta_1(k) \ge \beta_2(k) > 0, \quad 0 < \alpha(k) \le 1. \tag{7}$$

Here $\alpha(k)$ is the mixing probability for location k, which determines the frequencies with which the exponential distribution with the larger (β_1) or smaller (β_2) mean will be used to generate the next value in the $r_t(k)$ series. The condition $\beta_1 \ge \beta_2$ on the two scale parameters is not mathematically necessary to apply Eq. (7) to precipitation data at a single location, but is imposed here in order to allow the multisite generation scheme described below to better simulate certain aspects of the joint distribution of precipitation amounts at the multiple locations. For $\beta_1 = \beta_2$ or $\alpha = 1$ or $\alpha = 0$, Eq. (7) reduces to the ordinary single-parameter exponential distribution.

The explicit dependence of the three parameters on the location k in Eq. (7) emphasizes that different distributions are fit for each location. These parameters were estimated using maximum likelihood, separately for each of the 300 combinations of 25 locations and 12 calendar months. The resulting mixed exponential distributions provide very good fits to these nonzero daily precipitation data—much better than two-parameter gamma distributions, which are very commonly used for this purpose (e.g., Katz,

1977; Richardson and Wright, 1984; Wilks, 1989). This superior fit of the mixed exponential distribution is consistent with the findings of Foufoula-Georgiou and Lettenmaier (1987) and Woolhiser and Roldan (1982). The difference in log-likelihoods, averaged over the 300 locations and months, between the mixed exponential and gamma fits was 15.5. For perspective on this difference, note that the BIC (Schwarz, 1978) chooses the three-parameter mixed exponential as superior to the two-parameter gamma distribution for the present sample sizes when the difference in log-likelihoods is greater than about 6.5. The AIC (Akaike, 1974) chooses the mixed exponential when the difference is only 1 or greater. These fitted mixed exponential distributions also provide quite strong separations between the 'heavy' and 'light' precipitation days, with the average of the ratio β_1/β_2 being 4.8, and a typical mixing probability a being about 0.6. However, the $\alpha(k)$ are quite variable across locations even within a given calendar month, with standard deviations across locations within a month being generally between 0.1 and 0.2.

In addition to providing a good fit to the precipitation data considered here, the mixed exponential distribution is convenient and efficient for stochastic simulation. First, a uniform [0,1] random variable is compared to the mixing probability $\alpha(k)$. The mean $\beta_1(k)$ is chosen for day t if this uniform variate is smaller than $\alpha(k)$, and $\beta_2(k)$ is chosen otherwise. Since the exponential probability density function can be integrated analytically, the resulting cumulative distribution function can be solved for r as a function of $F(r)$, which follows a uniform distribution (e.g., Lindgren, 1976). Another uniform [0,1] variate, say $v_t(k)$, is then produced and used to generate the synthetic precipitation amount by inversion (e.g., Bratley et al., 1987), using

$$r_t(k) = r_{\min} - \beta \ln[v_t(k)], \tag{8}$$

where β is either $\beta_1(k)$ or $\beta_2(k)$ as indicated above, and $r_{\min}$ is the daily precipitation accumulation below which a day is recorded as dry.

3. Extension to simultaneous multisite precipitation generation

Generalization of the single-station stochastic model described in Section 2 to multiple locations is described in this section. The same location-specific parameters $p_{01}(k)$, $p_{11}(k), \beta_1(k)$, $\beta_2(k)$, and $\alpha(k)$ described above are retained for each of a collection of locations $k = 1,\ldots,K$. The basic idea is then to drive this collection of individual station models with *vectors* of uniform [0,1] variates $\mathbf{u}_t$ and $\mathbf{v}_t$ whose elements ($u_t(k)$ and $v_t(k)$, respectively) are correlated so that $\mathrm{Corr}[u_t(k), u_t(\ell)] \neq 0$ and $\mathrm{Corr}[v_t(k), v_t(\ell)] \neq 0$, but which are mutually and serially independent so that $\mathrm{Corr}[u_t(k), v_t(\ell)] = \mathrm{Corr}[u_t(k), u_{t+1}(\ell)] = \mathrm{Corr}[v_t(k), v_{t+1}(\ell)] = 0$. Here ℓ denotes an alternate station index: k, $\ell = 1,\ldots,K$. Nonzero correlations among the elements of $\mathbf{u}_t$ and $\mathbf{v}_t$ result in interstation correlations between the resulting synthetic precipitation series, while the fact that the marginal distributions of the variates $u_t(k)$ and $v_t(k)$ are uniform and independent ensures that each local stochastic process behaves in the same way as if it alone were being simulated in the conventional way.

3.1. Multisite occurrence process

Given a network of K locations, there are $(K)(K-1)/2$ pairwise location combinations, and accordingly there are this many correlations between the elements of the occurrence forcing vector $\mathbf{u}_t$ that must be determined in order to simulate precipitation occurrences across the network. The problem of finding these correlations is simplified by the fact that the uniform variates $u_t(k)$ can be derived from standard Gaussian variates $w_t(k)$ through the transformation $u_t(k) = \Phi[w_t(k)]$ or, equivalently, the occurrence processes can be driven directly using standard normal variates according to Eq. (6).

Denote the pairwise correlations between the w_t for the station pair k and ℓ as

$$\omega(k, \ell) = \mathrm{Corr}[w_t(k), w_t(\ell)]. \tag{9}$$

Together with the pairs of transition probabilities (Eq. (3a) and Eq. (3b)) for stations k and ℓ a particular $\omega(k,\ell)$ will yield a corresponding correlation between the synthetic binary series for the two sites,

$$\xi(k, \ell) = \mathrm{Corr}[X_t(k), X_t(\ell)]. \tag{10}$$

Let $\xi^o(k,\ell)$ denote the observed sample counterpart of $\xi(k,\ell)$, which will have been estimated from the observed binary series $X_t^o(k)$ and $X_t^o(\ell)$ at locations

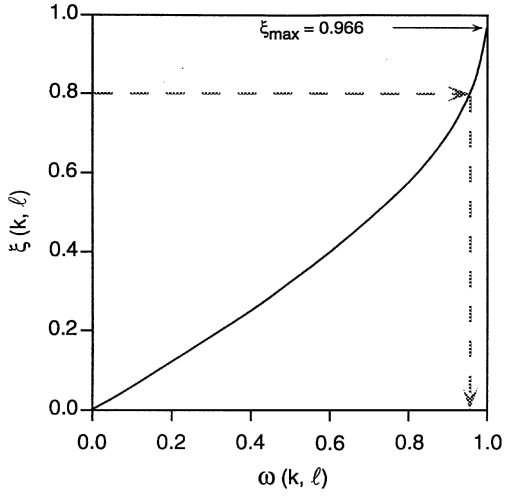

Fig. 1. Illustration of the relationship between the correlations $\omega(k,\ell)$ and $\xi(k,\ell)$ produced by the July Markov parameters for Freeville (p_{01} = 0.329, p_{11} = 0.462) and Ithaca (p_{01} = 0.320, p_{11} = 0.441). The observed correlation between the binary series at these sites is $\xi^{o}(k,\ell)$ = 0.800, so that choosing $w(k,\ell)$ = 0.957 will reproduce this correlation in the synthetic binary series. Forcing the Markov chain models for these locations with identical random number streams yields the correlation ξ_{max} = 0.966.

k and ℓ. This first aspect of the multisite precipitation generation problem is solved by individually finding the $(K)(K-1)/2$ pairwise correlations $\omega(k,\ell)$ which, together with the corresponding pairs of transition probabilities in Eq. (3a) and Eq. (3b), reproduces $\xi(k,\ell) = \xi^{o}(k,\ell)$ for each pair of stations.

Direct computation of $\omega(k,\ell)$ from $\xi^{o}(k,\ell)$ appears to be infeasible, since the thresholding process in Eq. (5) or Eq. (6) implies that the hypothetical series $u_t^{o}(k)$ or $w_t^{o}(k)$ corresponding to an observed binary precipitation series $X_t^{o}(k)$ cannot be computed. However, one finds empirically that there is a monotonic relationship between $\omega(k,\ell)$ and the resulting $\xi(k,\ell)$, for given pairs of the transition probabilities $[p_{01}(k), p_{11}(k)]$ and $[p_{01}(\ell), p_{11}(\ell)]$. Fig. 1 shows one such relationship, computed through stochastic simulations using the July transition probabilities for Freeville (p_{01} = 0.329, p_{11} = 0.462) and Ithaca (p_{01} = 0.320, p_{11} = 0.441), which are among the closest stations (12 km apart) in the network. This relationship passes through the origin, since independent forcing noise will produce independent synthetic series, and rises to a maximum of $\xi_{max}(k,\ell)$ = 0.966 when the Markov chain models for the two locations are forced with identical random number streams. For these two stations in July, the observed $\xi^{o}(k,\ell)$ = 0.800, and the dashed arrows in Fig. 1 indicate that the Markov models for these two stations reproduce this binary series correlation when forced by bivariate standard Gaussian noise with $\omega(k,\ell)$ = 0.957. In general, the maximum correlation ξ_{max} decreases as the pairs of transition probabilities (Eq. (3a) and Eq. (3b)) for the stations in question become increasingly dissimilar, so that it is not possible to produce very large synthetic correlations $\xi(k,\ell)$ for station pairs with very different Markov chain parameters. However, in practice stations with even moderately different precipitation occurrence climates (as characterized by the transition probabilities) will tend to exhibit only modest $\xi^{o}(k,\ell)$. For the data considered here, $\xi^{o}(k,\ell) < \xi_{max}$ for all station pairs and all months.

In practice, one can invert the relationship between $\omega(k,\ell)$ and $\xi(k,\ell)$ using a nonlinear rootfinding algorithm (e.g., Press et al., 1986), by defining the function $Q(\omega) = \xi^{o} - \xi(\omega)$ and finding the $\omega(k,\ell)$ for which $Q(\omega) = 0$. Since $\xi(\omega)$ is evaluated on each iteration with a stochastic simulation, more stable results are achieved by using long synthetic series, and beginning each new evaluation of $\xi(\omega)$ with the same random number seed.

Realizations of the vector $\mathbf{w}_t$ may be generated from the multivariate normal distribution having mean vector $\mathbf{0}$ and variance–covariance matrix $[\mathbf{\Omega}]$, the elements of which are the correlations $\omega(k,\ell)$. The individual $w_t(k)$ which result will each be distributed as (univariate) standard normal, and accordingly may be used in Eq. (6) to generate the precipitation occurrence time series $X_t(k)$. Each synthetic series $X_t(k)$ will then necessarily follow the specified Markov chain defined by Eq. (3a) and Eq. (3b), by construction; so that each individual series will also exhibit the correct overall relative frequency of wet days (Gregory et al., 1993). Fig. 2 indicates that the full joint distribution of simultaneous precipitation occurrence across the 25-station network is also well represented by the synthetic series. These panels compare the joint probabilities from the observed and synthetic daily series for all station pairs k and ℓ and all months, that both locations are dry (Fig. 2a), and that both locations are wet (Fig. 2b). Clearly the performance is very good, for both the large-scale precipitation systems of winter and the smaller scale convective precipitation of summer.

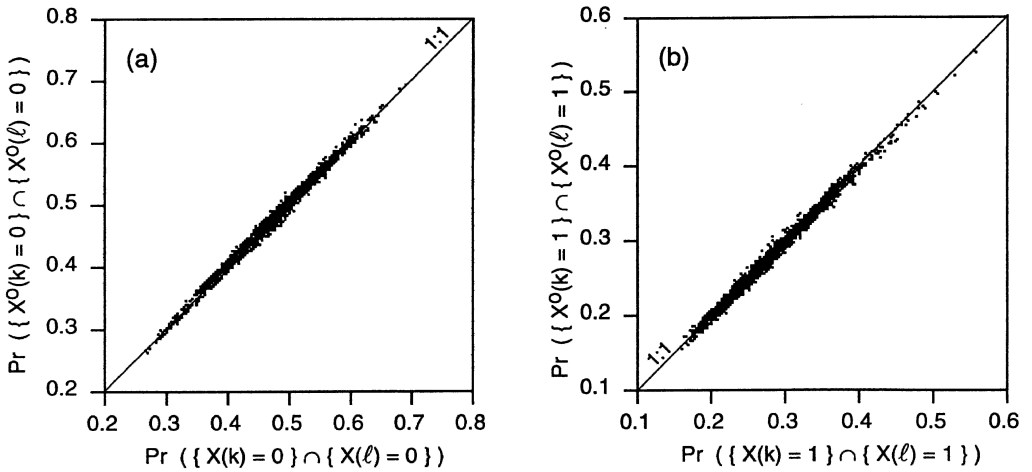

Fig. 2. Joint probabilities that station pairs are both (a) dry, or (b) wet, on a given day, in the synthetic series (horizontal) and the observed series (vertical), for all 3600 combinations of station pairs and all 12 months.

A deficiency in the precipitation occurrence generation is indicated in Fig. 3, which shows cross-correlations for precipitation occurrence lagged by 1 day (i.e., Corr[$X_t(k)$, $X_{t+1}(\ell)$], $k \neq \ell$), for the 7200 combinations of station pairs (each both leading and lagging) and months. Here it can be seen that the synthetic occurrence series tend to under-represent some of the larger lagged cross-correlations in the observations, which arise from the progression of synoptic-scale weather systems across the region. That is, there is a clear tendency for precipitation systems to move from west to east across the region. However, because the size of the study area is not large in relation to the typical 1-day movement of the systems, the magnitudes of these lagged correlations are quite modest, particularly when viewed from the perspective of proportion of variance accounted for.

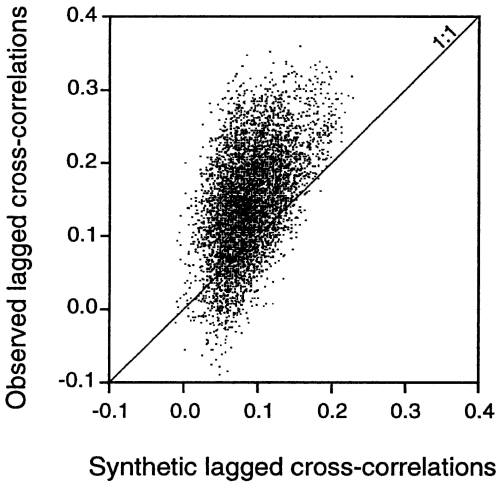

Fig. 3. Cross-correlations for precipitation occurrence lagged by one day (Corr[$X_t(k)$, $X_{t+1}(\ell)$], $k \neq \ell$), for the 7200 combinations of station pairs and months.

3.2. Multisite amounts process

Good specification of the field $\mathbf{Y}_t$ requires more than generation of a field of rainfall amounts $\mathbf{r}_t$ in Eq. (2) independently of the occurrence process $\mathbf{X}_t$. This is because the probability distributions of precipitation amounts conditional on nearby stations being dry have smaller means than the corresponding distributions conditional on near neighbors being wet. This problem of smooth continuity between wet and dry areas was referred to as spatial intermittence by Bardossy and Plate (1992), who point out that failure to address it leads to unrealistically sharp transitions between wet and dry portions of the spatial domain.

As alluded to previously, one reason that the mixed exponential distribution (Eq. (7)) is convenient here is that it allows a simple and easily implemented way to accommodate this tendency of real precipitation fields to exhibit nonzero amounts with smaller expected values in locations close to dry areas. The precipitation means $\beta_1(k)$ or $\beta_2(k)$ are chosen at each location k according to the relationship between the uniform

variate for precipitation occurrence $u_t(k)$ and the mixing parameter $\alpha(k)$. Specifically,

$$\beta_t(k) = \begin{cases} \beta_1(k) & , \quad u_t(k)/p_c(k) \le \alpha(k) \\ \beta_2(k) & , \quad u_t(k)/p_c(k) > \alpha(k) \end{cases}, \tag{11}$$

where $u_t(k) = \Phi[w_t(k)]$. Thus the means of the distributions of the precipitation amounts $r_t(k)$ are related to the occurrence process $\mathbf{X}_t$ through Eq. (5) or Eq. (6). In particular, because of the strong spatial correlation in the $w_t(k)$, wet locations in proximity to areas of zero precipitation for day t will generally result from $u_t(k)$ smaller than but near $p_c(k)$, so that Eq. (11) will choose the smaller of the two mixed exponential scale parameters $\beta_2(k)$ for that day. Conversely, locations in the 'heart' of a wet area on day t are more likely to have been forced by a relatively small $u_t(k)$, in which cases Eq. (11) will choose the larger scale parameter $\beta_1(k)$. Note that the conditional distributions of $u_t(k)/p_c(k)$, given that $u_t(k) \le p_c(k)$, are uniform on [0,1], and thus Eq. (11) preserves the unconditional frequencies of $\beta_1(k)$ and $\beta_2(k)$ implied by Eq. (7) because nonzero precipitation is simulated at location k only if $u_t(k) \le p_c(k)$.

As will be seen below, the suppression of large precipitation amounts at the fringes of wet regions accomplished by Eq. (11) can be improved upon by continuously varying the larger of the two scale parameters according to

$$\beta_t(k) = \begin{cases} \beta_2(k) + 2[\beta_1(k) - \beta_2(k)]\left[1 - \dfrac{u_t(k)}{\alpha(k)p_c(k)}\right], \\ \beta_2(k), \end{cases} \quad \begin{matrix} u_t(k)/p_c(k) \le \alpha(k) \\ u_t(k)/p_c(k) > \alpha(k) \end{matrix}. \tag{12}$$

That is, for relatively large $u_t(k)$ the smaller scale parameter is chosen as in Eq. (11), but for $u_t(k) < \alpha(k)\, p_c(k)$ the resulting scale parameter $\beta_t(k)$ is tapered linearly from $\beta_2(k)$ when $u_t(k) = \alpha(k)p_c(k)$ to a maximum of $2\beta_1(k) - \beta_2(k)$ when $u_t(k) = 0$. This form of tapering for the scale parameter $\beta_1(k)$ has been chosen to preserve the unconditional precipitation means for mixed exponential (Eq. (7)) variates:

$$E[r_t(k)] = \alpha(k)\beta_1(k) + [1 - \alpha(k)]\beta_2(k). \tag{13}$$

However, use of Eq. (12) increases the variance somewhat, from

$$\mathrm{Var}[r_t(k) = \alpha(k)\beta_1^2(k) + [1 - \alpha(k)]\beta_2^2(k) + \alpha(k)[1 - \alpha(k)] \times [\beta_1(k) - \beta_2(k)]^2, \tag{14}$$

which pertains to the mixed exponential distribution generally, including when Eq. (11) is applied, to

$$\mathrm{Var}[r_t(k)] = \frac{4}{3}\alpha(k)\beta_1^2(k) + [1 - \alpha(k)]\beta_2^2(k) + \alpha(k) \times [1 - \alpha(k)][\beta_1(k) - \beta_2(k)]^2, \tag{14b}$$

which is produced by Eq. (12) (see, e.g., Katz and Parlange, 1996).

Having chosen the scale parameters $\beta_t(k)$ using either Eq. (11) or Eq. (12), the precipitation amounts $r_t(k)$ are simulated using Eq. (8) and a vector of correlated uniform variates $\mathbf{v}_t$. As in the case of simulating precipitation occurrence, it is convenient to compute the elements of this vector, $v_t(k)$, from a corresponding realization of correlated standard normal variates $z_t(k)$ according to $v_t(k) = \Phi[z_t(k)]$. This vector $\mathbf{z}_t$ is drawn from a multivariate normal distribution with mean $\mathbf{0}$ and variance–covariance matrix $[\mathbf{Z}]$, whose elements are

$$\zeta(k, \ell) = \mathrm{Corr}[z_t(k), z_t(\ell)]. \tag{15}$$

As was the case for the correlations $\omega(k,\ell)$ in Eq. (9), direct computation of $[\mathbf{Z}]$ from a particular data set is not feasible since the $\mathbf{z}_t$ are not observed (they depend on the unobserved $\mathbf{w}_t$ through Eq. (11) or Eq. (12)). The correlations in Eq. (15) are estimated here using an analogous procedure. Define the cross-correlations between daily precipitation at locations k and ℓ as

$$\eta(k, \ell) = \mathrm{Corr}[Y_t(k), Y_t(\ell)]. \tag{16}$$

A particular $\zeta(k,\ell)$, together with the corresponding correlation $\omega(k,\ell)$ and the Markov chain and mixed exponential parameters for the stations k and ℓ, yields a unique $\eta(k,\ell)$. However, in working with data from the present stations it was found that determining the correlations $\zeta(k,\ell)$ pairwise, in the manner described in for $\omega(k,\ell)$, yields matrices $[\mathbf{Z}]$ that are not positive definite (i.e., have one or more negative eigenvalues). That is, while it is possible to find a particular $\zeta(k,\ell)$ yielding $\eta(k,\ell) = \eta^{\circ}(k,\ell)$ for each pair of stations k and ℓ, the resulting collection of $K(K - 1)/2 = 300$ such correlation are not necessarily mutually

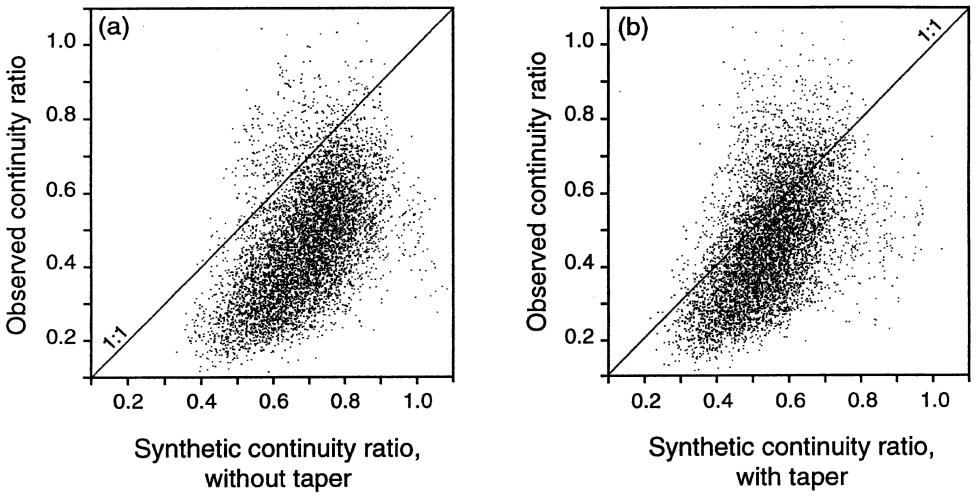

Fig. 4. Continuity ratios (Eq. (18)) from observations, versus those from synthetic series generated using (a) the untapered Eq. (11), and (b) tapered (Eq. (12)) mixed exponential scale parameters. Points are for all pairs of stations and all 12 months.

consistent. It is known that this problem can occur in situations where the correlations are not computed from a full set of consistent observations (e.g., Crosby and Maddock, 1970; Bras and Rodriguez-Iturbe, 1985; Bardossy and Plate, 1992). In the present analysis this problem may derive from the considerable flexibility provided by the three parameters in Eq. (7) for each location, and in particular that no constraints on the spatial consistency of these parameters have been imposed but their estimation is subject to sampling variations. In order to obtain a computationally feasible [**Z**] in the following, its components (Eq. (15)) are obtained as smooth functions of the distances between the station pairs, of the form

$$\zeta(k,\ell) = \exp[-d_1 D(k,\ell)^{d_2}], \tag{17}$$

where $D(k,\ell)$ is the horizontal distance between stations k and ℓ, and parameters d_1 and d_2 are chosen for

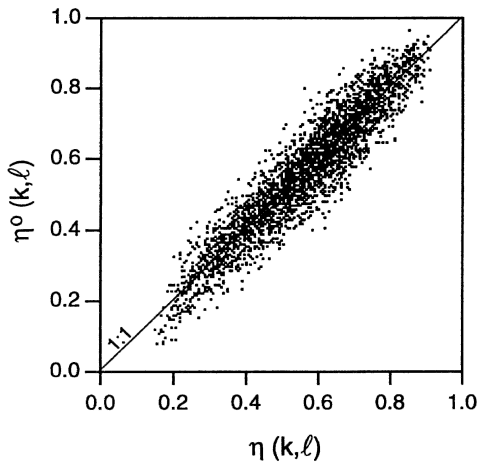

Fig. 5. Comparison of interstation correlations between daily precipitation amounts (Eq. (16)) in the observations (vertical) and synthetic series generated using Eq. (12) (horizontal), for all station pairs and months.

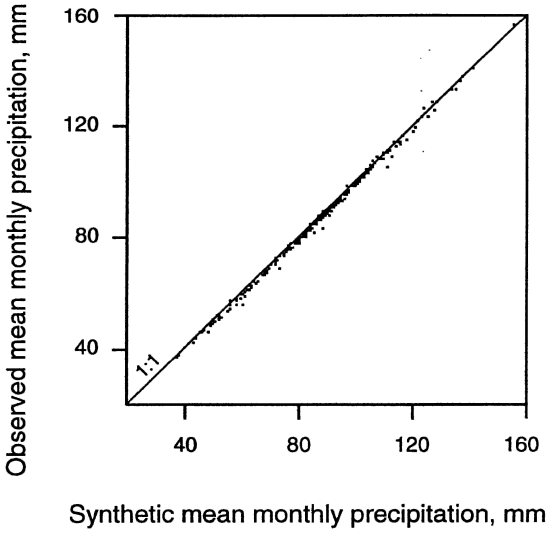

Fig. 6. Observed versus synthetic mean monthly precipitation, for all stations and all months.

each month which minimize $\Sigma_{k,\ell}[\eta(k,\ell)-\eta^{o}(k,\ell)]^2$. For the present station network the functions (Eq. (17)) are nearly flat as a function of interstation separation in winter, but decrease to less than 0.1 at the maximum interstation distance in summer. While not necessary from a purely computational perspective, it would be possible (and perhaps also desirable) to reduce drastically the number of parameters for precipitation occurrence in by fitting a function of the form of Eq. (17) to the correlations $\omega(k,\ell)$ in Eq. (9).

Fig. 4 illustrates both the 'continuity' relationship of the precipitation intensity and occurrence processes in the observations, and the capacity of the schemes Eq. (11) and Eq. (12) to capture this relationship at least in terms of the mean. The statistic plotted in Fig. 4 is the 'continuity ratio'; which is introduced here to express the ratio of the mean of the nonzero precipitation amounts at station k given zero precipitation at station ℓ, to the mean of the nonzero precipitation amounts at station k given nonzero precipitation at station ℓ,

$$\text{continuity ratio} = \frac{E[Y_t(k)|Y_t(k)>0,\ Y_t(\ell)=0]}{E[Y_t(k)|Y_t(k)>0,\ Y_t(\ell)>0]}. \tag{18}$$

The vertical axes in Fig. 4 show continuity ratios computed from the observed data, and the horizontal axes in Fig. 4a and Fig. 4b show the corresponding ratios for synthetic series computed using Eq. (11) and Eq. (12), respectively. Both panels show all combinations of station pairs for all 12 months. The degree to which mean precipitation on rain days is related to precipitation occurrence at other stations in the observed data can be appreciated by the vertical spread of points in these panels. If the occurrence- and amounts processes were independent at each of the sites, then the observed continuity ratio (Eq. (18)) for all stations would be 1. Similarly, a stochastic generation algorithm not including dependence between the occurrence and intensity processes would produce synthetic series exhibiting unit continuity ratios. In the observations, the more widely separated station pairs exhibit continuity ratios that suggest near-independence for the amounts and occurrence processes. For very close station pairs, average precipitation amounts given a dry day at the neighboring location are only 10–20% of those on days when both locations are wet. Fig. 4 shows that both of the schemes, Eq. (11) and Eq. (12), qualitatively reproduce the general nature of this dependence between the precipitation intensity and occurrence processes, and that use of the tapered scale parameters (Eq. (12)) in Fig. 5b does a somewhat better job. However, the fact that most of the points in both panels are below the 1:1 lines indicates that, on average, neither Eq. (11) nor Eq. (12) produces dependence between the synthetic precipitation amounts and occurrences that is as strong as in the observations.

More importantly, Fig. 5 shows the effectiveness of the scheme for reproducing the correlations $\eta(k,\ell)$ between the daily precipitation amounts $Y_t(k)$. The vertical axis shows values of this statistic $\eta^{o}(k,\ell)$ computed from the observations, and the horizontal axis shows the results for synthetic series constructed using Eq. (12) for tapering the mixed exponential distribution scale parameter. Clearly the procedure is successful in simulating this aspect of the observed precipitation statistics. The corresponding results obtained when using Eq. (11) are very similar, but tend to exhibit synthetic $\eta(k,\ell)$ that are too small for nearby station pairs during the winter months. Failing to include a mechanism for producing dependence between the amounts and occurrence processes yields daily $\eta(k,\ell)$ that are substantially too low, because large precipitation amounts are generated too frequently at locations close to dry ($Y_t(k) = 0$) stations.

3.3. Performance of the implied monthly precipitation statistics

While the stochastic process above has been fitted to the statistics of daily precipitation occurrences and amounts, it is of interest to investigate the degree to which they reproduce statistics of the observed precipitation climate on longer time scales. Accurate reproduction of monthly statistics might be of intrinsic interest for some applications, but in any case constitute a challenging test of the statistical properties of the daily multisite stochastic model. Fig. 6 shows the relationship between synthetic and observed monthly mean precipitation, for all 25 stations and all 12 months. Here the synthetic precipitation has been generated using the untapered scale

parameters in Eq. (11), although the results using Eq. (12) (or indeed any other distribution for daily precipitation amounts that preserves the mean) are essentially the same. The very close correspondence is not surprising, since the monthly means for the synthetic series are determined by the mean number of wet days per month and the mean precipitation per wet day (e.g., Todorovic and Woolhiser, 1975; Katz, 1977; Katz, 1985; Wilks, 1992), and these sample quantities are well reproduced by the Markov chain and mixed exponential models, respectively.

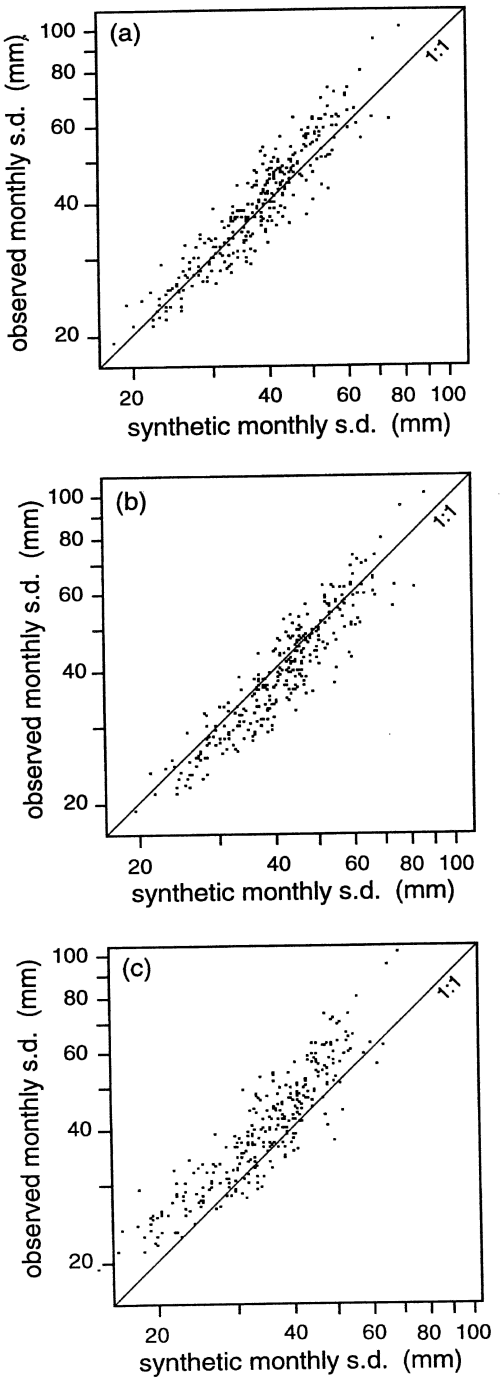

Fig. 7.

Similarly, Fig. 7 shows the standard deviations of monthly precipitation (characterizing the interannual variation in total monthly precipitation) in the observations versus synthetic series generated using (a) untapered mixed exponential distributions (Eq. (11)), (b) tapered mixed exponential distributions (Eq. (12)), and (c) gamma distributions also fit using maximum likelihood. Fig. 7a shows that use of the ordinary, untapered mixed exponential distribution produces interannual variability of monthly accumulated precipitation that is slightly too small, on average (i.e., most points are above the 1:1 line). This kind of stochastic precipitation model commonly under-represents interannual variability (e.g., Buishand, 1978; Wilks, 1989; Gregory et al., 1993; Katz and Parlange, 1993); and the source of this behavior is not completely clear (Katz and Parlange, 1998), although it is usually more pronounced than that shown in Fig. 7a. One possible reason for the relatively good results obtained here with respect to interannual variability is that the climate in New York state appears to respond relatively weakly to the El Niño–Southern Oscillation phenomenon (e.g., Livezey et al., 1997; Ropelewski and Halpert, 1996), so that year to year variations in the statistics of precipitation occurrence may be weaker here than in some other regions. Since use of the tapered mixed exponential scale parameter (Eq. (12)) increases the variance of daily nonzero precipitation amounts, it increases the monthly variance as well (e.g., Todorovic and Woolhiser, 1975; Katz, 1977, 1985; Wilks, 1992). Fig. 7b shows that this formulation for the

Fig. 7. Standard deviations of monthly total precipitation (reflecting interannual variability of accumulated monthly precipitation) in the observed data versus synthetic series generated using (a) mixed exponential distributions and Eq. (11), (b) mixed exponential distributions and Eq. (12), and (c) gamma distributions. Points are shown for all stations and all months.

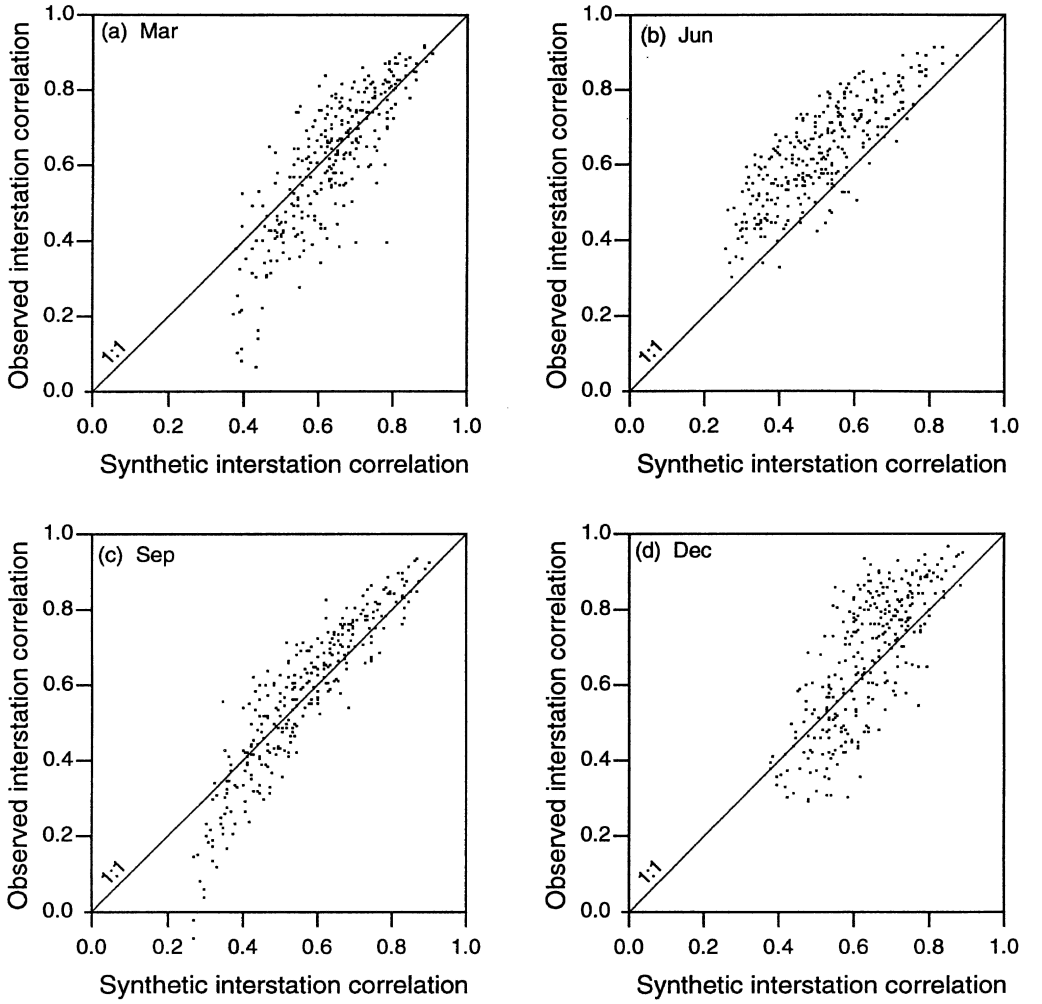

Fig. 8. Comparison of observed versus synthetic correlations for monthly total precipitation between all station pairs. (a) March, (b) June, (c) September, and (d) December.

stochastic model slightly over-represents the interannual variability, although both panels, Fig. 7a and Fig. 7b, indicate quite plausible portrayals of the observed monthly standard deviations. Fig. 7c, by contrast, shows that representing daily precipitation amounts with gamma distributions (which are used very commonly for this purpose) yields a very marked under-representation of the interannual variability.

Finally, Fig. 8 compares interstation correlations for the monthly total precipitation in the observations versus those generated by the stochastic model using Eq. (12), for four representative months. Here, not all of the scatter can be attributed to sampling variations associated with having only 46 years of observed data (the 95% confidence interval at $\rho = 0.5$ is approximately 0.24–0.7), but overall this aspect of the monthly precipitation climate is also reasonably well captured, even though the parameters have been fit to the underlying daily precipitation series.

4. Summary and conclusions

This paper has presented a multisite generalization of the well-known chain-dependent process model for daily precipitation observations, based on the simple

idea of driving individual station-level models with spatially correlated but temporally independent random vectors. Accordingly, the marginal properties of the local models are preserved while realistic spatial correlations among the variables at different sites are simulated. This scheme does not pretend to model directly the physical processes underlying the spatial distribution of weather variations, but rather is a utilitarian product that could be implemented relatively simply for stochastic simulation of weather in conjunction with a variety of agricultural, ecological and hydrological models.

Simple first-order Markov dependence has been retained here to model the time dependence of daily precipitation occurrence. While others have found this model to produce long dry spells with inadequate frequency (e.g., Buishand, 1978; Racsko et al., 1991; Lettenmaier, 1995), for the New York locations considered the BIC statistic chooses first- over higher-order Markov dependence in the overwhelming majority of cases. However, for other climates where first-order Markov dependence is clearly inadequate, the present scheme for multisite precipitation occurrence could be implemented as well under higher-order Markov dependence. In such cases the greater scope for intersite parameter heterogeneity would suggest that the maximum achievable intersite occurrence correlation $\xi_{\max}$ (c.f. Fig. 1) would be smaller.

The mixed exponential distribution (Eq. (7)) has been chosen to represent daily nonzero precipitation amounts. While a less common choice than the gamma distribution, it was found to provide substantially better fits to the data from this region. Using the mixed exponential in preference to the gamma distribution seems to ameliorate, at least for the locations considered here, the 'overdispersion', or inadequate recovery by the model of variability in the observations that is commonly found (e.g., Buishand, 1978; Wilks, 1989; Gregory et al., 1993; Katz and Parlange, 1998). In addition, this distribution allows a quite convenient means of representing the tendency for precipitation amounts near the boundaries of a wet part of the domain on a given day to be smaller than in the 'core' precipitation area. Dependence of the precipitation amounts distributions on the occurrence process produces some time dependence in consecutive nonzero precipitation amounts, although it is smaller than the dependence evident in the observations, which is statistically significant but practically modest, as has been reported elsewhere for other climates (Buishand, 1978; Foufoula-Georgiou and Lettenmaier, 1987; Gregory et al., 1993). While it is common to ignore this kind of serial correlation in stochastic precipitation models (as has been done here), it could in principle be included within the present framework by constructing a multivariate autoregression for the $\mathbf{z}_t$ series which forces the amounts process.

While not specifically a 'downscaling' model, the present scheme is similar in spirit to the multisite precipitation model of Bardossy and Plate (1992). An advantage of the present scheme over that of Bardossy and Plate (1992) is that it allows different correlation structures for precipitation occurrence and precipitation intensity, which has been found in observations both here and elsewhere (Hutchinson, 1995). Another connection to recent efforts in statistical downscaling (e.g., Hughes and Guttorp, 1994a) is that the random vectors $\mathbf{u}_t$ and $\mathbf{v}_t$ might be regarded as 'hidden' processes. It could also be possible to adapt the present scheme for simulation of multisite weather series under an assumed climate change, through station-level parameter adjustments of the kind described in Wilks (1992).

A potentially important extension to this work would be treatment of the full Richardson (1981) model in a similar way. Since the precipitation portion of that model is exactly the stochastic process treated in the present paper, this extension would require in addition that the trivariate (maximum temperature, minimum temperature, and solar radiation) autoregressions for each of K stations be simulated simultaneously. In principle, this could be done relatively easily, simply by constructing a first-order vector autoregression of dimension 3K, and this approach to simulation of these three nonprecipitation variables could be computationally feasible for networks up to a few hundred locations. Another possible extension could be interpolation of the parameters (including the intersite correlation functions) to a regular grid, in order to yield synthetic weather data on an arbitrarily defined spatial scale, regardless of the spatial density of the observations. Simple spatial correlation models like Eq. (17) are apparently adequate when the topographic relief is not too drastic, but inclusion

of elevation differences (e.g., Hanson et al., 1989 or Daly et al., 1994) might also be necessary for this purpose in more mountainous regions.

Acknowledgements

The initial idea for this project was conceived during conversations with Tom Wigley. Climate data was kindly provided by Keith Eggleston, of the Northeast Regional Climate Center, Cornell University. This work was supported by the Natural Resources Conservation Service, USDA, under Agreement 68-7482-7-250.

References

Akaike, H., 1974. A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19, 716–723.

Bardossy, A., Plate, E.J., 1992. Space-time model for daily rainfall using atmospheric circulation patterns, Water Resources Research, 28, 1247–1259.

Bras, R., Rodriguez-Iturbe, I., 1976. Rainfall generation: a nonstationary time varying multidimensional model, Water Resources Research, 12, 450–456.

Bras, R., Rodriguez-Iturbe, I., 1985. Random Functions and Hydrology. Addison-Wesley, Reading, MA, 559 pp.

Bratley, P., Fox, B.L., Schrage, L.E., 1987. A Guide to Simulation, Springer-Verlag, Berlin, 397 pp.

Buishand, T.A., 1978. Some remarks on the use of daily rainfall models, Journal of Hydrology, 36, 295–308.

Cox, D.R., Isham, V., 1988. A simple spatial-temporal model of rainfall, Proceedings of the Royal Society London, A415, 317–328.

Crosby, D.S., Maddock, T. III , 1970. Estimating coefficients of a flow generator for monotone samples of data, Water Resources Research, 6, 1079–1086.

Daly, C., Neilson, R.P., Phillips, D.L., 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain, Journal of Applied Meteorology, 33, 140–158.

Foufoula-Georgiou, E., Lettenmaier, D.P., 1987. A Markov renewal model for rainfall occurrences, Water Resources Research, 23, 875–884.

Georgakakos, K.P., Kavvas, M.L., 1987. Precipitation analysis, modeling, and prediction in hydrology, Reviews of Geophysics, 25, 163–178.

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

Hanson, C.L., Osborn, H.B., Woolhiser, D.A., 1989. Daily precipitation simulation model for mountainous areas, Transactions of the American Society of Agricultural Engineers, 32, 865–873.

Hanson, C.L., Cumming, K.A., Woolhiser, D.A., Richardson, C.W., 1994. Microcomputer program for daily weather simulations in the contiguous United States. USDA/ARS, ARS-114, 38 pp.

Hay, L.E., McCabe, G.J. Jr., Wolock, D.M., Ayres, M.A., 1991. Simulation of precipitation by weather type analysis, Water Resources Research, 27, 493–501.

Hughes, J.P., Guttorp, P., 1994. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena, Water Resources Research, 30, 1535–1546.

Hughes, J.P., Guttorp, P., 1994. Incorporating spatial dependence and atmospheric data in a model of precipitation, Journal of Applied Meteorology, 33, 1503–1515.

Hughes, J.P., Lettenmaier, D.P., Guttorp, P., 1993. A stochastic approach for assessing the effect of changes in synoptic circulation patterns on gauge precipitation, Water Resources Research, 29, 3303–3315.

Hughes, J.P., Guttorp, P., Charles, S.P., 1998. A non-homogeneous hidden Markov model for precipitation occurrence. Applied Statistics (in press).

Hutchinson, M.F., 1986. Methods of generation of weather sequences. In: Bunting, A.H. (Ed.), Agricultural Environments, C.A.B. International, Wallingford, pp. 149–157.

Hutchinson, M.F., 1995. Stochastic space-time weather models from ground-based data, Agricultural and Forest Meteorology, 73, 237–264.

Katz, R.W., 1977. Precipitation as a chain-dependent process, Journal of Applied Meteorology, 16, 671–676.

Katz, R.W., 1981. Estimating the order of a Markov chain, Technometrics, 23, 243–249.

Katz, R.W., 1985. Probabilistic models. In: Murphy, A.H., Katz, R.W. (Eds.), Probability, Statistics, and Decision Making in the Atmospheric Sciences. Westview, Boulder, CO, pp. 261–288.

Katz, R.W., Parlange, M.B., 1993. Effects of an index of atmospheric circulation on stochastic properties of precipitation, Water Resources Research, 29, 2335–2344.

Katz, R.W., Parlange, M.B., 1996. Mixtures of stochastic processes: application to statistical downscaling, Climate Research, 7, 185–193.

Katz, R.W., Parlange, M.B., 1998. Overdispersion phenomenon in stochastic modeling of precipitation, Journal of Climate, 11, 591–601.

Lall, U., Rajagopalan, B., Tarboton, D.G., 1996. A nonparametric wet/dry spell model for resampling daily precipitation, Water Resources Research, 32, 2803–2823.

Lettenmaier, D., 1995. Stochastic modeling of precipitation with applications to climate model downscaling. In: von Storch, H., Navarra, A. (Eds.), Analysis of Climate Variability, Applications of Statistical Techniques, Springer-Verlag, Berlin, pp. 197–212.

Lindgren, B.W., 1976. Statistical Theory. Macmillan, London, 614 pp.

Livezey, R.E., Masutani, M., Leetmaa, A., Rui, H., Ji, M., Kumar, A., 1997. Teleconnective response of the Pacific–North

American region atmosphere to large central equatorial Pacific SST anomalies, Journal of Climate, 10, 1787–1820.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1986. Numerical Recipes, the Art of Scientific Computing. Cambridge, 818 pp.

Racsko, P., Szeidl, L., Semenov, M., 1991. A serial approach to local stochastic weather models, Ecological Modeling, 57, 27–41.

Richardson, C.W., 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation, Water Resources Research, 17, 182–190.

Richardson, C.W., Wright, D.A., 1984. WGEN: A model for generating daily weather variables. USDA/ARS, ARS-8, 83 pp.

Ropelewski, C.F., Halpert, M.S., 1996. Quantifying southern oscillation–precipitation relationships, Journal of Climate, 9, 1043–1059.

Schwarz, G., 1978. Estimating the dimension of a model, Annals of Statistics, 6, 461–464.

Smith, J.A., Karr, A.F., 1985. Parameter estimation for a model of space-time rainfall, Water Resources Research, 21, 1251–1257.

Smith, R.E., Schreiber, H.A., 1974. Point processes of seasonal thunderstorm rainfall. Part 2: rainfall depth probabilities, Water Resources Research, 10, 418–423.

Stern, R.D., Coe, R., 1984. A model fitting analysis of daily rainfall data, Journal of the Royal Statistical Society, A147, 1–34.

Todorovic, P., Woolhiser, D.A., 1975. A stochastic model of *n*-day precipitation, Journal of Applied Meteorology, 14, 17–24.

Waymire, E., Gupta, V.K., 1981. The mathematical structure of rainfall representations. 1. A review of the stochastic rainfall models, Water Resources Research, 17, 1261–1272.

Waymire, E., Gupta, V.K., Rodriguez-Iturbe, I., 1984. A spectral theory of rainfall at the meso-β scale, Water Resources Research, 20, 1453–1465.

Wilks, D.S., 1989. Conditioning stochastic daily precipitation models on total monthly precipitation, Water Resources Research, 23, 1429–1439.

Wilks, D.S., 1992. Adapting stochastic weather generation algorithms for climate change studies, Climatic Change, 22, 67–84.

Wilson, L.L., Lettenmaier, D.P., Skyllingstad, E., 1992. A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation, Journal of Geophysical Research, D3, 2791–2809.

Woolhiser, D.A., 1992. Modeling daily precipitation—progress and problems. In: Walden, A.T., Guttorp, P. (Eds.), Statistics in the Environmental and Earth Sciences. John Wiley, New York, pp. 71–89.

Woolhiser, D.A., Pegram, G.S., 1979. Maximum likelihood estimation of Fourier coefficients to describe seasonal variation of parameters in stochastic daily precipitation models, Journal of Applied Meteorology, 18, 34–42.

Woolhiser, D.A., Roldan, J., 1982. Stochastic daily precipitation models. 2. A comparison of distribution of amounts, Water Resources Research, 18, 1461–1468.

Young, K.C., 1994. A multivariate chain model for simulating climatic parameters from daily data, Journal of Applied Meteorology, 33, 661–671.