Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

A Nonstationary Multisite Model for Rainfall

Bruno Sansó & Lelys Guenni

To cite this article: Bruno Sansó & Lelys Guenni (2000) A Nonstationary Multisite Model for Rainfall, Journal of the American Statistical Association, 95:452, 1089-1100, DOI: 10.1080/01621459.2000.10474305

To link to this article: http://dx.doi.org/10.1080/01621459.2000.10474305

Published online: 17 Feb 2012.

A Nonstationary Multisite Model for Rainfall

Bruno SANSÓ and Lelys GUENNI

Estimation and prediction of the amount of rainfall in time and space is a problem of fundamental importance in many applications in agriculture, hydrology, and ecology. Stochastic simulation of rainfall data is also an important step in the development of stochastic downscaling methods where large-scale climate information is considered as an additional explanatory variable of rainfall behavior at the local scale. Simulated rainfall has also been used as input data for many agricultural, hydrological, and ecological models, especially when rainfall measurements are not available for locations of interest or when historical records are not of sufficient length to evaluate important rainfall characteristics as extreme values. Rainfall estimation and prediction were carried out for an agricultural region of Venezuela in the central plains state of Guárico, where rainfall for 10-day periods is available for 80 different locations. The measurement network is relatively sparse for some areas, and aggregated rainfall at time resolutions of days or less is of very poor quality or nonexistent. We consider a model for rainfall based on a truncated normal distribution that has been proposed in the literature. We assume that the data y_{it}, where i indexes location and t indexes time, correspond to normal random variates w_{it} that have been truncated and transformed. According to this model, the dry periods correspond to the (unobserved) negative values and the wet periods correspond to a transformation of the positive ones. The serial structure present in series of rainfall data can be modeled by considering a stochastic process for w_{it}. We use a dynamic linear model on $w_t = (w_{1t}, \ldots, w_{Nt})$ that includes a Fourier representation to allow for the seasonality of the data that is assumed to be the same for all sites, plus a linear combination of functions of the location of each site. This approach captures year-to-year variability and provides a tool for short-term forecasting. The model is fitted using a Markov chain Monte Carlo method that uses latent variables to handle dry periods and missing values.

KEY WORDS: Bayesian spatio-temporal models; Nonlinear time series; Non-stationary time series.

1. INTRODUCTION

Estimation and prediction of the amount of rainfall in time and space is a problem of fundamental importance in many applications in agriculture, hydrology, and ecology. In tropical and subtropical regions, rainfall variability is a strong limiting factor in food production, hydropower generation, and health-related issues. An increasing concern is to improve the understanding of the rainfall process under current and future scenarios of climate variability and change.

Usually, long-term high-resolution data in time and space are needed to evaluate the impact of rainfall variability in a particular region. Long-term records usually come from ground-based networks that are spatially sparse for many parts of the world, especially in the tropics and subtropics. Accumulated rainfall data usually are not available for periods of less than daily increments, and even daily rainfall datasets have many quality problems and missing values. Monthly rainfall datasets include more of stations and are more easily available from the different national data centers (Hutchinson 1995).

Modeling rainfall data from a statistical perspective presents the challenge of dealing with a time series that is nonstationary and has a skewed distribution that includes a point mass at 0. When observations are considered over a given area, the difficulties of adding a spatial component to the model must be dealt with. Under conditions of a changing environment induced by human activities, nonstationarities of rainfall are attributed to increasing concentrations of greenhouse gases and land use changes that affect land-atmosphere interaction processes including rainfall dynamics (see, e.g., Avissar and Liu 1996). This poses the challenge of using new rainfall modeling approaches involving time-varying rainfall parameters with measures of uncertainty capable of describing the evolution of rainfall under changing conditions. Traditional approaches in which the parameters of a rainfall model are assumed to be constant for a particular calibration period no longer will be valid in a changing climate.

Over the last 30 years, several models have been developed with increasing degrees of sophistication to capture the underlying physical dynamics that govern rainfall. One of the earliest attempts is that of LeCam (1961), further developed by Waymire and Gupta (1981), Cox and Isham (1988), and Rodríguez-Iturbe, Cox, and Isham (1987, 1988), in which stochastic point process–based models in space and time are used; more recently, Smith and Robinson (1997) presented a likelihood-based approach and used a Bayesian method. Despite the developments of such models and the number of statistical tools available to fit them, it often happens that the only available data come from ground-based networks with sparse spatial coverage, often spanning limited time periods; these are important limitations when sophisticated models are fitted and undermine their applicability. In tropical areas, usually coincident with developing countries, data quality problems make the analysis even more challenging.

A different approach was considered by Smith (1994), who following Stearn and Coe (1984), broke down the problem of describing the distribution of rainfall into two parts:

Bruno Sansó (*bruno@cesma.usb.ve*) is Associate Professor and Lelys Guenni (*lbravo@cesma.usb.ve*) is Full Professor, Department of Scientific Computing and Statistics, Universidad Simón Bolívar, Apartado 89000, Caracas 1080-A, Venezuela. Substantial parts of this work were carried out while the first author was visiting the Institute of Statistics and Decision Science (ISDS) of Duke University, the support received from ISDS as well as comments from Peter Müller, the editor, and two anonymous referees are gratefully acknowledged. Both authors have been partially supported by grant G97000592 of Consejo Nacional de Investigaciones Científicas y Tecnológicas.

a process of wet and dry periods and a positive skewed distribution for the amount of rainfall, given a wet period. Another approach that has been considered to model the distribution of rainfall is that of a truncated normal distribution (Stidd 1973). The basic idea is to let r be the observed rainfall at a certain site and time interval and suppose that there is a normal random variable w such that

$$r = \begin{cases} w^{\beta} & w > 0 \\ 0 & w \leq 0 \end{cases}, \tag{1}$$

where $\beta > 0$ produces a transformation of the truncated normal to allow for different shapes and tail behaviors of the distribution of r. Bardossy and Plate (1992) considered a multisite model based on (1) and a certain covariance structure for the underlying normal process. Glasbey and Nevison (1997) also considered a truncated normal distribution but with a different family of transformations. A related model common to the econometric literature is the Tobit model (for a Bayesian approach to Tobit models see, e.g., Chib 1992). The appealing feature of (1) is that it is based on a latent Gaussian process for which all of the traditional geostatistics and time series techniques can be used, making it suitable for a likelihood-based approach. Sansó and Guenni (1999a) considered an isotropic Gaussian random field with a seasonal drift as the underlying process and used Monte Carlo (MC) methods, which impute the latent variables, to explore the posterior distribution of the parameters. But even though the dataset comprised accumulated rainfall over long intervals (months), the model is not able to capture the lack of stationarity in the data.

A structurally simple and yet powerful way of modeling nonstationary time series is provided by dynamic linear models (DLM), as described by West and Harrison (1997). Time-varying seasonality, trends, and dependence on lagged values can easily be expressed by the locally linear structure used in DLMs. Furthermore, multivariate observations fit very naturally into the framework, allowing for spatiotemporal processes that are quite flexible but relatively easy to specify. The elegant basic theory of DLMs assumes linearity and normality, but extensions have been developed (see West and Harrison 1997, chap. 13, for a full discussion). Of particular interest are the recent developments that use MC methods to explore the distribution of the parameters of nonlinear non-Gaussian dynamic models like those of Carter and Kohn (1994) and Frühwirth-Schnatter (1994) or applications like the one presented by Cargnoni, Müller, and West (1997). Methods used to update the distributions of the parameters sequentially, as observations become available, were described by Gordon, Salmon, and Smith (1993), Pitt and Sheppard (1997), and Sheppard (1994). Recent applications of DLMs to spatiotemporal modeling were given by Tonellato (1997a,b).

In earlier work (Sansó and Guenni 1999b), we considered a DLM for the latent process w in (1) when there is only one station, so that no spatial structure is present in the model. In this article we extend that model to handle the spatial correlation for the joint distribution of several stations at each time period. In Section 2 we present the dataset that motivated this study, comprising rainfall collected in 80 stations during 16 years in a region of central Venezuela. We explain the model in Section 3. In Section 4 we discuss the problems related to fitting the proposed model, and in Section 5 we present the results. We end with a discussion of the results and future developments in Section 6.

2. VENEZUELAN RAINFALL DATA

We consider a data set collected in 80 stations in the state of Guárico, in central Venezuela, from January 1968 to December 1983. The data correspond to accumulated rainfall over 10-day periods. Rainfall accumulated over short periods (5, 7, or 10 days) can be suitable for agricultural planning as in the implementation of irrigation schedules, planting dates, and land preparation. A justification for this is that water stress conditions on the crops, which determine irrigation needs, and soil humidity conditions, which determine planting dates and land preparation, are the result of processes that occur over a few days. Monthly rainfall is very coarse for identifying time periods of crop failure and water stress or excess, making agricultural planning very ineffective on that time step. (Jackson 1989; Linacre 1992; Puche 1994; Ritchie 1992). Although a daily time step is more appropriate in most cases, Venezuelan daily rainfall databases have many quality problems, among them a substantial number of missing observations and the presence of values that correspond to rainfall accumulated over few days. Accumulated data are often due to observational problems.

Rainfall in the region is usually of the convective type, with high-intensity storms occurring late in the afternoon during the rainy season, producing highly variable amounts for consecutive periods of short duration. Data quality problems and the nature of rainfall itself lead us to the decision of working with rainfall accumulated over 10-day periods, which is a compromise between a monthly and a daily time step. In our data, if one or more days of the 10-day period are missing, then the whole 10-day observation is considered missing. Missing values are then handled as latent variables in the estimation process. Obviously, 10-day accumulated rainfall will be simpler to model than daily rainfall, because we will have a smaller proportion of 0 values. Furthermore, a higher level of aggregation produces distributions for the data that are more regular and closer to normality.

Information on the positions of the stations (in terms of their longitude and latitude, as well as altitude with respect to the sea level) is available. No other relevant covariate is available. The stations are scattered irregularly over an area of roughly 250×300 km north of the Orinoco river. Most stations are clustered in the northern part of the region, which is the most densely populated. Figure 1 shows the location of the 80 stations with respect to a map of Venezuela and the time series for three of the stations.

Analysis of the series clearly shows a marked seasonality, with alternating periods of dry weather and periods of rather intense rainfall in yearly cycles. This is typical of the

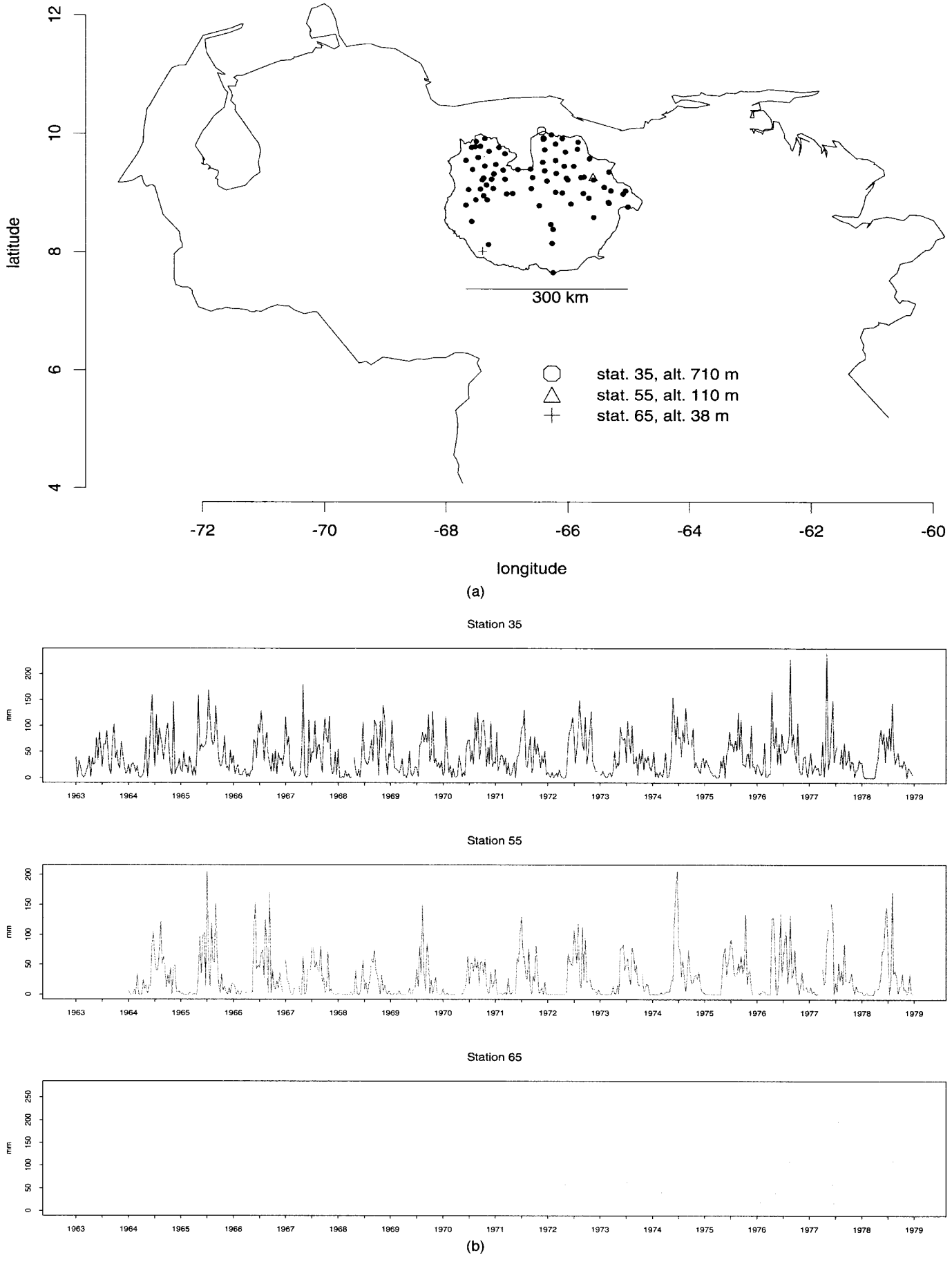

Figure 1. Locations of the 80 Stations Under Study and Observed Values of Rainfall at Three Different Locations.

dry–wet season behavior of tropical rainfall. The observed seasonality has a similar pattern in all the stations, which can be explained by the fact that the dominating mechanism for the production of rainfall in the area is the convective activity of the Inter-Tropical Convergence Zone. This also explains why the data show very little dependence on altitude, even though there is a decreasing gradient of altitude from north to south, from more than 700 to less than 100 meters. Several missing values are present, not only occasional ones, but also runs of missing values that span periods of several months. The interannual variability is an important feature of the data that is more evident after seasonality is taken into account and does not seem to correspond to any specific pattern or trend. This behavior can only be explained with a model that adapts dynamically to changes in observations from year to year.

3. A SPATIOTEMPORAL DYNAMIC MODEL

Let r_{it} be the rainfall collected at time $t = 1, \ldots, T$ on site $i = 1, \ldots, N$. Usually, r_{it} refers to rainfall accumulated over a specific time period. The application that we present herein considers rainfall accumulated over 10-day periods as discussed in the previous section. We propose the following model for r_{it}:

$$r_{it} = \begin{cases} w_{it}^{\beta} & \text{if } w_{it} > 0 \\ 0 & \text{if } w_{it} \leq 0, \end{cases} \tag{2}$$

where $\beta > 0$ and the latent variables w_{it} can be grouped as vectors $\mathbf{w}_t = (\mathbf{w}_{1t}, \ldots, \mathbf{w}_{Nt})'$, which follow a multivariate dynamic linear model,

$$\mathbf{w}_t = \mathbf{F}'\theta_t + \varepsilon_t \qquad \varepsilon_t \sim \mathrm{N}_N(0, \sigma^2 \mathbf{V}_t) \tag{3}$$

and

$$\theta_t = \mathbf{G}\theta_{t-1} + \eta_t \qquad \eta_t \sim \mathrm{N}_k(0, \mathbf{W}_t), \tag{4}$$

where $\sigma^2 > 0, \mathbf{V}_t \in \Re^{N \times N}, \theta_t \in \Re^k, \mathbf{G} \in \Re^{k \times k}, \mathbf{F}' \in \Re^{N \times k}$, and $\mathrm{N}_j(0, \mathbf{H})$ is used to denote a j-variate normal distribution with mean 0 and covariance matrix $\mathbf{H}$. We call (3) and (4) the observation equation and the evolution equation. The matrices $\mathbf{F}, \mathbf{G}$, and $\mathbf{V}_t$ can be appropriately chosen to model $\mathbf{w}_t$ as a multivariate nonstationary process with a seasonal pattern reflecting the seasonality of rainfall and a given spatial structure reflecting the correlation between sites. $\mathbf{W}_t$ determines the variability produced by each innovation in θ_t, the parameters that describe the trends in the process $\mathbf{w}_t$. The use of a transformed normal distribution is justified to provide a heavy-tailed distribution to accommodate extreme rainfall events.

We write $\theta_t = (\gamma_t, \theta_{t2})$ such that γ_t is a subvector that describes a spatial trend and θ_{t2} describes seasonality. A natural way to model the long-range spatial variation of the process $\mathbf{w}_t$ is to consider a trend that is expressed as a linear combination of functions of the locations of the stations, say $\mathbf{X}_{\gamma_t}$, so that for location i, we have $x_i'\gamma_t = (f_1(z_i), \ldots, f_q(z_i))'\gamma_t$, where γ_t is a vector of dimension q, z_i denotes some information about the location (usually longitude and latitude, but it could also include additional covariates), f_j are known functions, and x_i is the ith row of the $N \times q$ matrix $\mathbf{X}$. By letting γ_t depend on t, we are considering a trend that can vary with time in a way that will be controlled by the evolution equation.

To capture the seasonal variability of rainfall, we introduce θ_{2t}, the second set of components of θ_t, which will evolve periodically. One way to do this is to consider a Fourier representation of seasonality, which, following West and Harrison (1997), is achieved by an evolution matrix $\mathbf{G}$ that is a block-diagonal matrix of 2×2 blocks, where each block corresponds to a harmonic and the block associated with the rth harmonic is given by

$$\mathbf{G}_r = \begin{pmatrix} \cos(2\pi r/p) & \sin(2\pi r/p) \\ -\sin(2\pi r/p) & \cos(2\pi r/p) \end{pmatrix},$$

where $r = 1, \ldots, m$ and $p = 2m$ is the period. If l harmonics are used, then $\theta_{2t} \in \Re^{2l}$ and $k = q + 2l$ is the total number of parameters for a given time t. The matrix $\mathbf{F}'$ in the observation equation has a column of 1's for the first parameter of the harmonic block and a column of 0's for the second parameter. This will produce a seasonal trend that varies with time and adapts to small changes in phase and amplitude, but is identical for all of the stations, regardless of their position.

When the spatial and seasonal trends are considered together, we obtain a model for which $\mathbf{F}' = (\mathbf{X}, \mathbf{E}', \ldots, \mathbf{E}')$, where $\mathbf{E}$ is a $2 \times N$ matrix that has a first row of 1's and a second row of 0's and is repeated l times. On the other hand, if the evolution of γ_t follows a random walk, then $\mathbf{G} = \mathbf{BD}(\mathbf{I}_{q \times q}, \mathbf{G}_{r_1}, \ldots, \mathbf{G}_{r_l})$, where $\mathbf{BD}$ denotes a block-diagonal matrix, then $\mathbf{I}_{q \times q}$ is a q-dimensional identity matrix, and $r_1, \ldots, r_l$ are the harmonics being considered.

To capture short-range dependence between sites, we need to model the spatial correlation. Nonstationarity of the random field can be handled by considering a completely unknown correlation matrix $\mathbf{V}_t$. This leads to a highly parameterized model that calls for an informative prior distribution for $\mathbf{V}_t$. Brown et al. (1995) considered such an approach for spatial interpolation problems and discussed elicitation of the hyperparameters of the prior using an empirical Bayes technique. A simpler approach is to assume a parametric structure for $\mathbf{V}_t$; in particular, if we suppose that $\mathbf{w}_t$ is isotropic (i.e., the correlation between two sites depends only on the distance between them), then we can use the Matérn class of spatial correlation functions proposed by Handcock and Wallis (1994). This class is described by two parameters: ν controls the smoothness of the random field, and λ controls the range of the spatial correlation. Fixing $\nu = 1/2$ produces an exponentially decaying correlation, frequently used in the hydrological literature (see e.g., Bras and Rodríguez-Iturbe 1985), that, after dropping t, has the form $\mathbf{V}_{ij} = \exp(-\lambda d_{ij}), \lambda > 0$ and d_{ij} is the distance between sites i and j.

Spatial models commonly include an extra source of variability, commonly known as a *nugget effect,* which is equivalent to introducing an additional error for w_t, this can be

added to our original model to obtain

$$r_{it} = \begin{cases} w_{it}^{\beta} & \text{if } w_{it} > 0 \\ 0 & \text{if } w_{it} \le 0, \end{cases}$$

$$w_t = z_t + \nu_t \qquad \nu_t \sim \mathrm{N}_N(0, \tau^2 \mathbf{I}),$$

$$z_t = \mathbf{F}'\theta_t + \varepsilon_t \qquad \varepsilon_t \sim \mathrm{N}_N(0, \sigma^2 \mathbf{V}_t),$$

and

$$\theta_t = \mathbf{G}\theta_{t-1} + \eta_t \qquad \eta_t \sim \mathrm{N}_k(0, \mathbf{W}_t),$$

which implies that the spatial long-range dependencies, as well as the temporal fluctuations, are affected by normal perturbations before being transformed. (For a comprehensive discussion of geostatistical models, see, e.g., Cressie 1993.)

The covariance matrix $\mathbf{W}_t$ in the evolution equation can be determined dynamically by the use of discount factors (see West and Harrison 1997, chap. 6). A discount factor determines the amount of information loss through the process evolution in time. Let $P_t = \mathrm{var}(\mathbf{G}_t\theta_{t-1}|\mathbf{D}_{t-1})$; that is, the variance of $\mathbf{G}_t\theta_{t-1}$ where $\mathbf{D}_{t-1}$ denotes all the information available at time $t-1$. Then $\mathrm{var}(\theta_t|\mathbf{D}_{t-1}) = P_t + \mathbf{W}_t$. The use of a discount factor $0 < \delta \le 1$ sets $\mathrm{var}(\theta_t|\mathbf{D}_{t-1}) = P_t/\delta$ and, implicitly, $\mathbf{W}_t = P_t(1-\delta)/\delta$. Note that $\delta = 1$ corresponds to a static model where θ_t is fixed for all t, while small values of δ correspond to heavy discounting of the information available at time $t-1$. The usual range of values for δ is .8–1.

To complete the model, we need to specify a prior distribution for the parameters $\sigma^2, \tau^2, \theta_0, \lambda$, and β. We assume independence between the five parameters and propose a reference prior $1/\sigma^2$ for σ^2, a multivariate normal with mean m_0 and variance $\mathbf{C}_0$ for θ_0 and gamma densities with parameters a_τ, b_τ for τ^2, a_λ, b_λ for λ and a_β, b_β for β. Another hyperparameter must be specified; this is the discount factor δ. The choice of hyperparameters will be discussed as part of the analysis of the data.

4. FITTING THE MODEL

To fit the model presented in the previous section, we use a Markov chain Monte Carlo method (MCMC), as described by, for example, Smith and Roberts (1993), to obtain samples from the posterior distribution of $\sigma^2, \tau^2, \lambda, \beta, z = (z_1, \ldots, z_T)$ and $\theta = (\theta_1, \ldots, \theta_T)$, where each θ_i has dimension q. Analogous to the technique used to model censored observations, we consider latent variables v_t to model dry periods. The former means that an unknown quantity $v_{it} < 0$ is introduced in the likelihood every time that we observe a dry period; v_t is then treated as a parameter and inputed at each cycle of the Markov chain simulation. We use similar latent variables u_t to account for the missing values, without imposing any restrictions on u_{it}. Note that the dimensionality of u_t and v_t may change with time.

We can now redefine $\mathbf{w}_{it}$ as

$$\mathbf{w}_{it} = \begin{cases} u_{it} & \text{if } r_{it} \text{ is missing} \\ r_{it}^{1/\beta} & \text{if } r_{it} > 0 \\ v_{it} & \text{if } r_{it} = 0, \end{cases}$$

and observe that $\mathbf{w}_t$ is a vector for which some of the components are transformations of the observed (positive) data, some are unknown negative values, and some are unknown unrestricted values. Clearly, for each iteration of the MCMC, we will have imputations of β, v_t, and u_t, and so actual values for the components of $\mathbf{w}_t$ will be available. Our goal is to sample from a posterior distribution that is proportional to

$$\left(\frac{1}{\sigma^2}\right)^{NT} \left(\frac{1}{\rho^2}\right)^{NT/2} |\mathbf{V}(\lambda)|^{-T/2} \times \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} \frac{1}{\rho^2} \|\mathbf{w}_t - z_t\|^2 + (\mathbf{w}_t - \mathbf{F}'\theta_t)'\mathbf{V}(\lambda)^{-1}(\mathbf{w}_t - \mathbf{F}'\theta_t) - \frac{1}{2}\sum_{t=1}^{T} (\theta_t - \mathbf{G}\theta_{t-1})'\mathbf{W}_t^{-1}(\theta_t - \mathbf{G}\theta_{t-1}) \right) \times \left(\prod_{r_{it}>0} \frac{r_{it}^{1/\beta_t - 1}}{\beta_t} \right) p(\theta_0, \lambda, \beta, \sigma^2, \rho^2),$$

where $\rho^2 = \tau^2/\sigma^2$.

The full conditionals for σ^2 and ρ^2 can be easily derived as inverse gamma distributions, so it is straightforward to obtain samples from them. The full conditionals for λ and β do not correspond to the functional form of any known density and so, following Müller (1991), we propose the use of a Metropolis step. This proceeds as follows: We first transform λ to the log scale, then we obtain a proposal from a uniform distribution centred at the log of the last imputed value of λ, and finally we accept the proposal with a probability that is given by the usual Metropolis–Hastings algorithm (see, e.g., Chib and Greenberg 1995). We proceed in a similar way to obtain samples of β.

To obtain samples of u_t, let $\mathbf{w}_t = (u_t, \mathbf{w}_{t2})$ and, accordingly, split z_t as (z_{t1}, z_{t2}), the full conditional of u_t is a multivariate normal with mean z_{t1} and covariance matrix $\sigma^2\rho^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

By a similar argument, we can see that the full conditional distribution of v_t is a truncated multivariate normal with mean obtained by splitting z_t and covariance matrix a multiple of the identity, which facilitates the sampling. Nevertheless, care must be taken when the components of z_t are large and positive, because then the distribution is concentrated in a region away from the negative values.

The full conditional of z_t is a multivariate normal with covariance matrix $\mathbf{C} = (\mathbf{I}/\rho^2 + \mathbf{V}(\lambda)^{-1})^{-1}$ and mean $\mathbf{C}(\mathbf{w}_t/\rho^2 + \mathbf{V}(\lambda)^{-1}\mathbf{F}'\theta_t)$.

Samples of $\theta = (\theta_1, \ldots, \theta_T)$ can be obtained using the algorithm proposed by Carter and Kohn (1994) and Frühwirth-Schnatter (1994), which produces a sample from the joint density of θ. The procedure is known as *forward filtering backward sampling* (FFBS) and begins with calculating the posterior moments $E(\theta_t|\mathbf{D}_t, \sigma^2, \beta, \lambda) = m_t$ and $\mathrm{var}(\theta_t|\mathbf{D}_t, \sigma^2, \beta, \lambda) = \mathbf{C}_t$, using the standard results from

dynamic linear models (forward filtering). The backward sampling starts by obtaining a sample $\theta_T \sim \mathrm{N}_k(m_T, \mathbf{C}_T)$ and then recursively sampling $\theta_t \sim \mathrm{N}(m_{t|t+1}, \mathbf{C}_{t|t+1})$, where $m_{t|t+1} = m_t + A_t(\theta_{t+1} - \mathbf{G}m_t), \mathbf{C}_{t|t+1} = \mathbf{C}_t - A_t Q_t A_t', Q_t = \mathbf{G}\mathbf{C}_t\mathbf{G}' + \mathbf{W}_t$, and $A_t = \mathbf{C}_t\mathbf{G}'Q_t^{-1}$, for $t = T-1, \ldots, 1$.

5. RESULTS

We used the model and the MCMC algorithm described in the previous sections to fit the Guárico rainfall data. For the hyperparameters needed to specify the priors, we observed that posterior inference was not sensitive to the specific choice of m_0 and $\mathbf{C}_0$, so we chose m_0 as a vector of 0's and $\mathbf{C}_0$ as the identity matrix. The prior distribution of β was chosen as a gamma density highly concentrated around 3, using the information provided by previous fits. The prior distribution of λ was chosen as a gamma density concentrated between 4 and 8 with mean 6. To obtain this prior, we considered observations r_t at a specific t, with few missing data and no 0 values, transformed them using $1/\beta$ with $\beta = 3$, and estimated λ using the empirical variogram; the parameters of the prior were chosen to provide fairly vague prior information around the estimated value. The prior for ρ^2 was taken as a fairly vague gamma distribution. The discount factor δ is regarded as a tuning parameter in the fitting process. We considered three different discount factors: one for the "intercept," a second one for the rest of the parameters that correspond to the spatial trend, and a third one for the seasonality parameters. The values were chosen as .85, .90, and .95, reflecting the fact that the spatial trend is allowed to vary quite substantially while the amplitude and the phase of the seasonal components are expected to be fairly stable, given that changes from one season to the other occur quite regularly.

Figure 2 shows some posterior quantiles for the parameters of the dynamic part of the model from 1968 to 1981. The last 2 years where omitted and were used to assess the behavior of the predictions. Let $\mathbf{x}_i = (x_{i1}, x_{i2})$ denote the centered longitude and latitude of station i. We modeled the spatial trend as

$$x_i'\gamma_t = \gamma_{0t} + \gamma_{1t}x_{i1} + \gamma_{2t}x_{i2} + \gamma_{3t}x_{i1}^2 + \gamma_{4t}x_{i2}^2 + \gamma_{5t}x_{i1}x_{i2},$$

a second-order polynomial chosen based on graphical exploration and previous analyses of the data. We observe that the spatial trend for these data is essentially driven by an intercept that varies substantially over time, plus first- and second-order effects due to the latitude. It is interesting to notice the very strong seasonal pattern present in the coefficient of the latitude, despite the fact that most of the seasonality in the data is captured by two Fourier harmonics. The pattern observed provides evidence that a linear effect of latitude is more relevant at the beginning of the year, during the dry season. On the other hand, there is no significant multiplicative effect of longitude and latitude.

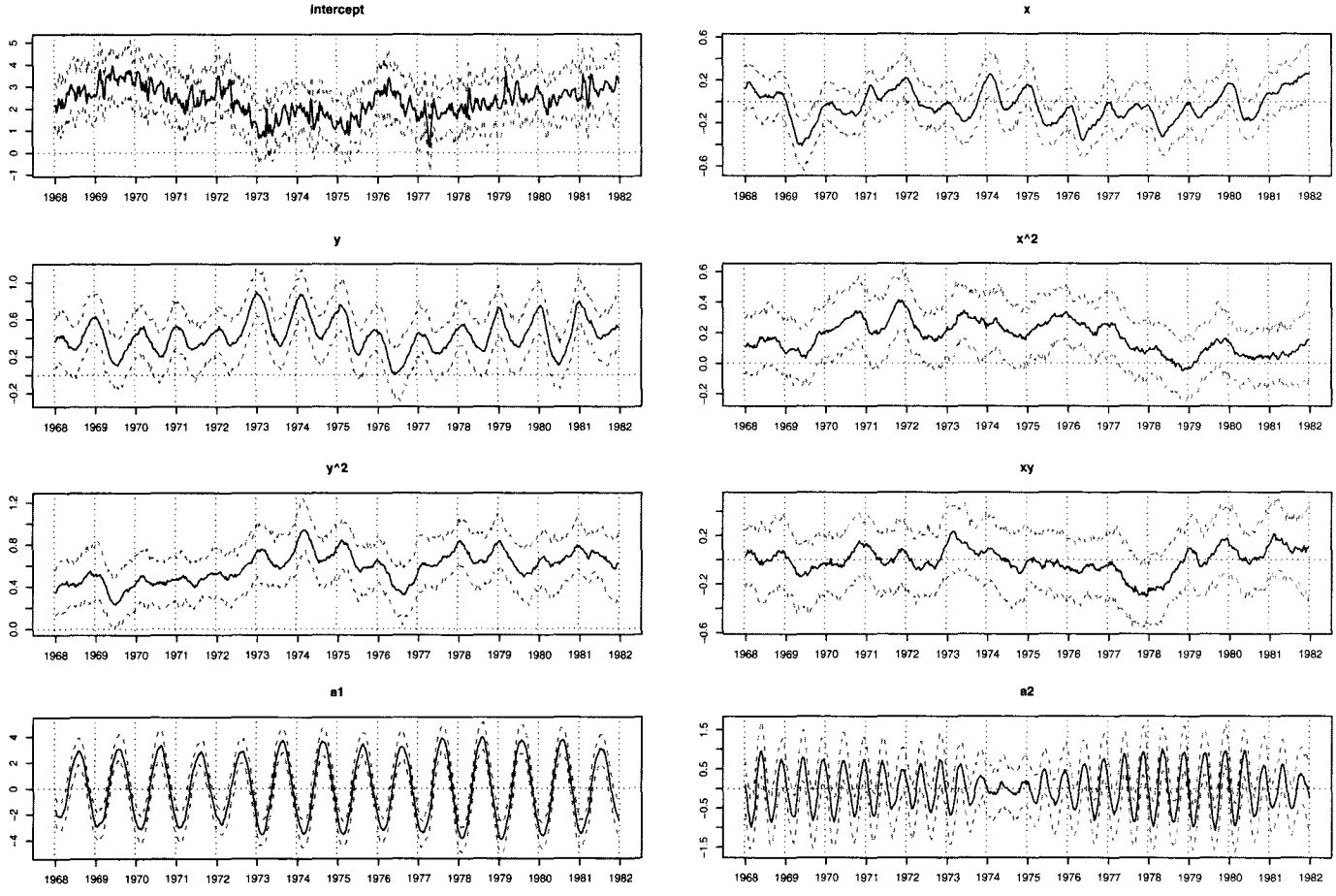

Figure 2. Posterior Distributions of the Parameters. Dotted lines indicate the boundaries of 90% probability intervals; solid lines indicate medians.

Table 1. Posterior Quantiles Based on 1,800 Samples Taken Every 50 Iterations After a Burn-In of 4,000 Samples

	2.5%	50%	97.5%
σ^2	1.646	1.751	1.859
λ	1.526	1.630	1.740
β	2.343	2.365	2.388
ρ^2	.666	.713	.765

To assess the number of harmonics to be used in the model, we considered the first four harmonics and observed that the first two were the only harmonics to have an effect significantly different from 0 most of the time. In the plot we show the first component of each harmonic indicated by a_1 and a_2. It is interesting to note that whereas the first harmonic is fairly regular along time, the second disappears during 1972–1976 and 1981–1982. This reflects the fact that for these years, the second harmonic did not play an important role in describing precipitation. For this region, two precipitation maxima might occur due to the presence of a rather long period of low rain in the middle of the rainy season. This period is a characteristic feature of rainfall in the Venezuelan central plains, as discussed by Puche (1994). For some years, this feature is not evident from the data, and a single harmonic is enough to represent the seasonal cycle.

The results that we present are based on 90,000 samples from a chain after a burn-in of 4,000 iterations. We assessed convergence by running the chain for four different starting values, changing the initial seed of the random number generator, and monitoring the samples of σ^2, ρ^2, β, and λ. The values of the Brooks, Gelman, and Rubin convergence diagnostic for each of the parameters, as well as the multivariate test, were all below 1.11, a threshold suggested by Gamerman (1997) as satisfactory, after 4,000 iterations. We noticed that samples stabilize after few iterations, although a strong autocorrelation is present. We ran the Raftery and Lewis convergence test (Raftery and Lewis 1992) and obtained that for all four parameters, the suggested burn-in was less than 800, the suggested thinning was less than 20, and fewer than 50,000 iterations were needed to achieve an accuracy of $\pm.01$, with probability .95, in the estimation of the .025 and .95 quantiles. All of these results where obtained running the set of routines BOA for R, developed by Brian Smith (http://www.public-health.uiowa.edu/boa/).

Quantiles from the posterior distributions of σ^2, λ, β, and τ^2 are given in Table 1. To assess the width of the uniform distribution used as a proposal for the samples of λ and β,

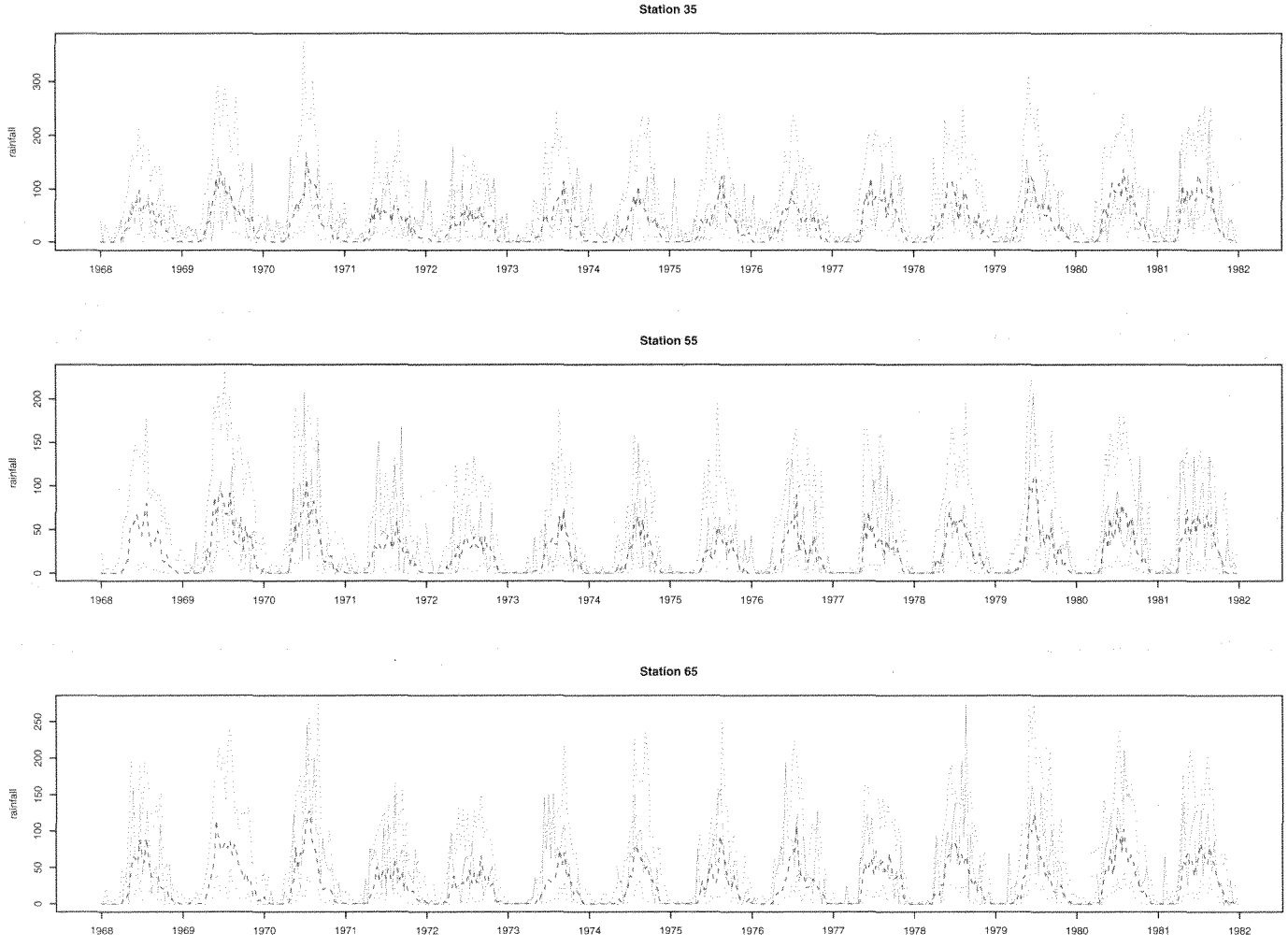

Figure 3. Median and 90% Interval of the Posterior Predictive Distribution of Rainfall. Solid lines correspond to observed data; dashed lines, to median; and dotted lines, to extremes of the 90% interval.

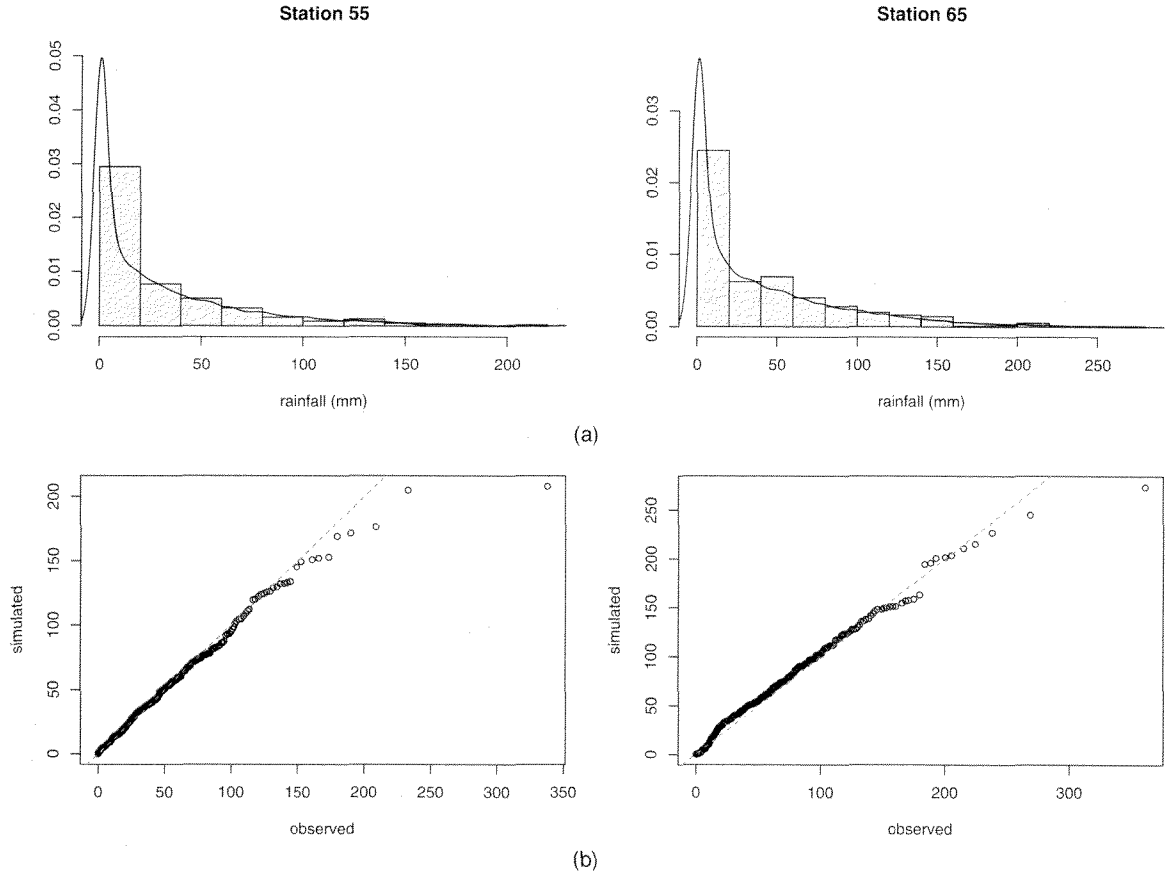

Figure 4. Estimated Density of Samples From the Predictive Posterior of Rainfall at Two Given Stations Superimposed Over the Histogram of the Corresponding Observed Data (a) and Quantile-Quantile Plots of Sampled Versus Observed Data (b). The dotted line has intercept 0 and slope 1.

we estimated the variance of the samples in the burn-in. We experimented with different values of the width and also with a normal instead of a uniform jumping distribution, and obtained very similar results. The median value of $\lambda = 1.63$ is such that only when stations are more than 180 km apart is the correlation below .05. Note that correlations are referred to w_t, and that the nonlinear transformation that gives r_t will lower them. In fact, simulations based on the values in Table 1 show that a given correlation between pairs of r_t requires a distance between sites that is 15–25% smaller than the distance needed to produce the same correlation between pairs of w_t.

The fit of the model can be explored by simulating from the posterior predictive distribution. This is done by imputing the values of $\mathbf{w}$ and then truncating all negative values and transforming the positive ones. Quantiles of these simulations are compared with the observed data in Figure 3 for the three stations considered in Figure 1. We considered only 100 samples from the chain, because of the difficulty in storing and handling the large file produced by the full output of the chain. These samples were taken every 100 iterations after burn-in. We observed that the historical rainfall values are mostly within the limits of the 90% interval of the posterior predictive distribution. Figure 4 graphically compares the distributions of the observed and the sampled data, and Table 2 gives a numerical comparison for the probability of dry periods and the exceedence probabilities of rainfall greater than 130 mm for 10 of the stations.

Simulated rainfall values are usually required for many applications in which rainfall information must be used as the driving input variable in many hydrological, agricul-

Table 2. Probabilities of a Dry Period and Probabilities of More Than 130 mm of Rain, for 10 Stations Chosen at Random

		Station									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
	2.5%	.220	.225	.218	.202	.155	.207	.221	.183	.204	.225
Dry	Observed	.272	.226	.250	.236	.070	.216	.287	.221	.251	.279
	97.5%	.274	.273	.266	.255	.230	.248	.277	.227	.255	.265
	2.5%	.016	.010	.017	.017	.031	.017	.014	.026	.020	.016
High	Observed	.026	.020	.021	.012	.026	.021	.030	.022	.023	.011
	97.5%	.039	.037	.037	.045	.054	.044	.039	.046	.042	.039

NOTE: The 2.5% and 97.5% quantiles are estimated from samples of the predictive posterior, and the empirical estimates are obtained from the observed data; 130 mm is the 97.5% quantile of the whole dataset.

tural, and ecological models. Figure 5 shows simulations from the predictive posterior distributions of some stations. We observe fairly similar patterns to those in Figure 1, and an ensemble of such simulations could be used as input data for the aforementioned models, with the additional advantage of not having missing information and with similar statistical properties of the original rainfall data. The spatial behavior of the fitted rainfall can be observed in Figure 6, where a more complex pattern is observed in the northern part of the region and a smoother trend of the median isolines is observed toward the southern part of the state. Although this is expected because of the more complex terrain toward the north, there is also an effect due to the larger data sparsity in the south than in the north.

To explore the goodness of the fit obtained by our model, we considered as "residuals" for iteration i the difference between the imputed values of $\mathbf{w}_t^i$ and those of $\mathbf{F}'\theta_t^i$, standardized with the imputed values of σ^{2i} and ρ^{2i}. Note that although some of the components of $\mathbf{w}_t^i$ may be generated by the sampler as normal variates, most of them correspond to a transformation of the actual rainfall, and so these residuals contain information about the goodness of fit. We explored several features of these residuals, including normal plots and autocorrelation functions, which provide evidence that they have a behavior close to white noise in time. To assess the spatial fit, we estimated the spatial correlogram from the residuals and compared it to the fitted exponential correlogram, obtaining acceptable results.

5.1 Dry Periods and Aerial Rainfall

The sampling-based method that we use is well suited for estimating two important measures of rainfall activity over a specific area: the probability of a dry period at a given time and location and the amount of aerial rainfall; that is, the average rainfall depth over a region. The first measure can be estimated as the proportion of negative $\mathbf{w}_t$ in a sample from the posterior predictive distribution. The second, defined as

$$a(t) = \frac{1}{\nu(A)} \int_A r(s,t)\, ds,$$

where s denotes location and $\nu(A)$ denotes the area of a given region A, can be estimated by considering a grid of points in A, say (x_i, y_i) for $i = 1, \ldots, L$; sampling the pos-

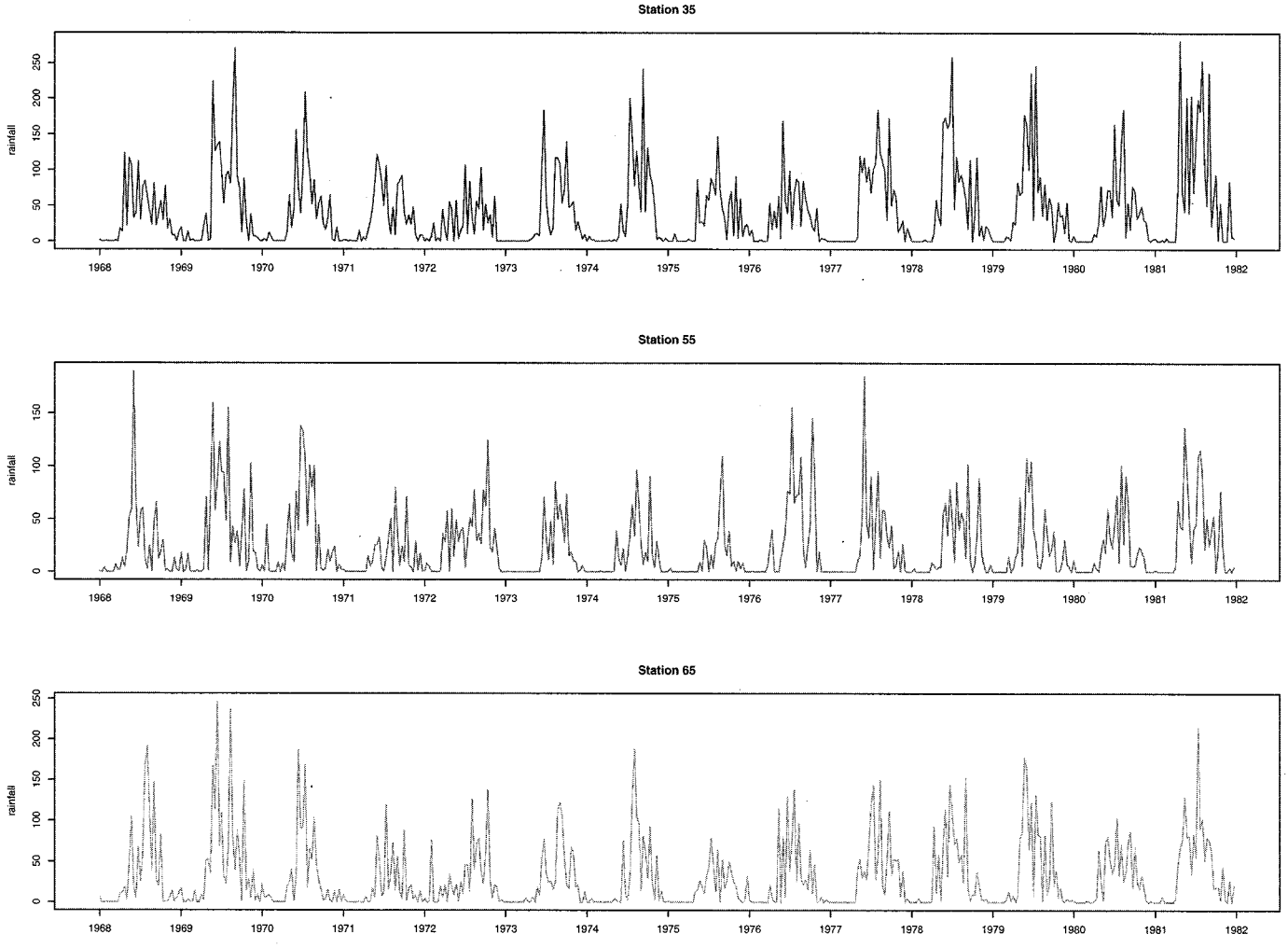

Figure 5. Simulations From the Predictive Posterior Distribution of Rainfall for Three Different Stations.

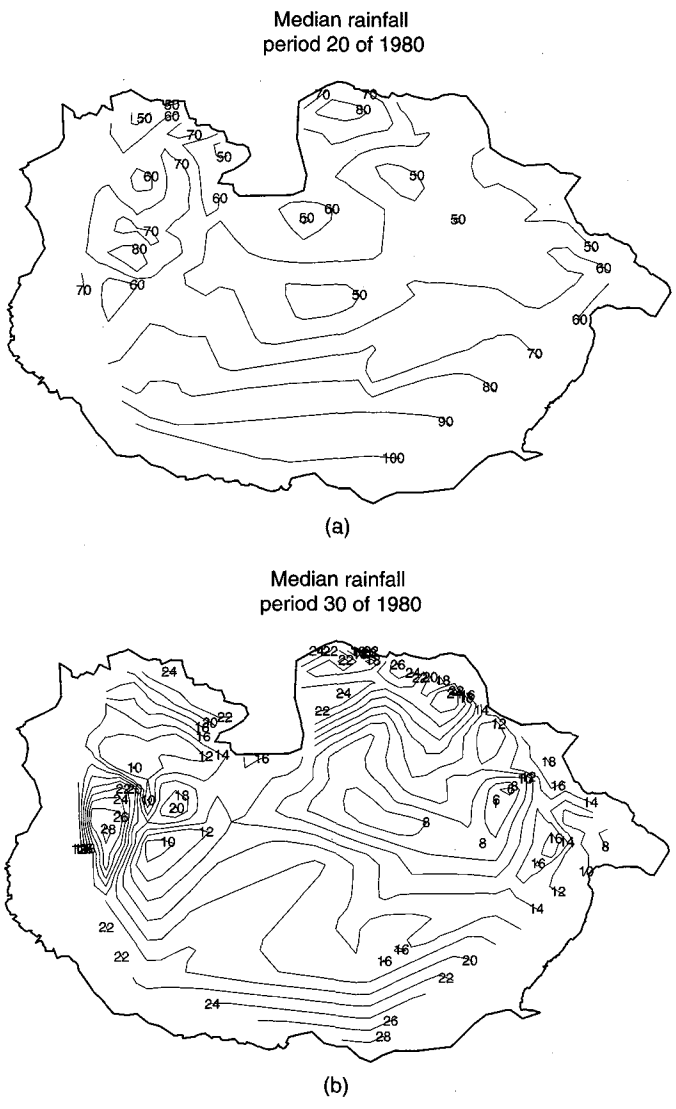

Figure 6. Median of the Predictive Posterior Distribution of Rainfall for Two Different Periods of 1980.

terior predictive distribution of the rainfall at those points; and then averaging the results. This process can be repeated several times to obtain various samples of $a(t)$.

5.2 Predicting Future Observations

The model and the methodology that we have presented in the preceding two sections are well suited for producing short-term predictions. In fact, given samples from the posterior distribution of θ_T, we can simulate from the distribution of θ_{T+h} using the evolution equation. To obtain the samples that correspond to θ_T^j, we use the covariance matrix $\mathbf{W}_{T+h}^j = \mathbf{G}\mathbf{C}_Y^j\mathbf{G}'(1-\delta)/\delta$; that is, no further discount is considered after the last observation has been observed. Once a sample of θ_{T+h} becomes available, we can obtain a sample from the predictive density of the rainfall at time $T+h$ by sampling from a N-variate normal distribution centered at $\mathbf{F}'\theta_{T+h}$, with covariance matrix $\sigma^{2j}\mathbf{V}(\lambda^j)+\tau^2\mathbf{I}$. The results of following this procedure are shown in Figure 7 for the same three stations considered before. Median values represent the seasonal pattern reasonably well. Most observed values are within the probability intervals.

6. CONCLUSIONS

We have considered a model based on the idea that rainfall can be described by a fairly simple mechanism consisting of a truncated and transformed process that evolves according to a multivariate dynamic linear model (DLM). The multivariate structure of the DLM allows for the modeling of time-varying long-range spatial dependencies and seasonal trends, using linear structures that evolve with time. Short-range spatial dependencies are modeled using an exponentially decaying correlation based on the assumption of isotropy. The model is fitted using Monte Carlo methods that provide the means of exploring the posterior distribution of the parameters, as well as the predictive distribution of any quantity of interest such as rainfall, aerial rainfall, or probability of a dry period.

The results show that despite the simplifying assumptions on which the model is based, it is possible to obtain a description of the behavior of rainfall for a moderately large number of stations, some of which present data that are not of the highest quality. The model also produces predictions that are well within acceptable ranges.

Several extensions and refinements for the actual model are possible. It would be desirable to have a more systematic assessment of the values of the discount factors δ. The approach suggested by West and Harrison (1997) involves evaluating the marginal distribution of the data, as a function of the discount factor, and then maximizing it in a maximum likelihood fashion. For this example, such an approach presents the difficulty of evaluating the marginal from a random sample of a high-dimensional parameter.

Assuming that a time-varying λ is also a desirable refinement, an evolution equation for λ could be specified as $\log(\lambda_t) = \log(\lambda_{t-1})+$ error, for a normal error, using a structure similar to that in the Metropolis step used to impute λ. Introducing a time-varying σ^2 also could be considered. An obvious extension of the correlation structure is to consider a wider family of correlation functions, like the Matérn class or a family based on a spectral representation. But it is doubtful that substantial improvements in model fit will result from this, given that, on average, stations are quite distant (only 14% of all distances are less than 50 km) and thus short-distance dependence is hard to estimate. Moreover, the observations result from a nonlinear transformation of the latent variables, which dampens the correlation.

The proposed model is fitted with a Monte Carlo method that is computationally rather intensive (several hours of CPU time on a two-processor Sun Enterprise 450 to produce a few thousand iterations), so it is important to explore efficient methods of updating the posterior once new observations become available without having to redo the entire simulation. This is the subject of current research.

In terms of model applications, the possibility of producing short-term forecasting with a measurement of uncertainty is appealing when such information needs to be used as input for models in hydrology and agriculture. Simulated values from the predictive posterior distribution in space

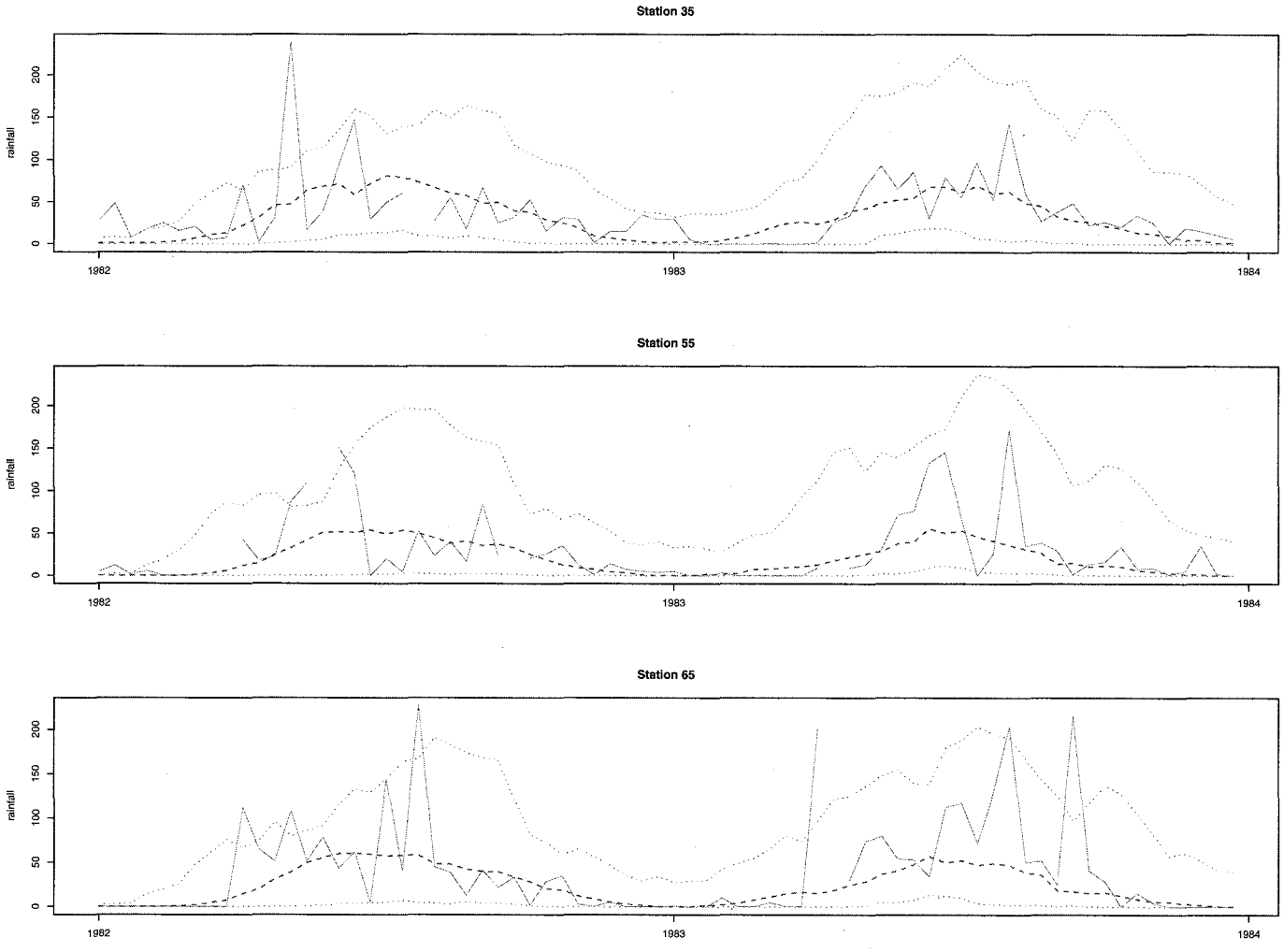

Figure 7. Forecasts for the Years 1982 and 1983 That Were Excluded From the Model Fit. Solid lines correspond to the observed data, dashed lines to the median and dotted lines to a 90% interval.

and time also reproduce the seasonal and spatial behavior of the rainfall data within the observed limits.

[Received May 1998. Revised June 2000.]

REFERENCES

Avissar, R., and Liu, Y. (1996), "A Three-Dimensional Numerical Study of Shallow Convective Clouds and Precipitation Induced by Land-Surface Forcing," *Journal of Geophysical Research*, 101, 7499–7518.

Bardossy, A., and Plate, E. J. (1992), "Space-Time Model for Daily Rainfall Using Atmospheric Circulation Patterns," *Water Resources Research*, 28, 1247–1259.

Bras, R. L., and Rodríguez-Iturbe, I. (1985), *Random Functions and Hydrology*, Reading, MA: Addison-Wesley.

Brown, P. J., Le, N. D., and Zidek, J. V. (1995), "Multivariate Spatial Interpolation and Exposure to Air Pollutants," *The Canadian Journal of Statistics*, 22, 489–509.

Cargnoni, C., Müller, P., and West, M. (1997), "Bayesian Forecasting of Multinomial Time Series Through Conditionally Gaussian Dynamic Models," *Journal of the American Statistical Association*, 92, 587–606.

Carter, C. K., and Kohn, R. (1994), "On Gibbs Sampling for State-Space Models," *Biometrika*, 81, 541–553.

Chib, S. (1992), "Bayes Inference in the Tobit Censored Regression Model," *Journal of Econometrics*, 51, 79–99.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis–Hastings Algorithm," *The American Statistician*, 49, 327–335.

Cox, D. R., and Isham, V. (1988), "A Simple Spatial-Temporal Model of Rainfall" (with discussion), *Proceedings of the Royal Society of London*, Ser. A, 415, 317–328.

Cressie, N. A. C. (1993), *Statistics for Spatial Data* (rev. ed.), New York: Wiley.

Frühwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models," *Journal of Time Series Analysis*, 15, 183–202.

Gamerman, D. (1997), *Markov Chain Monte Carlo*, London: Chapman and Hall.

Glasbey, C. A., and Nevison, I. M. (1997), "Rainfall Modelling Using a Latent Gaussian Variable," in *Modelling Longitudinal and Spatially Correlated Data*, eds. T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren, and R. D. Wolfinger, New York: Springer-Verlag, pp. 233–242.

Gordon, N. J., Salmon, D. J., and Smith, A. F. M. (1993), "A Novel Approach to Non-Linear and Non-Gaussian Bayesian State Estimation," *IEEE Proceedings*, F 140, 107–133.

Handcock, M. S., and Wallis, J. R. (1994), "An Approach to Statistical Spatial-Temporal Modeling of Meteorological Fields" (with discussion), *Journal of the American Statistical Association*, 89, 368–390.

Hutchinson, M. F. (1995), "Interpolation of Mean Rainfall Using Thin Plate Smoothing Splines," *International Journal of Geographical Information Systems*, 9, 385–403.

Jackson, J. T. (1989), *Climate, Water and Agriculture in the Tropics* (2nd ed.), London: Longman.

LeCam, L. (1961), "A Stochastic Description of Precipitation," In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, Berkeley, CA: University of California Press, pp. 165–186.

Linacre, E. (1992), *Climate Data and Resources*, London: Routledge.

Müller, P. (1991), "Metropolis-Based Posterior Integration Schemes," Technical Report 91-09, Purdue University, Dept. of Statistics.

Pitt, M. K., and Sheppard, N. (1997), "Filtering via Simulation: Auxiliary Particle Filters," Technical Report 1997-W13, University of Oxford, Nuffield College.

Puche, C. (1994), "Evaluation of the Water Region of Rainfed Agriculture in Areas of Seasonal Rainfall in Venezuela," doctoral thesis, University of Reading, U.K.

Raftery, A. E., and Lewis, S. M. (1992), "How Many Iterations in the Gibbs Sampler?," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 765–776.

Ritchie, J. T. (1992), "Classification of Crop Simulation Models," in *Crop Modelling and Related Environmental Data: A Focus on Applications for Arid and Semiarid Regions in Developing Countries*, eds. P. F. Uhlir and G. C. Carter, Paris: CODATA, pp. 3–14.

Rodríguez-Iturbe, I., Cox, D. R., and Isham, V. S. (1987), "Some Models for Rainfall Based on Stochastic Point Processes," *Proceedings of the Royal Society of London*, Ser. A, 410, 269–288.

——— (1988), "A Point Process Model for Rainfall: Further Developments," *Proceedings of the Royal Society of London*, Ser. A, 417, 283–298.

Sansó, B., and Guenni, L. (1999a), "Venezuelan Rainfall Data Analysed Using a Bayesian Space-Time Model," *Applied Statistics*, 48, 345–362.

——— (1999b), "A Stochastic Model for Tropical Rainfall at a Single Location," *Journal of Hydrology*, 214, 64–73.

Sheppard, N. (1994), "Partial Non-Gaussian State Space," *Biometrika*, 81, 115–131.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computations via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 3–25.

Smith, R. L. (1994), "Spatial Modelling of Rainfall Data," in *Statistics for the Environment 2: Water Related Issues*, eds. V. Barnet and K. Feridum Turkman, New York: Wiley, pp. 19–42.

Smith, R. L., and Robinson, P. J. (1997), "A Bayesian Approach to the Modelling of Spatial-Temporal Precipitation Data," in *Case Studies in Bayesian Statistics III*, eds. G. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch, P. Rossi, and N. D. Singpurwalla, New York: Springer-Verlag, pp. 237–269.

Stearn, R. D., and Coe, R. (1984), "A Model Fitting Analysis of Rainfall Data," *Journal of the Royal Statistical Society*, Ser. A, 147, 1–34.

Stidd, C. K. (1973), "Estimating the Precipitation Climate," *Water Resources Research*, 9, 1235–1241.

Tonellato, S. F. (1997a), "Bayesian Dynamic Linear Models for Spatial Time Series," technical report, Venezia University, Dept. of Statistics.

——— (1997b), "A Space-Time Analysis of Carbon Monoxide Atmospheric Concentrations," technical report, Venezia University, Dept. of Statistics.

Waymire, E. D., and Gupta, V. K. (1981), "The Mathematical Structure of Rainfall Representations. 3. Some Applications of the Point Process Theory to Rainfall Processes," *Water Resources Research*, 17, 1287–1294.

West, M., and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models* (2nd ed.), New York: Springer-Verlag.