Incorporating Spatial Dependence and Atmospheric Data in a Model of Precipitation

JAMES P. HUGHES

Department of Biostatistics, University of Washington, Seattle, Washington

PETER GUTTORP

Department of Statistics, University of Washington, Seattle, Washington

(Manuscript received 12 August 1993, in final form 30 April 1994)

ABSTRACT

Nonhomogeneous hidden Markov models (NHMM) provide a method of relating synoptic atmospheric measurements to precipitation occurrence at a network of rain gauge stations. In previous work it was assumed that, conditional on the current atmospheric pattern (termed a "weather state"), rain gauge stations in a network could be considered spatially independent. For a spatially dense network, this assumption is not tenable. In the present work, the NHMM is extended to include the case of spatial dependence by postulating an autologistic model for the conditional probability of rainfall given the weather state. Methods for fitting the parameters, assessing the goodness of fit of the model, and generating rainfall simulations are presented. The model is applied to a network of 24 stations in the Puget Sound region of western Washington State.

1. Introduction

General circulation models (GCMs) are, at present, the best tool available for modeling large-scale atmospheric circulation patterns and for determining the consequences of changes in the atmosphere (e.g., increased CO_2 concentration). Recent work (Zorita et al. 1995) has shown that GCMs can replicate key aspects of large-scale atmospheric circulation patterns. However, GCMs have been less successful in modeling local or regional-scale phenomena such as precipitation (see, e.g., Rind et al. 1989). This problem is especially acute in areas such as the Pacific Northwest, where local topography (on scales that cannot, at present, be integrated into GCMs) heavily influences precipitation patterns. Nonetheless, accurate simulations of precipitation are necessary to determine the potential effects of alternative climate hypotheses such as the greenhouse effect.

In an effort to improve the quality of precipitation sequences, Hay et al. (1991) introduced a class of stochastic models known as "weather state" models, which link synoptic-scale information such as sea level pressure with local precipitation. The idea is to explicitly classify each day into one of a small number of characteristic weather states based on synoptic atmospheric data. Precipitation is then assumed to be conditionally (temporally) independent given the weather states. Precipitation probabilities and distributions are estimated separately for each weather state based on historical data. The effect of altered climate hypotheses on precipitation may be evaluated by using altered climate atmospheric data as input into the fitted model to generate altered climate precipitation sequences. The validity of the fitted model is evaluated by generating simulated precipitation sequences using historical atmospheric data as input and comparing the statistics of the simulated precipitation sequence to those of the historical precipitation data. One such statistic that seems to be particularly difficult to reproduce is the distribution of the length of wet and dry periods (the duration distributions) at each station. The typical weather-state model tends to fit a distribution that does not produce enough long runs of wet and dry days. Various ad hoc corrections have been used to overcome this problem (see, e.g., Wilson et al. 1992; and Hughes et al. 1993).

Hughes and Guttorp (1994), building on the earlier work of Zucchini and Guttorp (1991), propose an alternative model formulation that includes the weather-state models as a special case. In their approach, which they refer to as the nonhomogeneous hidden Markov model (NHMM) [see Rabiner and Juang (1986) for a basic reference on hidden Markov models], the weather state is not explicitly defined. Rather the weather state is viewed as an unobserved (or hidden) stochastic process that evolves as a first-order Markov chain. The transition probabilities from one time to the next (and hence the probability of being in any particular state) are determined by the previous day's state and the current values of the synoptic atmospheric

Corresponding author address: Dr. James P. Hughes, Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195.

variables. Thus, on any given day, instead of specifying a single weather state the probability distribution of the weather states is given. However, through appropriate parameterizations this probability distribution may be made degenerate giving the explicit weather state models described above as a special case. The distribution of rainfall or rain occurrence is, as before, assumed to depend solely on the present day's weather state. Conceptually, the NHMM acts as an automatic classifier—it defines weather states that correspond to particular patterns of precipitation at the stations. This is in contrast to the weather state models described previously in which the weather states had to be explicitly defined a priori. In practice, we have found that the implicitly defined weather states of the NHMM also correspond to distinct patterns in the atmospheric data. Thus, the model serves to link particular patterns in the atmospheric data with particular rainfall patterns.

Hughes and Guttorp (1994) reported good results when this model was applied to 24 years of rainfall occurrence data from a network of four stations in Washington State. In particular, they found that the NHMM gave a better fit to the duration distributions of wet and dry runs than similar explicit weather-state models. In addition, no arbitrary modifications to the conditional independence assumption of the rainfall occurrence process were required to achieve this result. However, their model assumed that, conditional on the weather state, the precipitation occurrence processes at the stations were spatially independent. This assumption appeared to be satisfied for the four widely separated stations included in their example. However, one would not expect this conditional spatial independence assumption to be satisfied when a network of closely spaced stations is analyzed. In this paper we extend the methods of Hughes and Guttorp (1994) by proposing a model for multivariate rainfall occurrence data that preserves both the first- and second-order (conditional) moments. This model is imbedded in the NHMM framework to provide a way of connecting the synoptic atmospheric measures to local networks of rain gauge stations.

In the next section the model is formulated. In section 3 the method of parameter estimation is outlined. In section 4 a method of simulating sequences of multistation precipitation data from the model is described and these simulations are used to obtain Monte Carlo estimates of model statistics. In section 5 the model is applied to a network of 24 stations in the Puget Sound region of western Washington.

2. Model

Everything will be written in terms of probabilities; if some of the measures are absolutely continuous then densities can be substituted. Let $\mathbf{R}_t$ be the measurement of precipitation at time t, S_t the weather state at time t (unobserved), and $\mathbf{X}_t$ the measurement (or summary) of the atmospheric data at time t for $1 \leq t \leq T$. In the following, $\mathbf{R}_t$ will be assumed to be a multivariate binary measure corresponding to rain/no rain at a network of sites. The notation $\mathbf{X}_1^T$ will be used to indicate the sequence of atmospheric data from time 1 to T. Lowercase will be used to indicate the actual values of random variables [i.e., $P(\mathbf{R}_t = \mathbf{r})$]. All vectors are row vectors.

In its most general form, the NHMM is defined by the following assumptions:

$$P(\mathbf{R}_t | S_1^T, \mathbf{R}_1^{t-1}, \mathbf{X}_1^T) = P(\mathbf{R}_t | S_t), \qquad \text{(M1)}$$

$$P(S_t | S_1^{t-1}, \mathbf{X}_1^T) = P(S_t | S_{t-1}, \mathbf{X}_t), \qquad \text{(M2)}$$

and $P(S_1 | \mathbf{X}_1^T) = P(S_1 | \mathbf{X}_1)$.

The first assumption (M1) states that the rainfall process, $\mathbf{R}_t$, is conditionally independent given the weather state. In other words, all the temporal persistence in precipitation is captured by the persistence in the weather state described in (M2). Hughes (1993) discusses some extensions in which rainfall is assumed to be conditionally Markov.

Assumption (M2) states that, given the history of the weather state up to time $t-1$ and the entire sequence of the atmospheric data (past and future), the weather state at time t depends only on the previous weather state and the current atmospheric data. Note that the atmospheric data need not be, and typically will not be, raw atmospheric data. Rather, it is usually more convenient to use some sort of summary atmospheric measure such as a principal component (empirical orthogonal function) or an investigator-defined summary measure. To relate this model to the explicitly defined weather state models of Hay et al. (1991) and others, let $\mathbf{X}_t$ be the investigator's explicitly defined weather state and set $P(S_t | S_{t-1}, \mathbf{X}_t) = P(S_t | \mathbf{X}_t) = 1$ if $S_t = \mathbf{X}_t$ and 0 otherwise.

Specific NHMMs are defined by parameterizing $P(\mathbf{R}_t | S_t)$ and $P(S_t | S_{t-1}, \mathbf{X}_t)$. In Hughes and Guttorp (1994), the following parameterization for $P(\mathbf{R}_t | S_t)$ for n-station precipitation occurrence data was used. Let $\mathbf{R}_t = \{R_t^1, \cdots, R_t^n\}$ with observed value of $\mathbf{r}_t = \{r_t^1, \cdots, r_t^n\}$. Typically, $r_t^i = 1$ if rain occurs on day t at station i and 0 otherwise. Then Hughes and Guttorp (1994) define the independence model for $P(\mathbf{R}_t | S_t)$ as

$$P(\mathbf{R}_t = \mathbf{r}_t | S_t = s) = \prod_{i=1}^{n} p_{is}^{r_t^i} (1 - p_{is})^{1 - r_t^i}. \qquad (1)$$

This model assumes that there is some probability of rain, p_{is}, which varies from station to station and weather state to weather state. However, the occurrence process is assumed to act independently between stations conditional on the weather state (unconditionally, however, the stations will be correlated due to the influence of the common weather state that affects all

stations). Hence, the joint rainfall occurrence distribution has the simple form shown. If there are m weather states then there are nm parameters to estimate in this model for $P(\mathbf{R}_t = \mathbf{r}_t | S_t = s)$.

For a widely dispersed network, the independence model will usually be adequate since most or all of the correlation between stations will be induced by the common weather state. For a densely packed network, however, local topography and climatic features may induce correlations between stations that cannot be explained by the synoptic-scale weather state. For this situation we propose using (a generalization of) the autologistic model (Cressie 1991) for multivariate binary data:

$$P(\mathbf{R}_t = \mathbf{r} | S_t = s) \propto \exp\Big(\sum_{i=1}^{n} \alpha_{is} r^i + \sum_{j<i} \beta_{ijs} r^i r^j\Big), \quad (2)$$

where both α_{is} and β_{ijs} must be finite and $\beta_{iis} = 0$ to ensure model identifiability. In reporting results, the parameterization $p_{is} = \exp(\alpha_{is})[1 + \exp(\alpha_{is})]^{-1}$ will often be used for compatibility with the independence model (see below).

The parameters β_{ijs} measure the spatial dependence. When β_{ijs} is positive, stations i and j are positively correlated (within weather states). A negative value for β_{ijs} implies negative correlation between stations i and j (within weather states). When each β_{ijs} is 0, (2) reduces to the independence model with $\alpha_{is} = \log[p_{is}(1 - p_{is})^{-1}]$. Model (2) contains $mn(n-1)/2$ free parameters—a substantial number if n is at all large. However, the number of parameters can be reduced by modeling the β_{ijs} as a function of the distance and direction between stations i and j. Some possibilities are

$$1)\quad \beta_{ijs} = \begin{cases} \beta_s, & d_{ij} < h_s \\ 0, & d_{ij} > h_s \end{cases}$$

$$2)\quad \beta_{ijs} = \begin{cases} \beta_{ijs}, & d_{ij} < h_s \\ 0, & d_{ij} > h_s \end{cases}$$

$$3)\quad \beta_{ijs} = \begin{cases} \dfrac{\beta_s}{c_s + d_{ij}}, & i \neq j \\ 0, & i = j, \end{cases}$$

where d_{ij} is some measure of the distance only (for isotropic models), or the distance and direction (for anisotropic models) between stations i and j and h_s is a prespecified constant for each weather state. Use of the parameter c_s in the third example allows the correlation between arbitrarily close stations to be less than unity. Often c_s will be set to 0. However, if local topology tends to make the precipitation process highly spatially irregular, or if weather state s corresponds to scattered showers or local thunderstorms, then it might make sense to allow c_s to be nonzero. Also, a nonzero c_s can be used to account for measurement error.

Parameterizations for $P(S_t | S_{t-1}, \mathbf{X}_t)$ are discussed in detail in Hughes (1993) and Hughes and Guttorp (1994). The parameterization that we have found most satisfactory is motivated by Bayes' formula:

$$P(S_t = j | S_{t-1} = i, \mathbf{X}_t) \propto P(S_t = j | S_{t-1} = i) P(\mathbf{X}_t | S_{t-1} = i, S_t = j) \quad (3)$$

$$= \gamma_{ij} \exp\Big[-\frac{1}{2}(\mathbf{X}_t - \boldsymbol{\mu}_{ij}) \Sigma^{-1} (\mathbf{X}_t - \boldsymbol{\mu}_{ij})'\Big]. \quad (4)$$

The matrix $\gamma_{ij} = P(S_t = j | S_{t-1} = i)$ may be thought of as the baseline transition matrix of the weather-state process. The exponential term quantifies the effect of the atmospheric data on the baseline transition matrix. The mean vector of the atmospheric variables, $\boldsymbol{\mu}_{ij}$, is assumed to vary according to the current and past weather state, but the variance–covariance matrix Σ is assumed constant (in general, Σ could be defined separately for each weather state). To ensure identifiability of the parameters, the constraints $\sum_j \gamma_{ij} = 1$ and $\sum_j \boldsymbol{\mu}_{ij} = \boldsymbol{\mu}_i = 0$ are imposed.

3. Parameter estimation

For small networks (less than about 15 stations), or when the independence model for $P(R_t | S_t)$ is used, parameter estimation is accomplished by maximum likelihood. The likelihood can be written in matrix form as [see Hughes and Guttorp (1994) for a detailed derivation]

$$L = \boldsymbol{\delta}(\mathbf{x}_1)\mathbf{B}(\mathbf{r}_1)\mathbf{A}(\mathbf{x}_2)\mathbf{B}(\mathbf{r}_2)\cdots\mathbf{A}(\mathbf{x}_T)\mathbf{B}(\mathbf{r}_T)\mathbf{1}', \quad (5)$$

where $A_{ij}(\mathbf{x}) = P(S_t = j | S_{t-1} = i, \mathbf{X}_t = \mathbf{x})$, $\boldsymbol{\delta}(\mathbf{x})$ is the solution to $\mathbf{A}(\mathbf{x})\boldsymbol{\delta}'(\mathbf{x}) = \boldsymbol{\delta}'(\mathbf{x})$, and $\mathbf{B}(\mathbf{r})$ is an $m \times m$ diagonal matrix with the elements $\pi_{\mathbf{r}}(s) = P(\mathbf{R}_t = \mathbf{r} | S_t = s)$, $s = 1, \cdots, m$ along the diagonal. If one has several independent sequences of data (e.g., several years of January rainfall data), then the likelihoods for each year are multiplied together to form the overall likelihood. Missing values in the precipitation data may be incorporated by summing the likelihood over all possible values of the missing data. That is, let $\{\mathbf{R}_1^T\} = \{\mathbf{R}^{\text{obs}}, \mathbf{R}^{\text{miss}}\}$, where $\{\mathbf{R}^{\text{obs}}\}$ is the observed data and $\{\mathbf{R}^{\text{miss}}\}$ is the missing data. Then the likelihood may be expressed as

$$L = P(\mathbf{R}^{\text{obs}} | \mathbf{X}_1^T) = \sum_{\mathbf{R}^{\text{miss}}} P(\mathbf{R}^{\text{obs}}, \mathbf{R}^{\text{miss}} | \mathbf{X}_1^T).$$

Missing values for the atmospheric data are more difficult to handle and essentially require the specification of a model for the distribution of the missing data given the observed data [i.e., $P(\mathbf{X}^{\text{miss}} | \mathbf{X}^{\text{obs}})$]. When the amount of missing atmospheric data is small, we use a simple interpolation scheme to "fill in" the missing atmospheric data. If, however, a large proportion of the atmospheric data is missing, then a defensible model for $P(\mathbf{X}^{\text{miss}} | \mathbf{X}^{\text{obs}})$ will need to be developed before the NHMM can be fit.

When the number of rain stations in the network is large and the autologistic model [Eq. (2)] for $P(\mathbf{R}_t|S_t)$ is used, it is computationally impracticable to compute the likelihood directly. This is because computation of the normalizing constant in (2) requires summing over 2^n terms where n is the number of stations. This summation must be redone every time the parameters change during the maximization of the likelihood. Thus, an alternative method of parameter estimation is needed.

If the weather states were known, the method of maximum pseudolikelihood (Besag 1975) could be used in place of maximum likelihood for parameter estimation. The *pseudolikelihood* of the autologistic model is defined as

$$\prod_{i=1}^{n} P(R_t^i | R_t^{j\neq i}, S_t = s) = \prod_{i=1}^{n} \frac{\exp(\alpha_{is} r_t^i + \sum_{j\neq i} \beta_{ijs} r_t^i r_t^j)}{1 + \exp(\alpha_{is} + \sum_{j\neq i} \beta_{ijs} r_t^j)}. \quad (6)$$

The advantage of the pseudolikelihood is that each of the conditional probabilities in (6) is easily determined—there is no complex normalizing constant. The disadvantage is that the pseudolikelihood is less efficient than the true likelihood (unless, of course, all the β_{ijs} are equal to 0, in which case the pseudolikelihood is equal to the likelihood). Numerical studies (Besag 1977) have suggested that this efficiency loss is relatively small when the spatial autocorrelation is small (i.e., the β_{ijs} are small) but may be substantial when the spatial autocorrelation is large.

Unfortunately, the weather states are not known and direct substitution of the pseudolikelihood for $P(\mathbf{R}_t|S_t)$ in Eq. (5) yields abberrant results (Hughes 1993). Instead, the alternating estimation–restoration (ER) method of Besag (1986) was adapted to this problem. The idea is to start with some parameter values and "restore" (i.e., estimate) the underlying, unobserved weather-state sequence [using, e.g., the Viterbi algorithm (Forney 1973) outlined in the appendix] and then reestimate the parameters of $P(\mathbf{R}_t|S_t)$ as if the estimated weather state sequence is true. The procedure is then iterated. In the present application, the parameters obtained by first maximizing the likelihood (5) using the independence model for $P(\mathbf{R}_t|S_t)$ provide good starting values. The weather-state sequence is estimated using the Viterbi algorithm. The parameters of $P(\mathbf{R}_t|S_t)$ are then reestimated using the pseudolikelihood, the weather-state sequence is restored using the new parameters, and the procedure is iterated. This estimation method will be referred to as ER/MPLE. Note that, unlike Besag (1986), the parameters of $P(S_t|S_{t-1}, \mathbf{X}_t)$ are fixed at their starting values (which were obtained under the assumption of spatial independence). Hughes (1993) applied this approach to a small network in which both the MLE (maximum likelihood estimate) and ER/MPLE (maximum pseudolikelihood estimate) parameters could be estimated and found that the ER/MPLE approach gave values quite close to the MLE estimates.

4. Simulation and summary statistics

The fitted model determines certain summary statistics that provide a useful description of the rainfall field and may also be used to ascertain the goodness of fit of the NHMM. These include the marginal probability of rain at each station, the spatial and temporal cross correlations between and within stations, and the duration distribution of wet and dry days at each station. When the independence model for $P(\mathbf{R}_t|S_t)$ is used, these statistics may be expressed in a computationally simple form (see Hughes and Guttorp 1994). When the autologistic model for $P(\mathbf{R}_t|S_t)$ is used with a large number of stations, however, direct computation of these statistics is intractable. As an alternative, the model-based estimates of these statistics may be obtained by Monte Carlo simulation. The idea is to generate multiple, independent precipitation sequences from the fitted model (conditional on the observed atmospheric data) and compute the summary statistics for each sequence. By averaging over a large enough number of sequences one can obtain an arbitrarily precise estimate of the true values of the summary statistics for the fitted model. Comparison of the model-based statistics with the actual observed values of the statistics in the original data gives a measure of the goodness of fit of the model.

Generating a simulated realization of an NHMM is conceptually simple. The most straightforward approach is to generate the process in two steps: in the first step a realization of the (nonhomogeneous) Markov chain, S_1^T, is generated using the transition matrices $\mathbf{A}(\mathbf{x}_t)$, $t = 2 \cdots T$ and the initial probabilities $\delta(\mathbf{x}_1)$. Call this realization s_1^T. In the second step, $\mathbf{R}_t$ is randomly generated according to the probabilities $\pi_r(s_t)$, $t = 1 \cdots T$. A practical problem arises, however, if the support of $P(\mathbf{R}_t|S_t)$ is large. In that case, drawing a random number according to the probabilities $\pi_r(s_t)$ may not be easy. This occurs, for instance, when the autologistic model for $P(\mathbf{R}_t|S_t)$ is used with large n since the probability distribution of $\mathbf{R}_t$ has support on 2^n points. Selecting a random value from this distribution is not directly possible for large n (note that this problem does not arise if the independence model is used since each binary R_t^i can be generated separately). Instead, importance sampling (Rubin 1988) may be used to generate deviates from the probability distribution $P(\mathbf{R}_t|S_t = s)$. To use importance sampling, one must first identify a probability distribution from which random deviates can be readily generated and that is "near" (in a sense to be defined

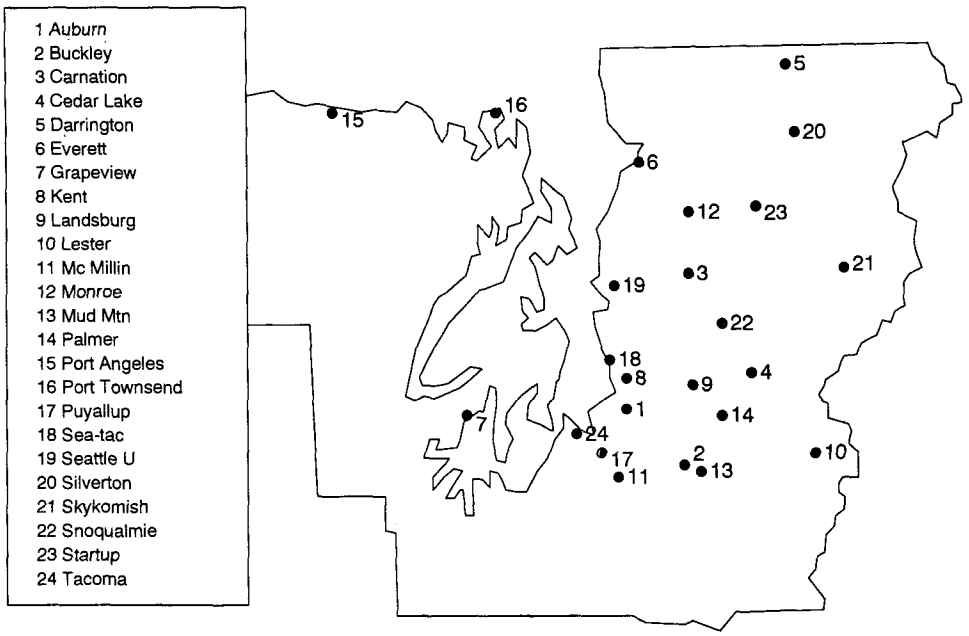

FIG. 1. Locations of rain gauge stations around the Puget Sound region of western Washington State.

below) to the distribution of interest, $P(\mathbf{R}_t | S_t = s)$. Let $f'_s(\mathbf{R})$ be such a distribution and let $f_s(\mathbf{R}) \propto P(\mathbf{R}_t | S_t = s)$. Then the idea behind importance sampling is to generate a large sample, $\mathbf{R}'_1, \ldots, \mathbf{R}'_M$ from $f'_s(\mathbf{R})$, form weights $w_i = f_s(\mathbf{R}_i)/f'_s(\mathbf{R}_i)$, and then draw a sample $\mathbf{R}_1, \ldots, \mathbf{R}_T$ from $\mathbf{R}'_1, \ldots, \mathbf{R}'_M$ (with replacement) with weights proportional to w_i. Rubin (1988) shows that, as $M/T \to \infty$, the distribution of $\mathbf{R}_t$ converges to $f_s(\mathbf{R}_t)$. Note that $f_s(\mathbf{R}_t)$ need only be known up to a constant of proportionality as is the case for the autologistic model.

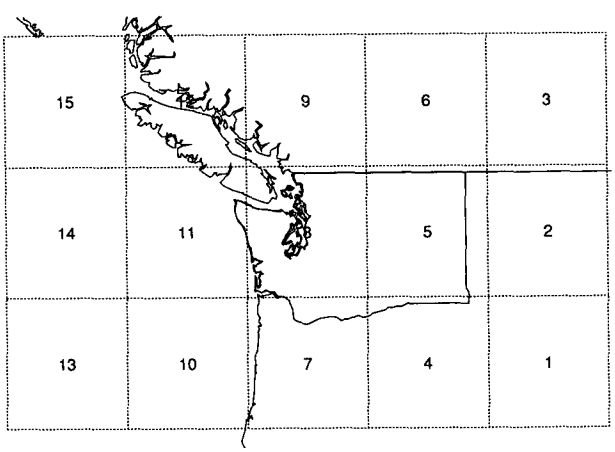

FIG. 2. Grid on which atmospheric data were collected. Numbers are positioned at the actual nodes to which the atmospheric measures were interpolated.

The efficacy of importance sampling depends strongly on the choice of $f'_s(\mathbf{R})$. The distribution $f'_s(\mathbf{R})$ must have the same support as $f_s(\mathbf{R})$, and $f_s(\mathbf{R})/f'_s(\mathbf{R})$ should not be "too large" or "too small" for any $\mathbf{R}$ in the support set. Fortunately, the structure of the autologistic model provides us with a natural candidate for $f'_s(\mathbf{R})$. For the autologistic model

$$f_s(\mathbf{R} = \mathbf{r}) \propto \exp\Big(\sum_{i=1}^{n} \alpha_{is} r^i + \sum_{i,j} \beta_{ijs} r^i r^j\Big).$$

Let $\beta_{ijs} = 0$ for all i and j to get

$$f'(\mathbf{R} = \mathbf{r}) = \prod_{i=1}^{n} \frac{\exp(\alpha_{is} r^i)}{1 + \exp(\alpha_{is})}.$$

Provided that the β_{ijs} are not too large, $f'_s(\mathbf{r})$ will be close to $f_s(\mathbf{r})$.

5. Example

In this example, rainfall occurrence (defined as 0.05 in. of precipitation or more) in a densely packed network of 24 stations around the Puget Sound area of Washington State is analyzed. Twenty-one years (1965–85) of daily winter data (the first 90 days of each year) are examined. Figure 1 shows the names and locations of each of the stations.

The atmospheric measures (described below) were based on sea level pressure and 500-mb geopotential height data collected by the U.S. National Meteorological Center (NMC) and made available on a CD-ROM by the Department of Atmospheric Sciences at

TABLE 1. Definitions of the atmospheric measures. Node numbers correspond to Fig. 2.

Measure	Interpretation	Definition
Sea level pressure		
P1	Mean	(node 5 + node 7 + node 8 + node 9 + node 11)/5
P2	North–south gradient	node 9 − node 7
P3	East–west gradient	node 5 − node 11
P4	Laplacian	node 8 − (node 5 + node 7 + node 9 + node 11)/4
500-mb geopotential height		
Z1	Mean	(node 5 + node 7 + node 8 + node 9 + node 11)/5
Z2	North–south gradient	node 9 − node 7
Z3	East–west gradient	node 5 − node 11
Z4	Laplacian	node 8 − (node 5 + node 7 + node 9 + node 11)/4
Z5	West Laplacian	node 11 − (node 8 + node 10 + node 12 + node 14)/4
Z6	East Laplacian	node 5 − (node 2 + node 4 + node 6 + node 8)/4

the University of Washington (Mass et al. 1987). The data were interpolated to a 3° × 4° latitude–longitude grid (see Fig. 2) from the original NMC octagonal grid using software developed at the National Center for Atmospheric Research. The percentage of missing data varies from 0% to 56% for the rain stations (median 3%) and is 2% and 3% for sea level pressure and 500-mb geopotential height, respectively.

Four measures were derived from the sea level pressure field and six measures were derived from the 500-mb geopotential height field. These measures are defined in Table 1 and include the mean field, the north–south gradient, the east–west gradient, and the Laplacian for both fields, and the east Laplacian and west Laplacian for the 500-mb geopotential height field. The gradient fields provide a measure of wind strength (the greater the gradient, the stronger the wind), while the Laplacians provide a measure of the convergence/divergence. Overall, then, there are 10 candidate atmospheric measures for inclusion in the NHMM. Each atmospheric measure was centered prior to analysis.

Table 2 shows the correlations among the atmospheric variables and Table 3 shows the correlations between the atmospheric variables and the rain occurrence process at each station. Mean pressure (P1) and north–south gradient (P2) seem to be most correlated with rainfall among the sea level pressure variables, whereas Z2 (north–south gradient) and Z3 (east–west gradient) show the greatest correlation with rainfall among the 500-mb geopotential height variables. Of these four measures, only P1 and Z3 and P2 and Z2 exhibit significant correlations with each other (−0.59 and 0.60, respectively). A reasonable strategy for choosing variables to enter in the NHMM is to choose measures that are most correlated with rainfall and least correlated with other measures that have already been entered in the model. On this basis, P1, Z2, P2, Z3 will be entered in that order.

Fitting the NHMM to a network of 24 stations is computationally demanding. Therefore, several strategies were adopted to reduce the number of models examined. First, models were fit hierarchically. That is, the model order (the number of weather states) was determined by fitting models with no atmospheric variables [the standard hidden Markov model (HMM) discussed in Zucchini and Guttorp (1991)]. Then the atmospheric measures were added in the sequence described above to determine the "best" [under the assumption of the independence model for $P(\mathbf{R}_t|S_t)$] model. Finally, an autologistic version of this final model was fit. Since the models are nested at each step, this approach provides good starting values for the optimization routine at each stage. The Bayes information criterion (BIC) (Katz 1981) (a penalized likelihood

TABLE 2. Correlation matrix of atmospheric variables.

	P1	P2	P3	P4	Z1	Z2	Z3	Z4	Z5	Z6
P1	1.00	0.23	−0.12	0.26	0.72	0.12	−0.59	0.31	0.51	0.13
P2	0.23	1.00	0.28	−0.31	0.10	0.60	−0.20	0.06	0.10	−0.04
P3	−0.12	0.28	1.00	−0.09	0.11	0.17	0.48	0.26	0.00	0.33
P4	0.26	−0.31	−0.09	1.00	0.28	−0.10	−0.04	0.22	0.27	0.12
Z1	0.72	0.10	0.11	0.28	1.00	0.09	−0.23	0.57	0.65	0.49
Z2	0.12	0.60	0.17	−0.10	0.09	1.00	−0.06	−0.02	−0.01	−0.06
Z3	−0.59	−0.20	0.48	−0.04	−0.23	−0.06	1.00	−0.04	−0.46	0.32
Z4	0.31	0.06	0.26	0.22	0.57	−0.02	−0.04	1.00	0.64	0.65
Z5	0.51	0.10	0.00	0.27	0.65	−0.01	−0.46	0.64	1.00	0.33
Z6	0.13	−0.04	0.33	0.12	0.49	−0.06	0.32	0.65	0.33	1.00

TABLE 3. Pearson correlation between each atmospheric measure and the 24 rain gauge stations. Rain stations are listed in the order given in Fig. 1.

	P1	P2	P3	P4	Z1	Z2	Z3	Z4	Z5	Z6
R1	−0.28	−0.25	0.27	0.00	−0.12	−0.33	0.29	0.08	−0.02	0.15
R2	−0.32	−0.34	0.17	−0.01	−0.14	−0.35	0.31	0.04	−0.09	0.14
R3	−0.30	−0.30	0.19	0.04	−0.16	−0.30	0.26	0.05	−0.06	0.12
R4	−0.36	−0.41	0.08	0.04	−0.26	−0.38	0.28	−0.08	−0.18	0.04
R5	−0.35	−0.49	0.11	0.09	−0.22	−0.40	0.32	−0.03	−0.13	0.08
R6	−0.37	−0.37	0.08	−0.06	−0.21	−0.35	0.33	−0.11	−0.21	0.04
R7	−0.43	−0.38	0.22	0.02	−0.28	−0.33	0.40	−0.03	−0.17	0.09
R8	−0.34	−0.28	0.27	−0.01	−0.16	−0.31	0.35	0.05	−0.11	0.16
R9	−0.27	−0.32	0.20	0.06	−0.14	−0.34	0.25	0.05	−0.05	0.12
R10	−0.26	−0.46	0.02	0.04	−0.13	−0.47	0.17	−0.00	−0.04	0.10
R11	−0.42	−0.38	0.16	−0.07	−0.23	−0.36	0.40	−0.10	−0.24	0.09
R12	−0.34	−0.35	0.20	0.00	−0.17	−0.33	0.30	0.03	−0.10	0.11
R13	−0.27	−0.31	0.17	0.03	−0.12	−0.34	0.24	0.02	−0.07	0.11
R14	−0.34	−0.39	0.10	0.04	−0.22	−0.35	0.28	−0.05	−0.14	0.02
R15	−0.31	−0.39	0.17	0.00	−0.18	−0.38	0.31	−0.01	−0.13	0.02
R16	−0.32	−0.35	0.08	−0.00	−0.19	−0.34	0.26	−0.05	−0.14	0.02
R17	−0.36	−0.30	0.26	−0.02	−0.19	−0.34	0.35	0.02	−0.14	0.14
R18	−0.28	−0.25	0.25	0.05	−0.14	−0.29	0.29	0.04	−0.06	0.12
R19	−0.38	−0.31	0.25	−0.06	−0.19	−0.27	0.42	−0.04	−0.21	0.01
R20	−0.38	−0.48	0.12	0.08	−0.21	−0.36	0.35	−0.03	−0.14	0.10
R21	−0.34	−0.50	0.06	0.12	−0.16	−0.41	0.28	−0.04	−0.12	0.09
R22	−0.28	−0.34	0.17	0.07	−0.14	−0.33	0.26	0.04	−0.06	0.12
R23	−0.36	−0.50	0.01	0.02	−0.22	−0.41	0.29	−0.10	−0.19	0.04
R24	−0.31	−0.28	0.28	0.00	−0.13	−0.36	0.35	0.08	−0.08	0.16

criterion equal to $2l - v \log n$ where l is the log likelihood, v is the number of model parameters, and n is the total sample size) was used to compare models.

Table 4 gives the results of this exercise. Comparing the HMMs only, the BIC criterion suggests that a model with five or six hidden states is appropriate. However, when atmospheric measures are added to the six-state model, the BIC actually increases. That is, the best six-state model is the HMM, not an NHMM. If one's only need were to generate simulations of the observed process, then this six-state HMM would probably be the model of choice. In this study, however, we are interested in the relationships between the atmospheric fields and rainfall so a five-state NHMM will be chosen, even though it may not fit as well as the six-state HMM. Among the five-state models, the BIC declines when P1 and Z2 are added, and then increases when P2 and Z3 are added. Thus, the five-state NHMM with atmospheric variables P1 and Z2 is chosen for further investigation. This model is denoted by S5P1Z2 in the following.

To investigate the need for a spatial model for $P(\mathbf{R}_t|S_t)$ [all of the fits given so far have been based on the independence model for $P(\mathbf{R}_t|S_t)$], each day was classified into a state using the Viterbi algorithm (see appendix) based on the S5P1Z2 model. The correlation in the precipitation occurrence process was then calculated for all pairs of stations within each state (i.e., for all the days in each state) and plotted against distance. The results are shown in Fig. 3. This figure suggests that an additional parameter to model the spatial correlation would be useful for several of the states. The figure also suggests that the within-state spatial correlation falls off quickly with distance so that the parameterization $\beta_{ijs} = \beta_s / d_{ij}$, where β_{ijs} indexes the spatial correlation in the autologistic model [Eq. (2)] and d_{ij} is the distance between stations i and j, is a reasonable model. States 1 and 4, which show little or no relationship between spatial correlation and distance, are associated with a high probability of rain at all stations and a high probability of dry conditions at all stations, respectively. States 2, 3, and 5, which show more of a relationship between spatial correlation and distance, are associated with particular regional patterns of rainfall that will be described below. The one outlier

TABLE 4. Comparison of models. Model notation is as follows: the number following S is the number of hidden states; the numbers following P are the pressure variables used, and the numbers following Z are the geopotential height variables used. Here *df* indicates the number of model parameters and log*L* is the log likelihood. BIC is the Bayes information criterion described in the text.

Model	*df*	log*L*	BIC
S2	50	17 453	35 283
S3	78	15 371	31 330
S4	108	14 905	30 625
S5	140	14 527	30 110
S5P1	160	14 437	30 081
S5P1Z2	180	14 350	30 058
S5P12Z2	200	14 298	30 104
S5P12Z23	220	14 250	30 160
S6	174	14 295	29 903
S6P1	204	14 197	29 933
S6P1Z2	234	14 109	29 983

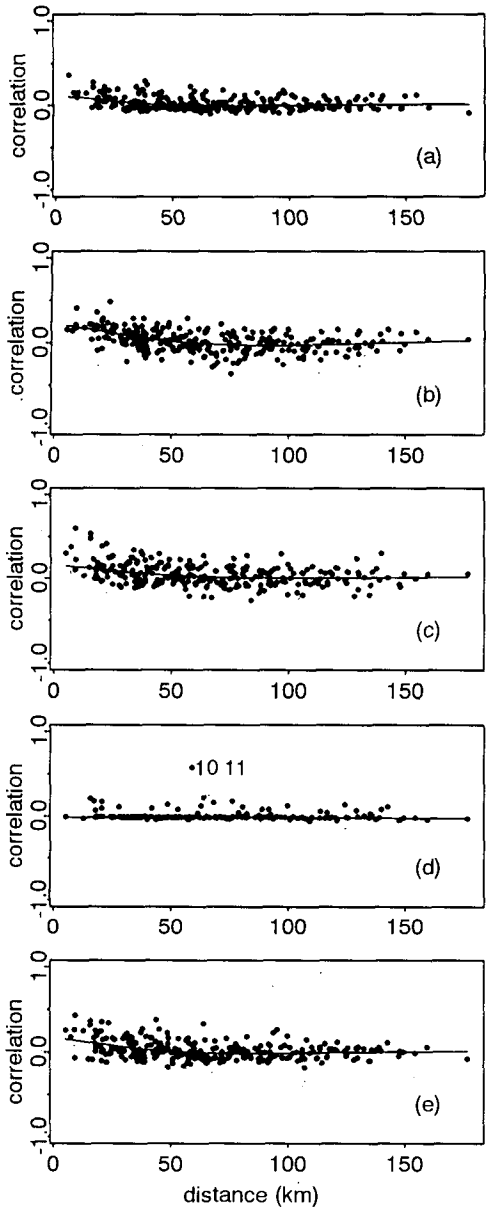

FIG. 3. Plot of pairwise correlations of the rainfall occurrence process versus distance within weather states 1–5 [panels (a)–(e), respectively]. Weather state was estimated on the basis of model S5P1Z2. The line represents a smooth through the points.

seen in state 4 represents the spatial correlation between stations 10 and 11 (Lester and McMillian). Due to missing values, this correlation is based on only 733 observations out of a possible 1890.

The ER/MPLE strategy described in section 3 was used to estimate the parameters of an autologistic model for $P(\mathbf{R}_t|S_t)$. Unfortunately, one cannot use the BIC criterion to compare this model to the independence model since the (true) likelihood is computationally intractable for the autologistic model. Instead the improvement in fit due to the autologistic model is evaluated graphically. At each iteration of the ER/MPLE procedure the predicted marginal probability of rain at each station and the predicted correlation between stations based on the current model were compared to the observed values of these statistics. Figure 4 gives the results for the independence model and the first three ER/MPLE iterations. The fit has clearly been improved after just one or two iterations. In particular, the high observed correlations, which were underestimated by the independence model (Fig. 4a), now appear to be modeled correctly. One problem is seen, however. As the ER/MPLE procedure is iterated, it appears as though the predicted correlations are being biased upward. There are two possible explanations for this. First, Qian and Titterington (1989) indicate that parameters obtained by Besag's (1986) estimation–restoration method are biased and that criticism may carry over to the ER/MPLE used here. Second, as the fitted model moves further and further from the independence model it becomes harder and harder to generate a valid sample using importance sampling (see section 4). Since the predicted marginal probabilities and correlations are estimated by generating repeated samples from the fitted model using importance sampling, the apparent bias in the predicted correlations may be an artifact of the way the predicted statistics were computed. In any event, the final model is taken to be the set of parameters from the second iteration. These seem to give the best trade-off between reproducing the observed sample statistics and providing a valid importance sample. This model will be designated S5P1Z2s. Note that β_4 was fixed at 0 (the independence model value) based on the results seen in Fig. 3.

The observed and predicted marginal probabilities of rain at each station for the model S5P1Z2s are shown in Fig. 5. The model does a very good job of reproducing the observed marginal rainfall probabilities. In addition, unlike the independence model, the fitted spatial model does a good job of reproducing the observed spatial correlation at both the high and low end of the spectrum (cf. Figs. 4a and 4c).

Figure 6 shows the observed and predicted survivor curves of the duration distribution of rainfall at each station (the survivor function evaluated at duration t is equal to $1 - F(t)$, where $F(t)$ is the cumulative distribution function; thus, $S(t)$ is the probability that a "storm" will last more than t days). In general, the model does a good job of reproducing the observed survivor curves. The worst fit is at station 7—Grapeview—which is relatively isolated. One speculative explanation is that there is an additional weather state that strongly affects Grapeview but that was not detected due to the lack of stations in this area. Alternatively, the poor fit at this station may result from the inability to draw on information from other nearby stations. A second situation where the

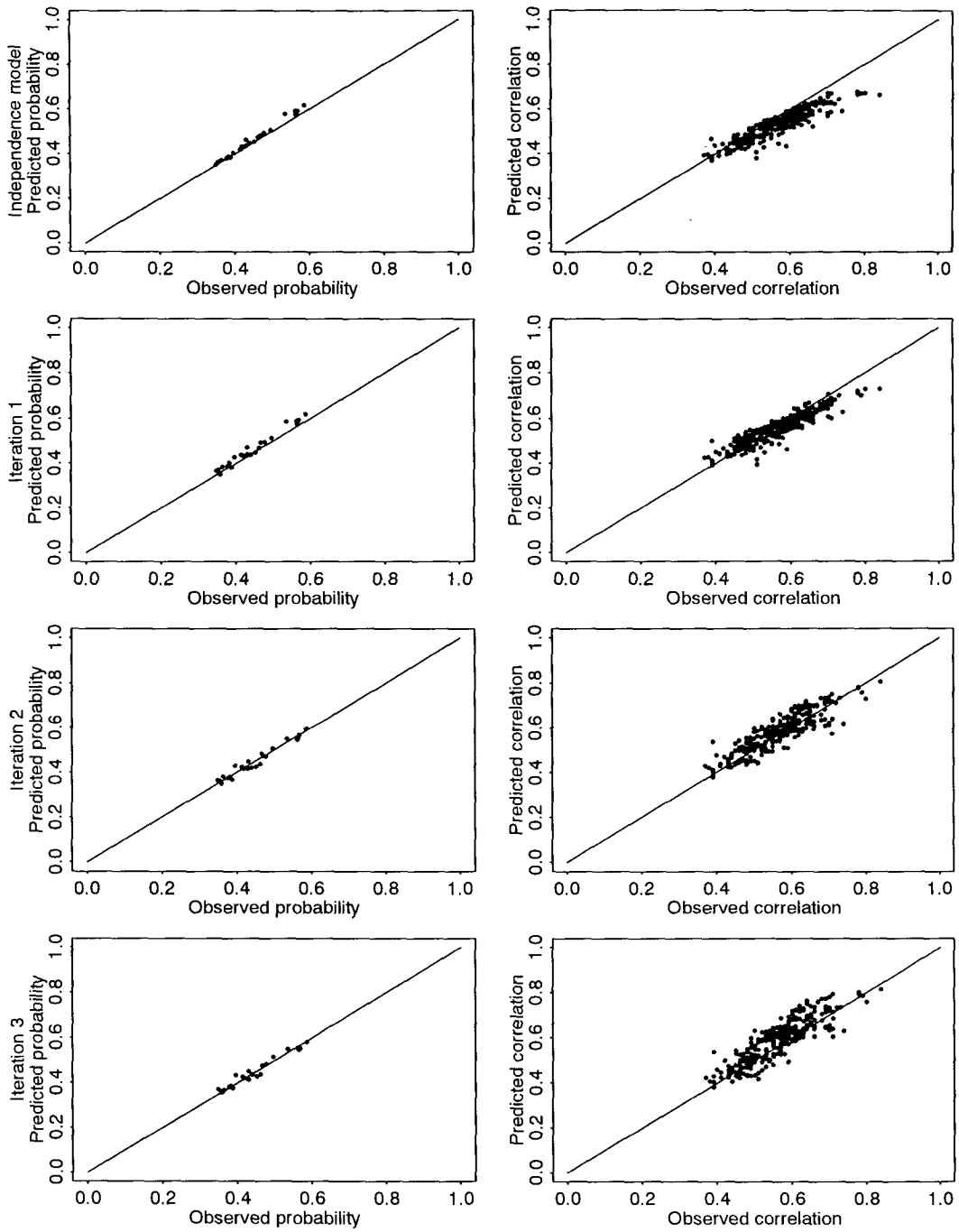

FIG. 4. Agreement between observed and predicted sample statistics for the independence model (S5P1Z2) and several ER/MPLE iterations of the spatial (S5P1Z2s) model.

fit is not as good as might be hoped for occurs when the observed survivor curve is concave (e.g., station 21). The predicted survivor curve fails to reproduce this behavior. Interestingly, this behavior seems to occur primarily in the inland stations (i.e., 10, 20, 21). In all cases where the fit is poor the model produces a distribution that is too light tailed—that is, not enough long storms. Plots of the interarrival times (not shown)—the times between storms—give similar results. The fit is generally good but when the model does perform poorly it tends to produce a distribution that is too light tailed. This implies that the overall persistence—of both wet and dry periods—is not being completely captured.

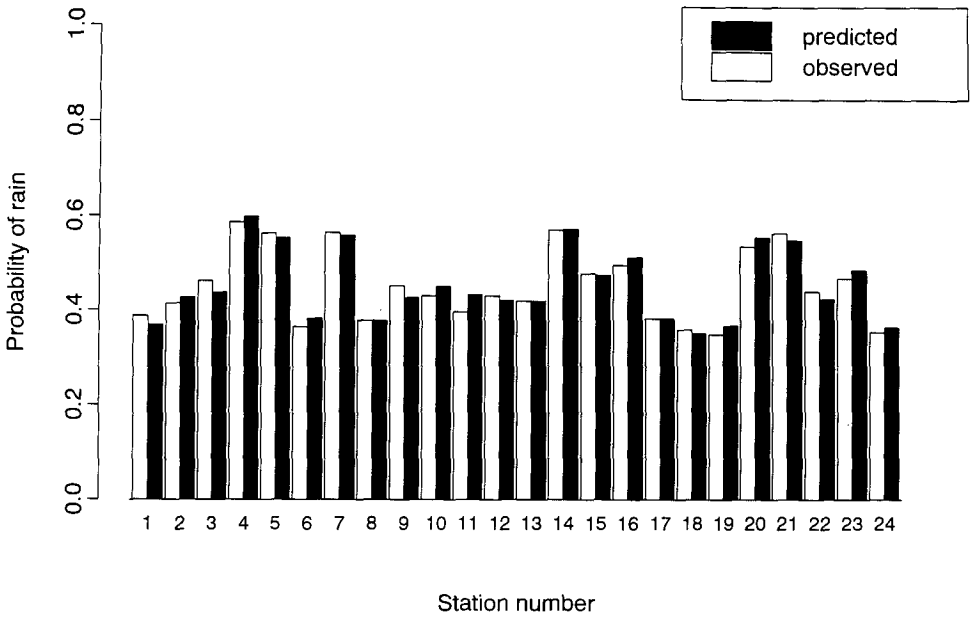

FIG. 5. Observed versus predicated probability of rain at each station. Refer to Fig. 1 for station locations.

There are five weather states in the final model. When the state for each day is estimated using the Viterbi algorithm, the relative frequencies of the estimated states are 26%, 4%, 19%, 28%, and 23%. Figure 7 shows the mean sea level pressure field, the mean 500-mb geopotential height, and the distribution of rainfall for each weather state. As before, the states are estimated from the final model and the plots represent an average of all days in a given state. State 1 corresponds to southwesterly flow over the Puget Sound region both at the surface and aloft. This leads to a high probability of rain at all stations. State 2 corresponds to a westerly flow at the surface but southwest flow at 500 mb. This corresponds to a high probability of rain to the north, relatively low probability of rain in the central Puget Sound region, and moderate to high probability to the south. State 3 is characterized by southwesterly flow at sea level and westerly flow at 500 mb. The winds at sea level are weaker than in state 1. The result is that the stations bordering Puget Sound are in a rain shadow, whereas the more eastern stations (i.e., Silverton, Skykomish, and Lester) have a high probability of rain. State 4 is characterized by a high pressure system over the Puget Sound region. Consequently, the probability of rain is low at all stations. Finally, state 5 is similar to state 2, although the 500-mb flow is more westerly. The rainfall pattern is similar to that seen in state 2. These results indicate that the model is finding sensible and interpretable weather states.

6. Discussion

A model has been described that allows one to link synoptic atmospheric information to rainfall in a densely packed local network. Previous models were either limited to regional networks, where the majority of the spatial correlation between stations is induced by synoptic weather patterns, or evoked ad hoc corrections for the spatial correlation. The distinguishing feature of this model is the concept of an unobserved weather state that serves to discretize the infinite number of atmospheric data patterns into a few canonical classes. All existing weather-type models of which the authors are aware represent particular parameterizations of the NHMM.

There are several ways that this model might be used. Once a model has been fit to historical data, simulations of multistation rainfall may be generated conditional on a given sequence of synoptic atmospheric data. In particular, the effect of various altered climate hypotheses on local rainfall may be tested by comparing rainfall simulations from a GCM control run and the corresponding GCM altered climate run. Such an exercise is, of course, highly speculative since one cannot know if the relationships between the synoptic atmospheric information and rainfall that were observed in the historical data will continue to hold under the altered climate. The problem is somewhat akin to predicting from a regression equation outside the range of one's data. A more testable use of the model might be for short-term weather prediction. Based on a fitted model and a short-term prediction of synoptic atmospheric conditions, one could make predictions about short-term precipitation probabilities. Thus far, we have not attempted to use the model for this purpose. Finally, simulations from

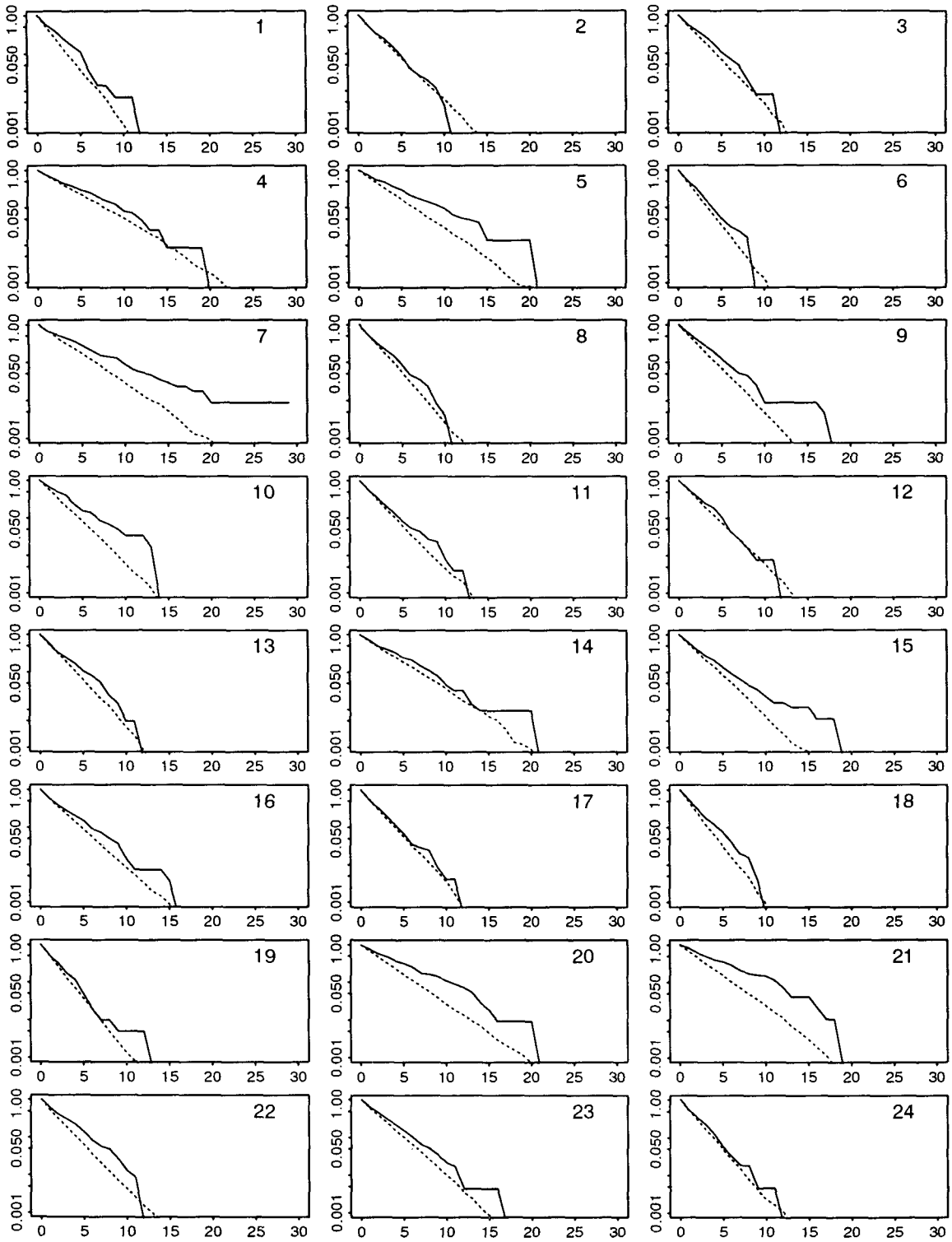

FIG. 6. Observed versus predicted duration distribution of rain at each station. Solid line is observed duration distribution; dotted line is predicted. The horizontal axis is time in days and the vertical axis is (on a log scale) the probability that a wet spell lasts for more than t days. Station number is in the upper-right corner of each plot.

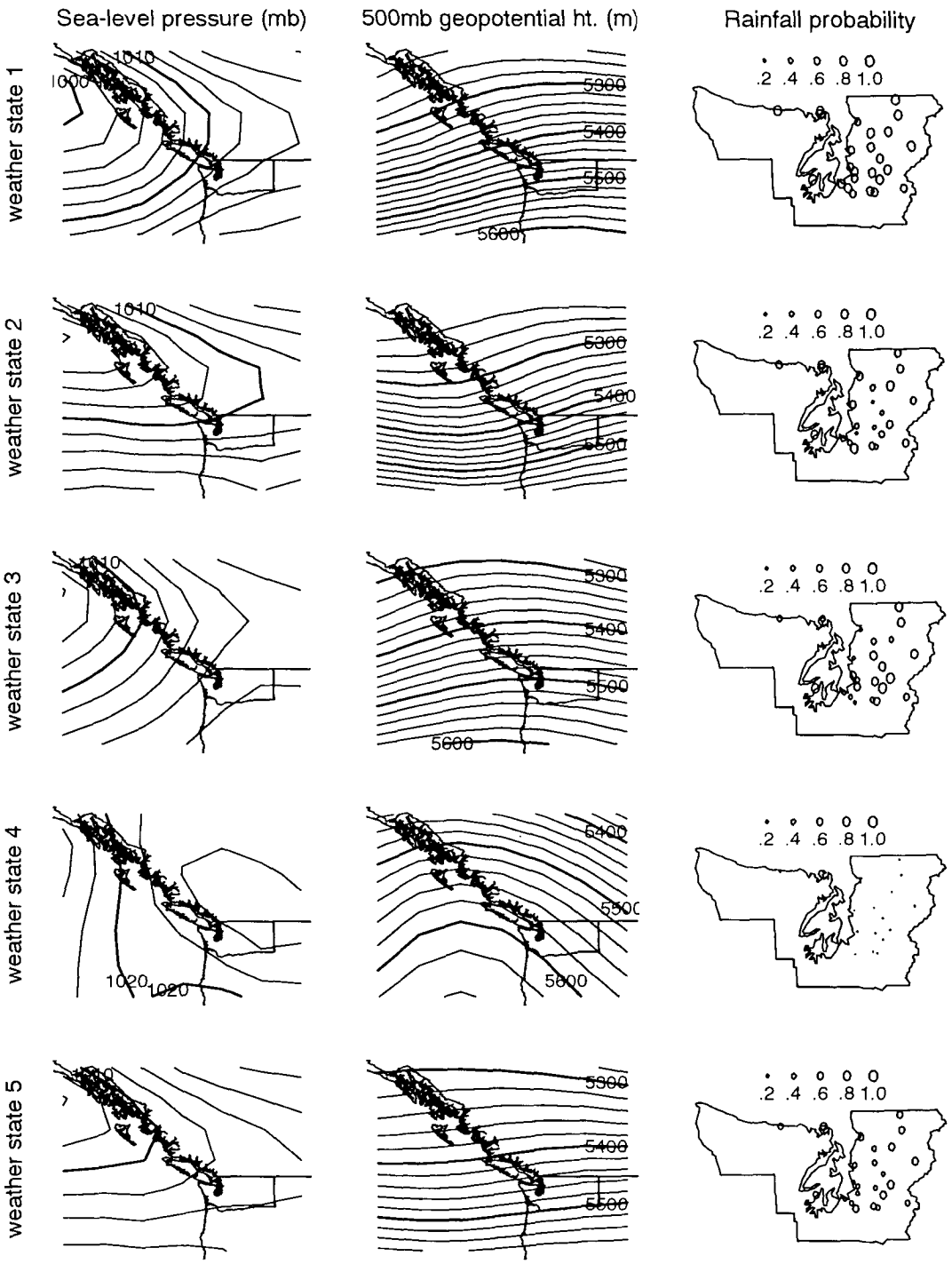

FIG. 7. Mean sea level pressure and 500-mb geopotential height averaged over all days classified in each weather state. Contour interval for pressure is 2 mb; contour interval for geopotential height is 20 m.

the model run under current climate conditions might be used as input into watershed models, crop models, etc.

All of the discussion to this point has dealt with precipitation occurrence data rather than amounts. However, all of the key results continue to hold if $\mathbf{R}_t$ is a continuous random variable. Then $P(\mathbf{R}_t | S_t)$ is a density—for instance, a mixture of a point mass (at 0) and a gamma distribution (Stern and Coe 1984) or a transformed multivariate normal distribution (Bar-

dossy and Plate 1992). The latter would probably be preferable for multistation data.

The NHMM as presented above applies to meteorologically homogeneous regions (on the synoptic scale). We are currently investigating the possibility of extending this approach to continental or subcontinental scales. One possible approach to this problem would be to identify K distinct precipitation and meteorological regions and then allow $\mathbf{S}_t = (S_t^1 \cdots S_t^K)$ to be a K-variate random process where the kth element corresponds to the weather state in the kth region. Then $P(\mathbf{S}_t|\mathbf{S}_{t-1}, \mathbf{X}_t)$ would be modeled as a multivariate Markov process.

Acknowledgments. The first author had partial support from a fellowship from the IBM Thomas Watson Research Center. The second author had partial support from National Science Foundation Grant DMS-9115756. Partial computing support was provided by the San Diego Supercomputing Center. The authors are grateful to Mike Wallace for helpful discussions.

APPENDIX

Estimating the Weather States

The Viterbi algorithm estimates the most likely weather-state sequence given a set of atmospheric and rainfall data and a fitted NHMM. Estimating the most likely sequence of weather states is equivalent to finding the values s_1^T that maximize the probability

$$P(S_1^T = s_1^T | \mathbf{R}_1^T = \mathbf{r}_1^T, \mathbf{X}_1^T, \theta), \qquad \text{(A1)}$$

where θ are the model parameters. The solution to this problem is well known for the standard hidden Markov model (see, e.g., Forney 1973). We have updated this method for the NHMM. Here is an explicit description of the Viterbi algorithm (using the notation introduced in sections 2 and 3) for reconstructing the state sequence $S_1, \ldots, S_T$ from the observed sequences $\mathbf{r}_1, \ldots, \mathbf{r}_T$ and $\mathbf{x}_1, \ldots, \mathbf{x}_T$.

1) Create the integer storage array $\mathbf{Q}$ with dimensions $m \times T$ where m is the number of states of S.

2) Create temporary real-valued storage arrays $\mathbf{f}$ and $\mathbf{g}$ of dimension $1 \times m$.

3) Let $f(s) = B_{ss}(\mathbf{r}_T)$ $s = 1, \cdots, m$.

4) For each $t = T - 1, \cdots, 1$ do the following:

   (a) For each $s = 1, \cdots, m$ set $j = \text{argmax}_i A_{si}(\mathbf{x}_{t+1}) f(i)$, $Q_{s,t+1} = j$, and $g(s) = \pi_{\mathbf{r}_t}(s) A_{sj}(\mathbf{x}_{t+1}) f(j)$.

   (b) Set $\mathbf{f} = \mathbf{g}$.

5) Set $j = \text{argmax}_i \delta_i(\mathbf{x}_1) f_1(i)$ and $Q_{j1} = j$.

6) The solution is $S_t = Q_{jt}$ $t = 1, \cdots, T$.

REFERENCES

Bardossy, A., and E. J. Plate, 1992: Space–time models for daily rainfall using atmospheric circulation patterns. *Water Resour. Res.*, **28,** 1247–1259.

Besag, J., 1975: Statistical analysis of non-lattice data. *Statistician,* **24,** 179–195.

——, 1977: Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika,* **64,** 616–618.

——, 1986: On the statistical analysis of dirty pictures. *J. Roy. Stat. Soc. B,* **48,** 259–302.

Cressie, N., 1991: *Statistical Analysis of Spatial Data.* John Wiley and Sons, 900 pp.

Forney, G. O., 1973: The Viterbi algorithm. *Proc. IEEE,* **61,** 268–278.

Hay, L., G. J. McCabe, D. M. Wolock, and M. A. Ayers, 1991: Simulation of precipitation by weather type analysis. *Water Resour. Res.*, **27,** 493–501.

Hughes, J. P., 1993: A class of stochastic models for relating synoptic atmospheric patterns to local hydrologic phenomena. Ph.D. dissertation, University of Washington, Seattle, 139 pp.

——, and P. Guttorp, 1994: A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour. Res.*, **30,** 1535–1546.

——, D. P. Lettenmaier, P. Guttorp, 1993: A stochastic approach for assessing the effects of changes in regional circulation patterns on local precipitation. *Water Resour. Res.*, **29,** 3303–3315.

Katz, R. W., 1981: On some criteria for estimating the order of a Markov chain. *Technometrics,* **23,** 243–249.

Mass, C. F., H. J. Edman, H. J. Friedman, N. R. Cheney, E. E. Recker, 1987: The use of compact disks for storage of large meteorological and oceanographic data sets. *Bull. Amer. Meteor. Soc.*, **68,** 1556–1558.

Qian, W., and D. M. Titterington, 1989: On the use of Gibbs Markov chain models in the analysis of images based on second-order pairwise interactive distributions. *J. Appl. Stat.*, **16,** 267–281.

Rabiner, L. R., and B. H. Juang, 1986: An introduction to hidden Markov models. *IEEE Acoust. Speech Signal Process. Mag.*, **3,** 4–16.

Rind, D., R. Goldberg, and R. Ruedy, 1989: Change in climatic variability in the 21st century. *Clim. Change,* **14,** 5–37.

Rubin, D. B., 1988: Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics 3,* J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Eds., Oxford University Press.

Stern, R. D., and R. Coe, 1984: A model fitting analysis of daily rainfall data. *J. Roy. Stat. Soc. A,* **147,** 1–34.

Wilson, L. L., D. P. Lettenmaier, and E. Skyllingstad, 1992: A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation. *J. Geophys. Res.*, **97,** 2791–2809.

Zorita, E., J. P. Hughes, H. von Storch, and D. P. Lettenmaier, 1995: Stochastic characterization of regional circulation patterns of climate model diagnosis and estimation of local precipitation. *J. Climate,* **8,** in press.

Zucchini, W., and P. Guttorp, 1991: A hidden Markov model for space–time precipitation. *Water Resour. Res.*, **27,** 1917–1923.