Vol. 15: 1–12, 2000 | CLIMATE RESEARCH Clim Res | Published May 15

A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts

Enrica Bellone[1,*], James P. Hughes[2], Peter Guttorp[1]

[1]Department of Statistics 354322 and [2]Department of Biostatistics 357232, University of Washington, Seattle, Washington 98195, USA

ABSTRACT: Nonhomogeneous hidden Markov models (NHMMs) provide a relatively simple framework for simulating precipitation at multiple rain gauge stations conditional on synoptic atmospheric patterns. Building on existing NHMMs for precipitation occurrences, we propose an extension to include precipitation amounts. The model we describe assumes the existence of unobserved (or hidden) weather patterns, the weather states, which follow a Markov chain. The weather states depend on observable synoptic information and therefore serve as a link between the synoptic-scale atmospheric patterns and the local-scale precipitation. The presence of the hidden states simplifies the spatio-temporal structure of the precipitation process. We assume the temporal dependence of precipitation is completely accounted for by the Markov evolution of the weather state. The spatial dependence of precipitation can also be partially or completely accounted for by the existence of a common weather state. In the proposed model, occurrences are assumed to be conditionally spatially independent given the current weather state and, conditional on occurrences, precipitation amounts are modeled independently at each rain gauge as gamma deviates with gauge-specific parameters. We apply these methods to model precipitation at a network of 24 rain gauge stations in Washington state over the course of 17 winters. The first 12 yr are used for model fitting purposes, while the last 5 serve to evaluate the model performance. The analysis of the model results for the reserved years suggests that the characteristics of the data are captured fairly well and points to possible directions for future improvements.

KEY WORDS: Hidden Markov model · Precipitation amounts model · Downscaling

1. INTRODUCTION

Stochastic models for precipitation have several important applications. For example, simulations from these models enter as input into flooding, runoff and crop growth models. Historically, rainfall modeling has followed 2 main themes. Some models were constructed to incorporate physical principles (e.g. Hobbs & Locatelli 1978), while others gave a more statistical description of the data. Along the lines of the former approach, point process models were developed by Le Cam (1961), Waymire et al. (1984) and Goodall & Phelan (1991). In the context of statistical descriptions of rainfall data, Gabriel & Neumann (1962) modeled precipitation occurrences as a first-order Markov chain. Their approach has been extended to allow seasonal differences (Stern & Coe 1984, Woolhiser 1992) by using time-varying parameters. Chain-dependent models have been developed where precipitation occurrences are assumed to follow a first-order Markov chain while precipitation amounts are taken to be conditionally independent given the occurrence process, with the amount distribution at each time point depending on occurrence at the present and preceding time points (Katz 1977, Katz & Parlange 1996).

Recently, the idea of relating daily precipitation to synoptic atmospheric patterns has led to the development of weather-state models. One motivation for including atmospheric variables in the model is the desire to assess the regional and local effects of global climate changes. General circulation models (GCMs), which typically operate on grids on the order of 3° latitude × 3° longitude, can capture large-scale atmospheric patterns and determine the effect of changes in the atmosphere on those patterns. However, GCMs are not as adequate for reproducing local and regional

*E-mail: bellone@stat.washington.edu

phenomena, such as rainfall (Giorgi & Mearns 1991). Thus, there is a need for models that can downscale the GCM predictions of global climate to local precipitation patterns. Stochastic models for rainfall that do not include synoptic atmospheric information cannot be used for this purpose, since they can only produce simulations under the current climate regime. In weather-state models, synoptic atmospheric patterns are the basis for classifying each day into a weather state and precipitation is then modeled within each state via multivariate distributions. Different versions of these models have been proposed by, for example, Hay et al. (1991), Bardossy & Plate (1992), Hughes et al. (1993) and Bartholy et al. (1995).

Our goal is to obtain a model that allows simulation of precipitation amounts, conditional on the value of some synoptic atmospheric variables. We base our approach on nonhomogeneous hidden Markov models (NHMMs), a class of models introduced by Hughes & Guttorp (1994). NHMMs extend the hidden Markov models (HMMs) used by Zucchini & Guttorp (1991) by incorporating synoptic atmospheric information. NHMMs assume the existence of weather states, but they differ from the weather-state models mentioned above in the way the states are defined. In weather-state models, each day is classified *a priori* into a state, according to synoptic patterns. Precipitation does not affect the state definition. In contrast, in NHMMs the states are identified as precipitation patterns that result from the model fitting procedure, while the role of synoptic atmospheric information is to influence the state transitions. In Hughes et al. (1994, 1999) NHMMs are used to model precipitation occurrences. Here we extend this approach to precipitation amounts.

In Section 2 we describe the assumptions that define an NHMM, the parameterization we use and the methods we apply to obtain estimates of the model parameters. Section 2 also explains our approach to the problem of determining the model order and to the treatment of the atmospheric variables. Section 3 describes an application of our methods to precipitation amounts at a network of gauges in Washington State. In section 4, we conclude with a discussion.

2. METHODS

2.1. Model assumptions

The NHMM assumes the existence of a *hidden* or unobservable stochastic process, which can take on a discrete number of states. In the context of precipitation modeling, we interpret this process as the 'state of the weather.' The adjective 'nonhomogeneous' derives from the assumption that the state of the weather at time t depends not only on the state of the weather at the previous time point, but also on the current value of some atmospheric variables. Thus, the state transition matrix varies in time with the atmospheric quantities. The assumptions for the hidden process can be summarized as:

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

where S_t is the weather state at time t and $\mathbf{X}_t$ is a vector of atmospheric variables at time t, $1 \le t \le T$. The notation $\mathbf{X}_1^T$ indicates all the values of $\mathbf{X}_t$ from time 1 to T and similarly S_1^{t-1} denotes all the values of S_t between time 1 and time $t-1$.

Assumption (1) asserts that, given the state of the weather at the previous time point and the current value of some atmospheric variables, the state of the weather at time t does not depend on any other history of states or on any other past or future values of the atmospheric quantities.

The parameterization we adopt for $P(S_t|S_{t-1}, \mathbf{X}_t)$ is

$$P(S_t = j \mid S_{t-1} = i, \mathbf{X}_t) \propto \gamma_{ij} \exp[-\tfrac{1}{2}(\mathbf{X}_t - \mu_{ij})\mathbf{\Sigma}^{-1}(\mathbf{X}_t - \mu_{ij})'] \quad (2)$$

where $\mathbf{\Sigma}$ is the variance-covariance matrix for the atmospheric data and all the atmospheric variables are centered around their mean. The $\boldsymbol{\mu}_{ij}$ parameters represent the mean vectors of the atmospheric variables when the state of the weather at the previous time point was state i and the current state of the weather is j, while the γ_{ij} parameters can be interpreted as baseline transition probabilities. It is necessary to impose the constraints $\sum_j \gamma_{ij} = 1$ and $\sum_j \boldsymbol{\mu}_{ij} = \mathbf{0}$, in order to ensure identifiability of the parameters. The latter constraint preserves the means of the (centered) atmospheric variables.

Eq. (2) has the general form of a multiplicative transition model—$P(S_t|S_{t-1}, \mathbf{X}) = \gamma \cdot g(\mathbf{X})$, where γ is the baseline transition matrix and $g(\mathbf{X})$ is any positive function. Since the atmospheric measures we use are usually continuous and symmetrically distributed, we have found the Gaussian kernel used in Eq. (2) to be a convenient functional form, but other forms are possible.

The other fundamental element in the NHMM is the *observed* stochastic process—in this context precipitation—which is assumed to be conditionally temporally independent, given the weather state. The hidden Markov model assumptions for the observed process can be summarized by

$$f_{\mathbf{R}_t|(S_1^T, \mathbf{R}_1^{t-1}, \mathbf{X}_1^T)}(\mathbf{r}) = f_{\mathbf{R}_t|S_t}(\mathbf{r}) \quad (3)$$

where f denotes a probability density function, $\mathbf{R}_t$ is the vector of precipitation amounts at a network of stations at time t and $\mathbf{R}_1^{t-1}$ indicates all the precipitation data from time 1 to $t-1$. Thus, given the current

weather state, precipitation is assumed independent from all the past precipitation values, all other past and future weather states and from any values of the atmospheric variables.

Assumptions (1) & (3) determine the temporal structure in the precipitation process. The definition of the spatial structure requires additional hypotheses. In the analysis presented here, we assume conditional spatial independence of both occurrences and amounts given the weather state, i.e. we hypothesize that all the dependence between rain gauges is induced by the common weather state. In the discussion we suggest possible extensions to this relatively simple dependence structure.

The parameterization for the observed process builds on the spatial independence model for precipitation occurrences of Hughes & Gattorp (1994). Amounts are introduced by modeling precipitation at each station, given the weather state, as a mixture of a point mass at zero and a gamma distribution. In other words, conditional on the current weather state and on the occurrences, we model the amounts at each gauge as a gamma distribution (with state-specific parameters). The resulting parameterization is

$$f_{\mathbf{R}_t|S_t=s}(\mathbf{r}) = \prod_{i=1}^{N}[p_{si}\boldsymbol{G}(r_t^i - c;\alpha_{si},\beta_{si})]^{1_{[r_t^i>c]}}(1-p_{si})^{1_{[r_t^i\le c]}} \quad (4)$$

where N is the number of rain stations, p_{si} is the precipitation probability at Stn i in state s and r_t^i is the precipitation amount at Stn i and time t. With $\boldsymbol{G}(r_t^i;\alpha_{si}, \beta_{si})$ we indicate the density at r_t^i of a gamma distribution with parameters α_{si} and β_{si} which depend on the state s and the Stn i:

$$\boldsymbol{G}(r;\alpha_{si},\beta_{si}) = \frac{\beta_{is}^{\alpha_{is}}}{\Gamma(\alpha_{is})} r^{\alpha_{is}-1}\mathrm{e}^{-r\beta_{is}} \quad (5)$$

where $\Gamma(\alpha_{is})$ is the gamma function with argument α_{is}. The indicator function $1_{[r_t^i > c]}$ takes on a value of 1 if the precipitation amount at time t and Stn i is above the prespecified cutoff c; it takes on a value of 0 if the precipitation amount is below c. Thus amounts below c are treated as no precipitation.

In the model described by Eqs. (2) & (4) the number of unconstrained parameters is

$$S(S-1)(M+1) + 3SN$$

where M is the number of atmospheric variables included in the model and S the number of weather states.

2.2. Parameter estimation

Parameter estimates are obtained by numerically maximizing the likelihood. The likelihood of the observed data given the atmospheric variables is

$$L(\boldsymbol{\theta}) = f_{\mathbf{R}_1^T|\mathbf{X}_1^T=\mathbf{x}_1^T,\boldsymbol{\theta}}(\mathbf{r}_1^T) \quad (6)$$

$$= \sum_{s_1,\ldots,s_T} f_{(\mathbf{R}_1^T,S_1^T)|\mathbf{X}_1^T=\mathbf{x}_1^T,\boldsymbol{\theta}}(\mathbf{r}_1^T, s_1^T) \quad (7)$$

$$= \sum_{s_1,\ldots,s_T} P(S_1 = s_1|\mathbf{X}_1) f_{\mathbf{R}_1|S_1=s_1}(\mathbf{r}_1) \times \prod_{t=2}^{T} P(S_t = s_t|S_{t-1},\mathbf{X}_t) f_{\mathbf{R}_t|S_t=s_t(\mathbf{r}_t)} \quad (8)$$

where $\boldsymbol{\theta}$ is the vector of the model parameters. In rainfall modeling, the number of observation times, T, is usually large. But even for small Ts, computation of the likelihood directly as in Eq. (8) is intractable. However, the calculation is possible using the recursive forward-backward algorithm, originally developed by Baum (1972). A useful tutorial on this algorithm and others related to HMMs is given in Rabiner & Juang (1986). The main idea is to write the likelihood as:

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

where $\mathbf{B}(\mathbf{r})$ is an $S \times S$ diagonal matrix, with $b_{ss}(\mathbf{r}) = f_{\mathbf{R}_t|S_t=s}(\mathbf{r})$, $\mathbf{A}(\mathbf{x})$ is an $S \times S$ transition matrix with $a_{ij}(x) = P(S_t=j\,|S_{t-1}=i, \mathbf{X}_t = \mathbf{x})$, $\mathbf{1}'$ is a length S column vector of ones and $\boldsymbol{\delta}(\mathbf{x})$ is a row vector of length S. The quantity $\boldsymbol{\delta}(\mathbf{x})$ is the solution to $\boldsymbol{\delta}(\mathbf{x})\mathbf{A}(\mathbf{x}) = \boldsymbol{\delta}(\mathbf{x})$ [i.e. it is the stationary distribution for $\mathbf{A}(\mathbf{x})$].

To maximize the likelihood, we apply the EM algorithm. Hughes et al. (1999) give a detailed description of this procedure.

2.3. Model order

Fitting an NHMM to precipitation data involves the choice of a model order and of the atmospheric variables to be included. We first determine the order of the NHMM, i.e. the number of hidden weather states, and include the atmospheric variables afterwards. The choice of the number of hidden states is a non-trivial issue. Standard likelihood-based methods—such as the Akaike information criterion (AIC) (Akaike 1974) and the Bayesian information criterion (BIC) (see Kass & Raftery 1995 for a review)—rely upon assumptions that do not hold for the order selection problem. Nonetheless, the Bayesian information criterion, defined as:

BIC = −2 loglikelihood
+ log(no. of observations)(no. of free parameters)

yields reasonable models in terms of interpretability and fit to the data (Hughes et al. 1999). Thus, BIC is one of the elements—but not the sole determining factor—that we use in choosing the number of weather states.

Models of different order can also be compared with respect to their capability of reproducing some key

features in the observed data. For example, an important characteristic we try to match is the distribution of the 'storm' durations at the different rain gauges, where 'storm' is defined as a string of consecutive days when precipitation occurred.

Another consideration is the increase in the number of parameters induced by an increase in the number of weather states. Choosing too many states can lead to an intractable model in terms of computer time.

2.4. Atmospheric variables

Atmospheric data are used to help determine the current (hidden) weather state (see Eqs. 1 & 2). To reduce the number of model parameters, we prefer to include relatively few atmospheric variables in the model. However, synoptic-scale atmospheric variables are typically available on regular grids and several grid nodes usually cover the region of interest. Thus, a method for summarizing the grid data into few values is needed. Our approach is based on the singular value decomposition (SVD) technique (von Storch & Zwiers 1999). For each atmospheric field **Y** we compute a matrix **C**, with element c_{ij} given by the correlation between the precipitation process at Stn i and the atmospheric variable **Y** at node j. This matrix is decomposed using the SVD method, to obtain

$$\mathbf{C} = \mathbf{U}\,\mathbf{W}\,\mathbf{V}' \qquad (9)$$

Letting N denote the number of rain gauges and G the number of grid nodes, **U** is a $N \times N$ matrix, **V** is a $G \times N$ matrix and **W** is a diagonal $N \times N$ matrix. The SVD technique ensures that $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$ and the diagonal elements of **W**, $w_1,\ldots,w_N$, are the singular values of the matrix **C**, in non-increasing order. If we standardize the atmospheric variable **Y** at each node j separately and call the resulting field $\mathbf{Y}^{\mathrm{std}}$, we can construct a summary of the original field by multiplying $\mathbf{Y}^{\mathrm{std}}$ by the ith column of the matrix **V**, $\mathbf{V}^{(i)}$. This summary variable explains $\frac{w_i^2}{\sum_k w_k^2}$ of the correlation between the precipitation process and the atmospheric field **Y**. The number of summary variables needed to explain a certain portion of the correlation depends on the relative magnitude of the singular values.

Once the SVD procedure has been applied to each of the atmospheric fields under consideration, the decision on how many and which of the resulting summary variables are to be included in the model is based on BIC.

3. AN APPLICATION

We used the NHMM to analyze precipitation amounts at a network of rain gauges in Washington state. The precipitation data set consists of daily precipitation amounts for the winters (November through March) 1973/74 to 1989/90, at the 24 rain gauges shown in Fig. 1. These data (available on CD ROM—Earthinfo Inc. 1990), were recorded by the National Weather Service and cooperators and corrected by the National Climatological Data center (NCDC) 'Validated Historical Daily Data' project. The cutoff c mentioned in Eq. (4) is 0 inches. The 12 winters 1973/74 to 1984/85 were used for model fitting, while the 5 winters 1985/86 to 1989/90 were reserved for model validation. The atmospheric data consists of daily geopotential height at 1000 and 850 mb, temperature at 850 mb and relative humidity at 1000 and 850 mb from the NCAR/NCEP (National Center for Atmospheric Research/National Center for Environmental Prediction) Reanalysis project, provided through the NOAA (National Oceanic and Atmospheric Administration) Climate Diagnostic Center. These variables are given on a 2.5° latitude × 2.5° longitude grid for the same period as the precipitation data. The area of interest spans 48 grid nodes.

The model fitting procedure was hierarchical. The number of weather states was first determined by fitting HMMs with 2 through 7 states to the occurrence data. Several considerations contributed to the decision to

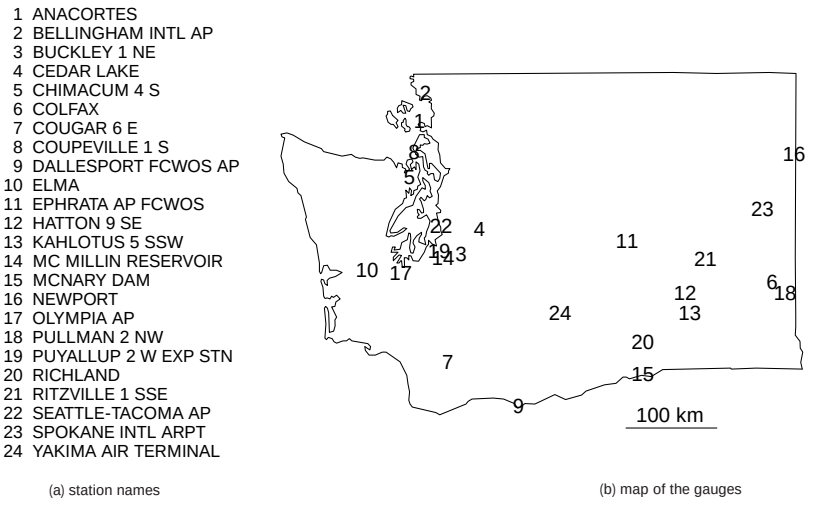


Fig. 1. Map of the rain gauges

include 6 states. The Bayesian information criterion suggested a 'large' number of states, since BIC decreased monotonically as the number of states increased, actually pointing at the 7-state model. However, the 7th state did not seem to improve the fit of the model to the observed storm duration distribution or any other important feature of the data. Thus we focused the remainder of our model building efforts on the 6-state model.

Atmospheric variables were added to the 6-state model after performing the SVD decomposition on each of the 5 fields—geopotential height at 1000 and 850 mb, temperature at 850 mb and relative humidity at 1000 and 850 mb—to summarize the 48 grid values into a few quantities. A few summary variables explain most of the correlation between each field and the precipitation process, as shown in Table 1.

As an example of the type of summary variables obtained with the SVD technique, consider the first linear combination variable for geopotential height at 1000 mb. The weights assigned to each grid node are highest just at the northwest of Washington state, and decay in all directions away from this region. The resulting summary variable can be interpreted as a weighted mean of the standardized 1000 mb geopotential height field.

Several NHMMs with 6 states were fit to precipitation occurrences using different combinations of the selected summary variables, and BIC was used to choose the best model. The model that minimizes BIC contains 2 atmospheric variables: the first summary variable for geopotential height at 1000 mb and the first summary variable for relative humidity at 850 mb.

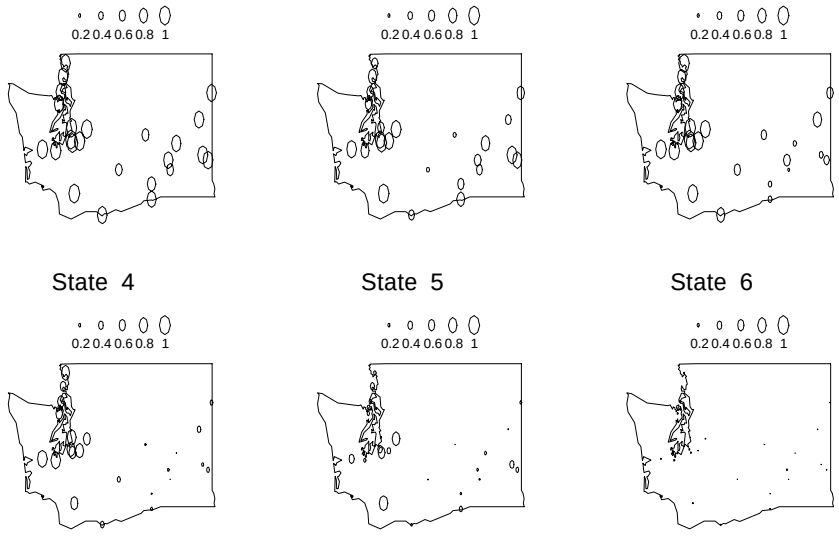
A NHMM with 6 states and including the first summary variables for geopotential height at 1000 mb and relative humidity at 850 mb was then fit to the precipitation amounts. The 6 states identified by the NHMM correspond to the precipitation patterns in Fig. 2a. States 1 and 6 are clear cut wet and dry respectively, for all the stations in the network. The other states correspond to intermediate patterns that reflect regional differences. Fitting the 6-state NHMM to occurrences only leads to very similar patterns, indicating that the inclusion of amounts does not seem to substantially change the state definitions, in terms of precipitation probabilities. The 6 weather states also correspond to different amount distributions. For each state, Fig. 2b shows the distribution of the positive precipitation amounts at Puyallap, in the South Puget region. Larger amounts correspond to the predominantly wet states, especially State 1, where the precipitation probability is large at all stations. In predominantly dry states, when precipitation occurs the amounts tend to be smaller. State 4, which is dry in Eastern Washington and relatively wet around Puget Sound, corresponds to smaller amounts with respect to the first 3 states, even at the stations where the precipitation probability remains fairly large.

To produce the amount distributions shown in Fig. 2b, each day has to be classified into 1 of the weather states defined by the NHMM. The Viterbi algorithm (Rabiner & Juang 1986) identifies the most probable sequence of states with resulting relative frequencies of the weather states 16, 15, 14, 18, 12 and 25%. Averaging the geopotential height at 1000 mb field over all days classified into a particular state gives the predominant pattern associated with that state. The same procedure leads to the predominant 850 mb relative humidity pattern associated with each of the 6 weather states. One may compare these atmospheric patterns to the corresponding precipitation patterns in Fig. 2a. Fig. 3 shows the contour plots for geopotential height at 1000 mb and relative humidity at 850 mb for all 6 states. State 6 is characterized by a high pressure system and low relative humidity over the Washington region, which correspond to low precipitation probability. In State 1, low pressure at the northwest of Washington State and high moisture over the entire region correspond to the high precipitation probability at all stations. The other atmospheric patterns are consistent with the observed precipitation patterns and suggest that some of the weather states might be regarded as 'transition states'. For example, we find that State 5 typically moves to either State 4 or more likely State 6, but tends not to persist.

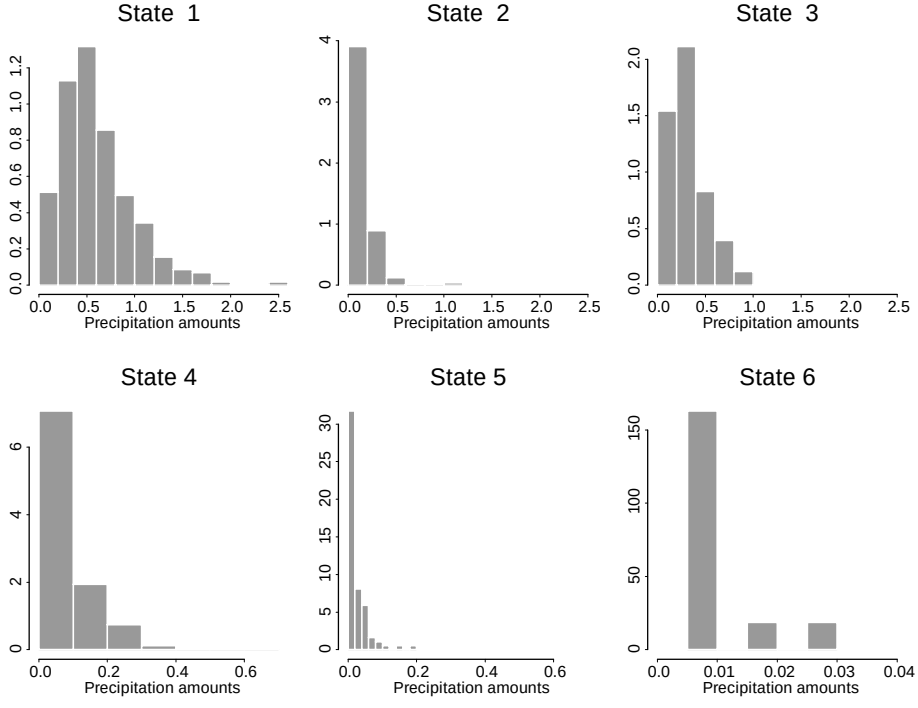
Some indications of how well the NHMM fits the data derives from the comparison between observed and model-based precipitation probabilities (Fig. 4a), and between observed and model-based log odds

Table 1. Proportion of correlation explained by the summary variables. gph: geopotential height; tem: temperature; hum: relative humidity

	gph 1000		gph 850		tem 850		hum 1000		hum 850	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
$\frac{w_i^2}{\sum_k w_k^2}$	0.95	0.03	0.96	0.03	0.84	0.11	0.91	0.07	0.96	0.03



(a) Precipitation probabilities at the 24 rain gauges



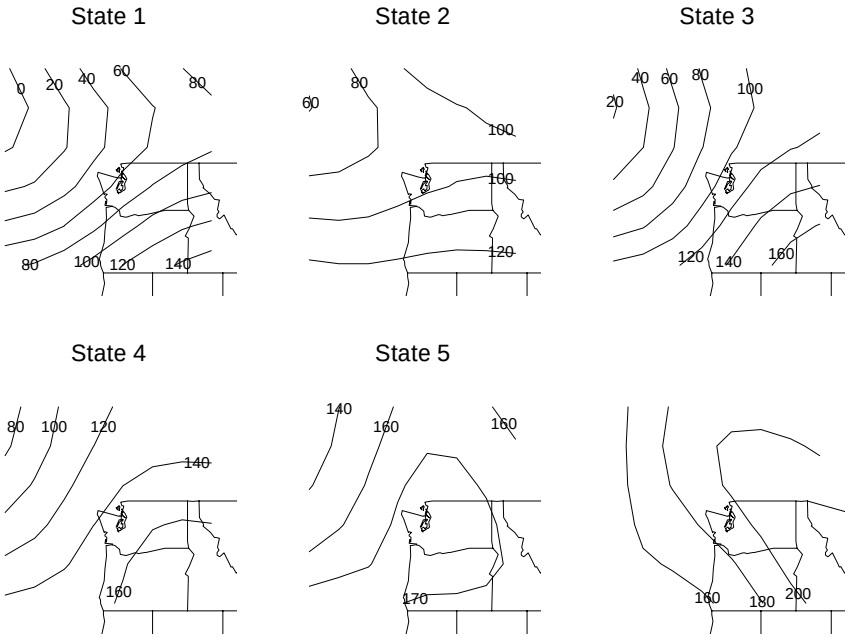
(b) Histograms of precipitation amounts at Puyallap (South Puget area)

Fig. 2. Precipitation probabilities and histograms of amounts (in inches) corresponding to the 6 weather states identified by the NHMM (nonhomogeneous hidden Markov model) including the first summary variables for geopotential height at 1000 mb and relative humidity at 850 mb

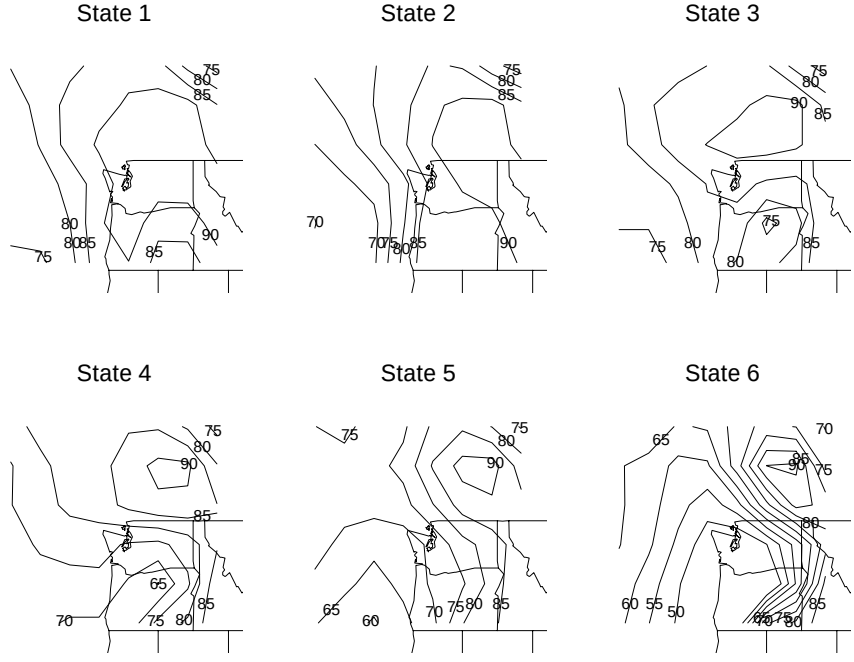
ratios (Fig. 4b). The log odds ratio is a common measure of association for binary variables and it is used in Fig. 4b to reflect the spatial correlation between occurrences at each pair of stations. The log odds ratio for Stns i and j can be defined as

$$\log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right) \tag{10}$$

where we denote n_{11} as the number of days when precipitation occurs at both Stns i and j, n_{00} as the number



(a) geopotential height at 1000 mb (meters)



(b) relative humidity at 850 mb (percentage)

Fig. 3. Contour of the geopotential height at 1000 mb and relative humidity at 850 mb fields for each weather state identified by the NHMM

of days when precipitation occurs neither at Stn i nor at Stn j, n_{10} as the number of days when precipitation occurs at Stn i but not at Stn j, and n_{01} as the number of days when precipitation occurs at Stn j but not at Stn i. The quantity in Eq. (10) takes on values in $(-\infty, +\infty)$, with large negative numbers indicating strong negative association, large positive numbers corresponding to strong positive association and values close to 0 reflecting a weak association. In Fig. 4, the precipitation probabilities are reproduced well, while the log

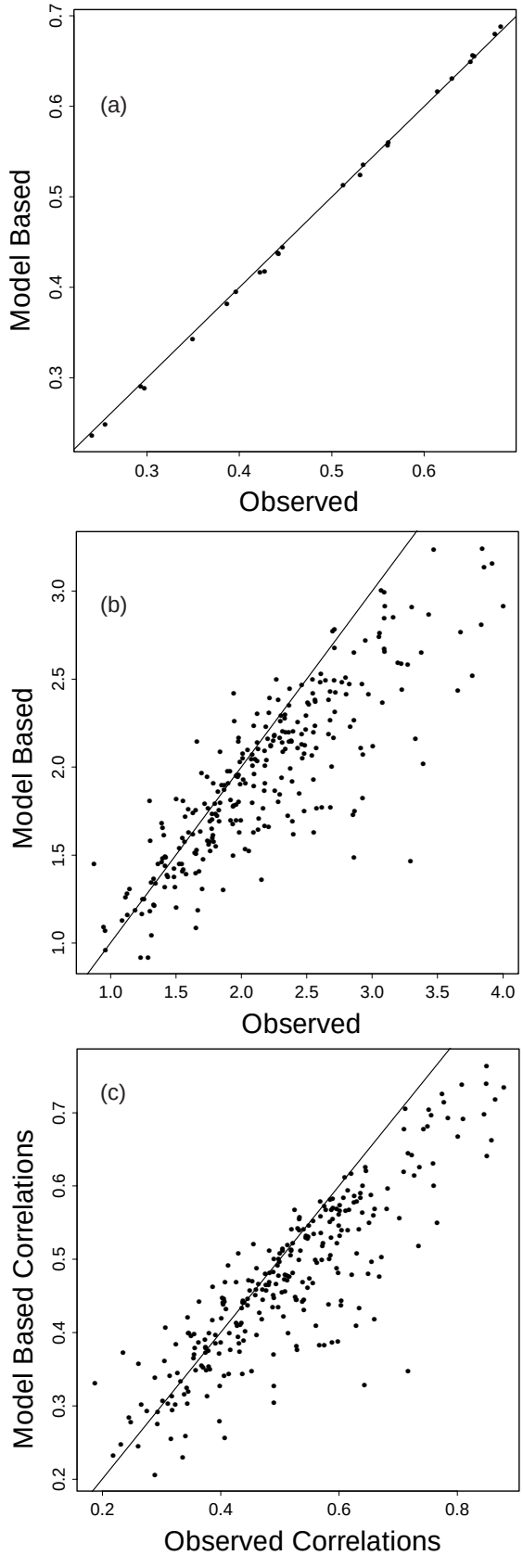


Fig. 4. Observed versus model based (a) precipitation probabilities, (b) log odds ratios and (c) correlations of positive amounts between all station pairs (Spearman coefficient). The model-based quantities are obtained by simulating data from the 6-state NHMM for amounts

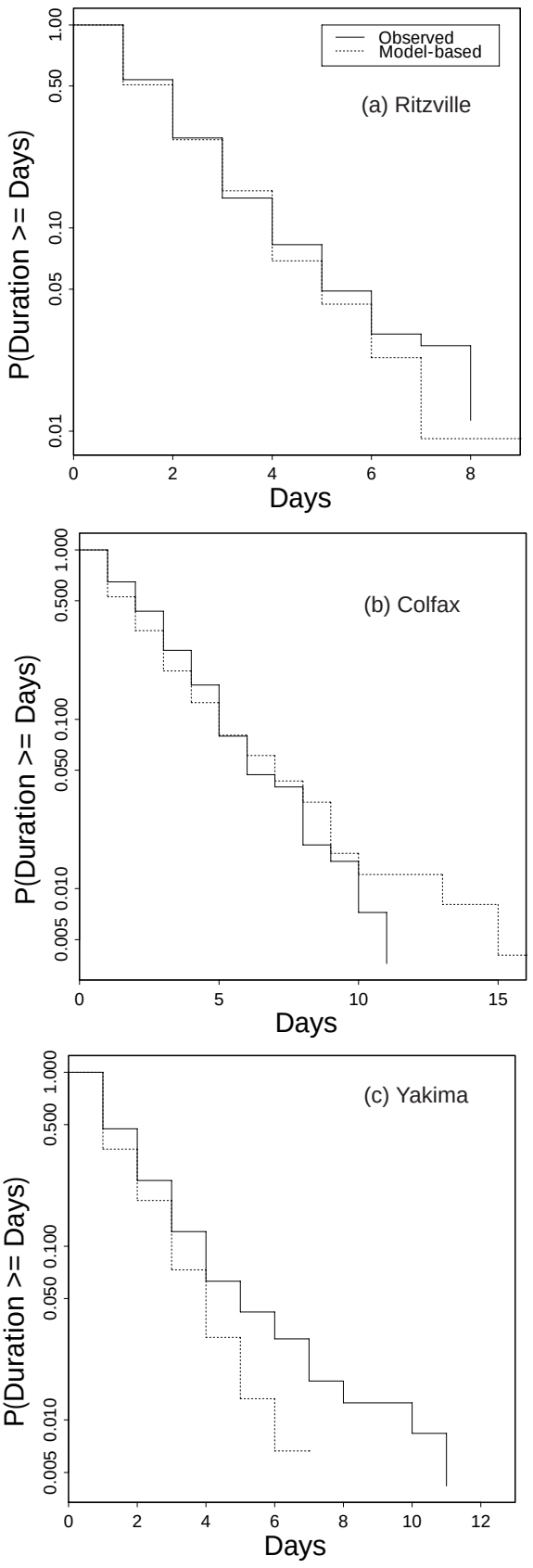


Fig. 5. Observed and model-based distributions of the storm durations. Model-based storm durations were computed from simulations produced by the 6-state NHMM

odds ratios are modeled less adequately, especially when the observed correlation is high. This indicates that the hypothesis of conditional spatial independence, given the weather state, may need to be modified. The common weather state seems to explain much of the correlation, but additional unexplained local spatial correlation remains. A similar conclusion is suggested by Fig. 4c, which shows the Spearman correlation coefficient corresponding to the precipitation amounts at each station pair. A relatively low spatial correlation between amounts, as well as between occurrences, can be adequately captured by the common weather state, but when the correlation between gauges is strong, the weather state is not sufficient to account for all of it.

Temporal correlation in the data is also a feature that the model should capture. The distribution of storm durations, which is often important in hydrological applications, seems to be reproduced quite well at most rain gauges. Fig. 5 shows examples from the Eastern Washington region chosen to represent from best to worst fit. Plots similar to Fig. 5a,b are typical of the Puget Sound region too, while Fig. 5c displays the worst fit among all the rain stations and is not representative of the fit at any other location.

Another issue is whether the gamma distribution is an appropriate choice to model the conditional distribution of precipitation amounts, given occurrence and the weather state. The fit varies from station to station; Fig. 6 shows quantile-quantile plots (qqplots) of observed versus model-based precipitation amounts at 3 representative stations from 3 geographical regions in Washington State. In general the distribution of precipitation amounts is best modeled at the stations in the South Puget area, while the North Puget stations correspond to the worst fit. The Eastern Washington region, which is the driest area, shows the largest variability in fit from gauge to gauge.

Plots similar to those in Figs. 4 & 6 were obtained using the reserved data. The final 6-state NHMM (which was fit using 1973 to 1985 data), together with geopotential height at 1000 mb and relative humidity at 850 mb for the 1985 to 1990 period, was used to generate precipitation amounts for the 1985 to 1990 winters. The SVD weights obtained previously were applied to form the summary atmospheric variables from the 1985 to 1990 geopotential height and relative humidity fields. The comparison of various statistics for the observed and generated 1985 to 1990 precipitation amounts indicates how well the model captures the characteristics of the reserved data. Fig. 7 shows the observed versus model-based precipitation probabilities at all stations. The model underestimates the precipitation probability at most stations. To determine if this consistent underprediction was due to model mis-

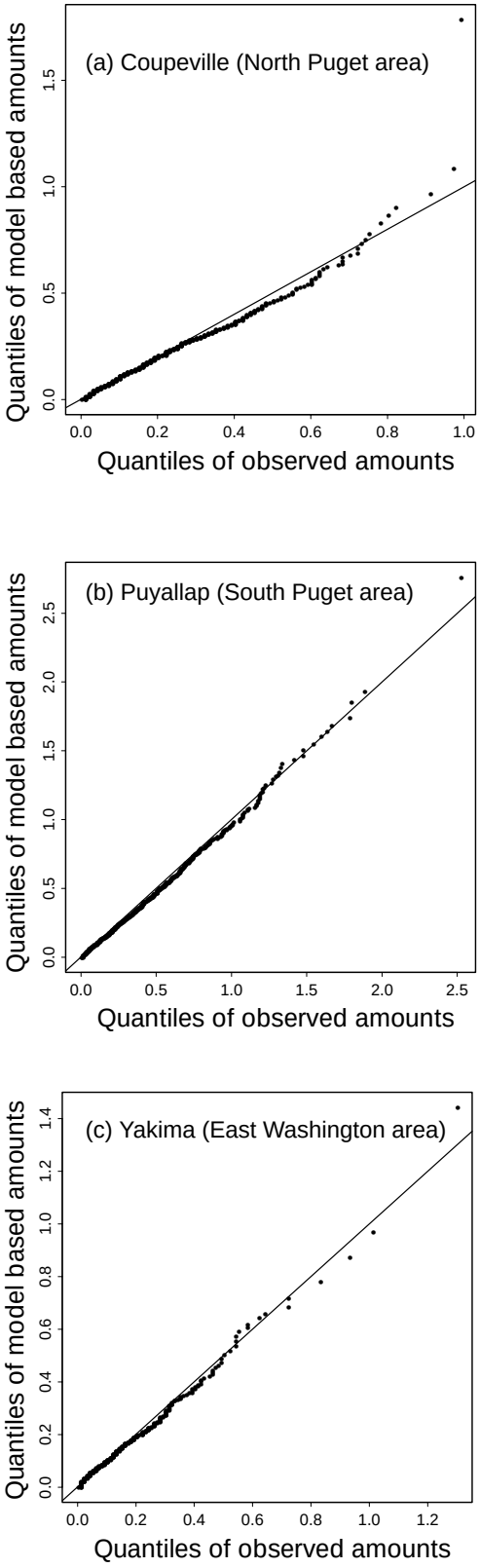


Fig. 6. Quantile-quantile plots of observed versus model-based amounts (in inches) at selected stations. The model-based amounts are simulated from the 6-state NHMM

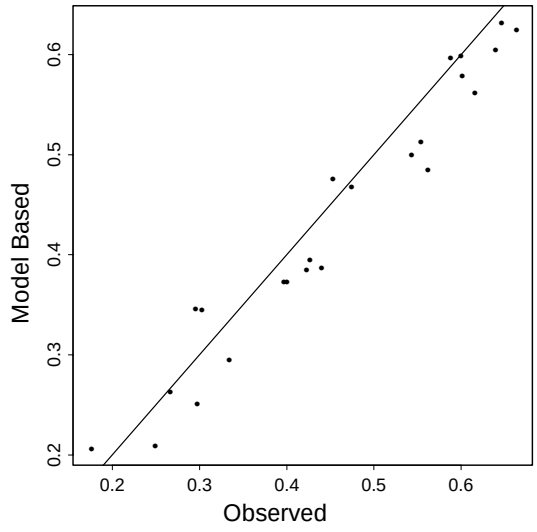


Fig. 7. Observed versus model-based precipitation probabilities for the reserved period, 1985/86 to 1989/90

specification, we generated several 5 yr realizations from the NHMM and compared the resulting 'observed' precipitation probabilities with the 'model-based' probabilities obtained by averaging over many sets of 5 yr realizations. Even for these cases, where the observations are a realization from the model, the 'observed probabilities' are typically mostly smaller or mostly larger than the 'model-based' ones. Thus the consistent underprediction of the 'model-based' precipitation probabilities can be explained by spatially correlated random variation and does not necessarily indicate model misspecification.

Another characteristic of the reserved data that should be captured by the model is the conditional distribution of precipitation amounts, given occurrence. Fig. 8 shows the qqplots of observed versus model-based precipitation amounts at the same stations as in Fig. 6. The model seems to reproduce the distribution of observed precipitation amounts reasonably well overall, although the fit varies from station to station.

4. DISCUSSION

The model described in this paper can be used to generate simulations of precipitation amounts that incorporate synoptic atmospheric information. The hidden Markov model assumptions simplify the temporal and spatial structures to be parameterized, since the common *weather state* accounts for the temporal dependence and much of the spatial correlation between rain gauges. Several possible improvements to the model are currently under investigation, includ-

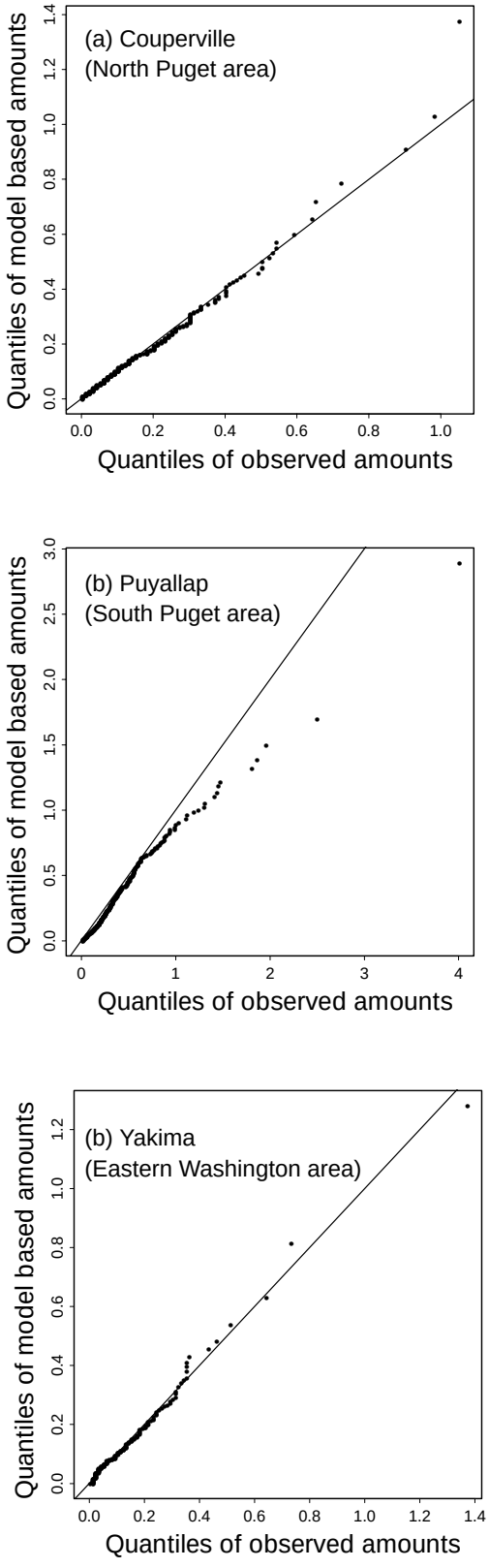


Fig. 8. Quantile-quantile plots of observed versus model-based amounts (in inches) at selected stations for the reserved period, 1985/86 to 1989/90

ing more realistic spatial dependence structures and reduced parameterizations.

The conditional spatial independence structure adopted in the present application is relatively simple. Although this assumption captures most of the correlation between rain gauges, Fig. 4b,c suggests the need to include some additional dependence in the model. We plan to investigate 2 alternative structures. The first step is to introduce dependence between precipitation *occurrences* and assume conditional spatial independence of *amounts* given occurrences and the weather state. The autologistic model of Hughes et al. (1999) can be adopted to describe the dependence of precipitation occurrences at different rain gauges. Precipitation amounts, conditional on occurrences, would be modeled independently at each gauge as in the previous sections.

If this structure still does not account for all the observed correlation between rain gauges, more complicated models which allow for interactions between both precipitation occurrences and amounts at different stations will be considered. The spatial dependence between occurrences could still be described by the autologistic model and, conditional on occurrences, the amounts could be modeled jointly at all stations through a multivariate gamma or exponential distribution.

The proposed modifications to the spatial dependence structure would increase the number of parameters, already large in the NHMM applied to the Washington State data. A possibility that will need to be investigated is the reduction of the number of parameters, both in the hidden and observed parts of the model. One reasonable modification of the hidden part is to have only 1 vector μ_j of means of the atmospheric variables for each state *j*, regardless of the state of the system at the previous time point. In the observed part of the NHMM, one could specify some function of the precipitation amount parameters to have a common value at all stations within a sub-region.

The likelihood for an NHMM is a non trivial function of a large number of parameters and may have several local maxima. In this situation, the estimates resulting from numerical maximization can only be guaranteed to correspond to a local maximum, which not necessarily is also the global one. It is then important to choose reasonable initial values for the maximization routines.

Models like the NHMM can be used to study the effect of *climate variability*. Repeated GCM simulations under current climate conditions can constitute different realizations of the atmospheric fields included in the model. The NHMM can be used to generate occurrences and amounts for each realization, thereby downscaling the effect of the variability in the synoptic-scale variables to precipitation. The effect of *climate change* is another issue that can be investigated using NHMMs. The output of GCM runs under altered climate conditions can serve as input into the downscaling model described here. Thus, the effects of the altered climate scenario could be downscaled to the local-scale precipitation processes by generating precipitation occurrences and amounts from the NHMM. For this application of the NHMM to be valid, the relationship between the synoptic-scale atmospheric variables and the local scale precipitation, as found under the model fitting conditions, would have to hold also under the altered climate. A promising result in this direction is given in Hughes et al. (1999). The authors fitted a NHMM to rainfall occurrence data in south-western Australia and verified that the model responded to shifts in atmospheric circulation in a reserved data set. Charles et al. (1999) discuss issues related to validation of downscaling models for studying climate change.

Acknowledgements. The research described in this article has been funded in part by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, and by NSF grant DMS-9524770. It has not been subjected to EPA's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no offical endorsement should be inferred.

LITERATURE CITED

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19:716–723

Bardossy A, Plate EJ (1992) Space-time models for daily rainfall using atmospheric circulation patterns. Water Resour Res 28:1247–1259

Bartholy J, Bogardi I, Matyasovszky I (1995) Effect of climate change on regional precipitation in Lake Balaton watershed. Theor Appl Climatol 51:237–250

Baum L (1972) An inequality and associated maximization technique instatistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8

Charles SP, Bates BC, Whetton PH, Hughes JP (1999) Validation of downscaling models for changed climate conditions: case study of southwestern Australia. Clim Res 12: 1–14

EarthInfo, Inc (1996) NCDC summary of the day. Available from EarthInfo, Inc, Boulder, CO

Gabriel KR, Neumann J (1962) A Markov chain model for daily rainfall occurrences at Tel-Aviv. Q J R Metorol Soc 88:85–90

Giorgi F, Mearns LO (1991) Approaches to the simulation of regional climate change: a review. Rev Geophys 29: 191–216

Goodall CR, Phelan MJ (1991) Edge-preserving smoothing and the assessment of point process models for Gate rainfall fields. In: Prabhu NU, Basawa IV (eds) Statistical inference in stochastic processes. Marcel Dekker, Inc, New York, p 35–66

Hay L, McCabe J, Wolock DM, Ayers MA (1991) Simulation of precipitation by weather type analysis. Water Resour Res 27:493–501

Hobbs PV, Locatelli D (1978) Rainbands, precipitation cores

and generating cells in a cyclonic storm. J Atmos Sci 35: 230–241
Hughes JP, Guttorp P (1994) A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water Resour Res 30:1535–1546
Hughes JP, Lettenmaier DP, Guttorp P (1993) A stochastic approach for assessing the effects of changes in regional circulation patterns on local precipitation. Water Resour Res 29:3303–3315
Hughes JP, Guttorp P, Charles S (1999) A nonhomogeneous hidden Markov model for precipitation occurrence. J R Stat Soc C 48:15–30
Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D (1996) The NCEP/NCAR reanalysis project. Bull Am Meteorol Soc 77:437–471
Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90: 773–795
Katz RW (1977) An application of chain-dependent processes to meteorology. J Appl Prob 14:598–603
Katz RW, Parlange MB (1996) Mixtures of stochastic processes: application to statistical downscaling. Clim Res 7: 185–193
Le Cam LM (1961) A stochastic description of precipitation. In: Neyman J (ed) Proc 4th Berkeley Symp Math Stat Prob 3:165–186
Rabiner LR, Juang BH (1986) An introduction to hidden Markov models. IEEE Acoustic Speech Signal Process Mag 3:4–16
Stern RD, Coe R (1984) A model fitting analysis of daily rainfall data. J R Stat Soc A 147:1–34
von Storch H, Zwiers F (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge
Waymire E, Gupta VK, Rodriguez-Iturbe I (1984) A spectral theory of rainfall intensity at the meso-β scale. Water Resour Res 20:1453–1465
Woolhiser DA (1992) Modeling daily precipitation—progress and problems. In: Guttorp P, Walden A (eds) Statistics in the environmental and earth sciences. Griffin, London, p 71–89
Zucchini W, Guttorp P (1991) A hidden Markov model for space-time precipitation. Water Resour Res 27:1917–1923

Editorial responsibility: Hans von Storch, Geesthacht, Germany

Submitted: February 22, 1999; Accepted: November 9, 1999
Proofs received from author(s): April 3, 2000