Appl. Statist. (1999)
48, *Part* 1, *pp.*15–30

A non-homogeneous hidden Markov model for precipitation occurrence

James P. Hughes and Peter Guttorp

University of Washington, Seattle, USA

and Stephen P. Charles

Commonwealth Scientific and Industrial Research Organisation, Wembley, Australia

[Received January 1997. Final revision January 1998]

Summary. A non-homogeneous hidden Markov model is proposed for relating precipitation occurrences at multiple rain-gauge stations to broad scale atmospheric circulation patterns (the so-called 'downscaling problem'). We model a 15-year sequence of winter data from 30 rain stations in south-western Australia. The first 10 years of data are used for model development and the remaining 5 years are used for model evaluation. The fitted model accurately reproduces the observed rainfall statistics in the reserved data despite a shift in atmospheric circulation (and, consequently, rainfall) between the two periods. The fitted model also provides some useful insights into the processes driving rainfall in this region.

Keywords: Climate change; EM algorithm; Hidden Markov model; Monte Carlo maximum likelihood; Precipitation model

1. Introduction

Stochastic models for precipitation can aid in understanding the probabilistic structure of precipitation and are important for generating simulations that can be used as input into models of, for example, flooding, run-off, stream flow and crop growth. Until recently, however, most stochastic precipitation models considered rainfall in isolation, without reference to the atmospheric processes that drive it (e.g. Gabriel and Neuman (1962), Stern and Coe (1984), LeCam (1961), Waymire and Gupta (1981) and Kavvas and Delleur (1981)). In part, this reflected the absence of good, long-term records of atmospheric circulation. Such models have several limitations, however. In developing hydrologic models, for instance, researchers may use information on temperature, solar radiation and other climatic factors in addition to precipitation. Ideally, the precipitation model should produce simulations which are consistent with these other inputs into the hydrologic model. Also, precipitation models which exclude atmospheric information can only be used to simulate rainfall under climatic conditions which are stochastically similar to those used to fit the model, yet the atmospheric processes that drive precipitation may be non-stationary, even over relatively short time periods (i.e. decades). Thus, the ability of these models to produce precipitation simulations for periods other than those used to fit the model (or even for subintervals of this period) is limited. In particular, a model which fails to incorporate atmospheric information will not be useful in studies of climate variability or climate change.

Address for correspondence: James P. Hughes, Department of Biostatistics, University of Washington, Seattle, WA 98195, USA.
E-mail: hughes@biostat.washington.edu

 0035–9254/99/48015

Over the past few decades advances in gathering data and our understanding of atmospheric circulation have led to the availability of high quality sets of atmospheric data of variable length (typically, 15–40 years). In addition, the development of physically based, three-dimensional dynamic models of global circulation — general circulation models (GCMs) — has led to the creation of realistic simulations of atmospheric circulation of essentially unlimited duration (some background on GCMs can be found in Intergovernmental Panel on Climate Change (1995)). To take advantage of these types of data, and to address the problems noted above, a new class of stochastic precipitation models known as 'weather state models' or 'downscaling models' has been developed. Recent efforts include papers by Hay *et al.* (1991), Bardossy and Plate (1992), Kidson (1994) and others. Weather state models condition precipitation on available atmospheric information. These models can be thought of as 'conditionally stationary' in that any non-stationarity in large scale atmospheric circulation is (hopefully) captured by the conditioning variables.

Weather state models can be used to generate realistic precipitation simulations by using historical sequences of atmospheric data. Such an approach guarantees that the precipitation simulations will be consistent with the observable atmospheric information. In addition, weather state models can be used with atmospheric simulations from GCMs to study the effects of climate variability on precipitation. In this respect, weather state models provide important data that cannot, at present, be obtained from GCM simulations. The spatial resolution of GCMs is constrained by both computational considerations as well as our understanding of atmospheric dynamics to scales of approximately 2–5° of longitude and latitude. Precipitation, however, varies on much more local scales. For these reasons, GCMs have been unable to generate realistic simulations of rainfall (Giorgi and Mearns, 1991). Weather state models have been proposed as a method of downscaling the GCM atmospheric simulations to local precipitation.

A final, more speculative, application of weather state models is to investigate the effect of hypothesized climate changes (e.g. the 'greenhouse effect') on precipitation. These predictions are based on experiments with GCMs in which a particular atmospheric condition (e.g. carbon dioxide level) is systematically altered. Potentially, a weather state model could be used to downscale the GCM results and to assess the effect of the altered climate on precipitation. Although this clearly represents an important application of weather state models, relatively little work has been done to investigate the validity of these models under climate regimes that are different from those used for model development.

In this paper we use the non-homogeneous hidden Markov model (NHMM) described by Hughes and Guttorp (1994) to relate atmospheric circulation to precipitation occurrence at 30 rain-gauge stations in south-western Australia. The study area and data are described in Section 2. Section 3 describes the model and an estimation procedure is developed which extends the utility of the NHMM to spatially dense networks of stations. In Section 4 we evaluate and interpret the model fitted and demonstrate that the model can detect shifts in precipitation frequency that result from changes in atmospheric circulation. In Section 5 we discuss future directions for research on downscaling models and the NHMM.

2. Study area and data

South-western Australia experiences a 'mediterranean' climate with mild wet winters and hot dry summers. 80% of the annual precipitation falls during the winter months of May–October. Winter rainfall is primarily driven by the successive passage of cold fronts lodged between high pressure systems at latitudes 30° and 35° south. Rainfall amounts are plentiful

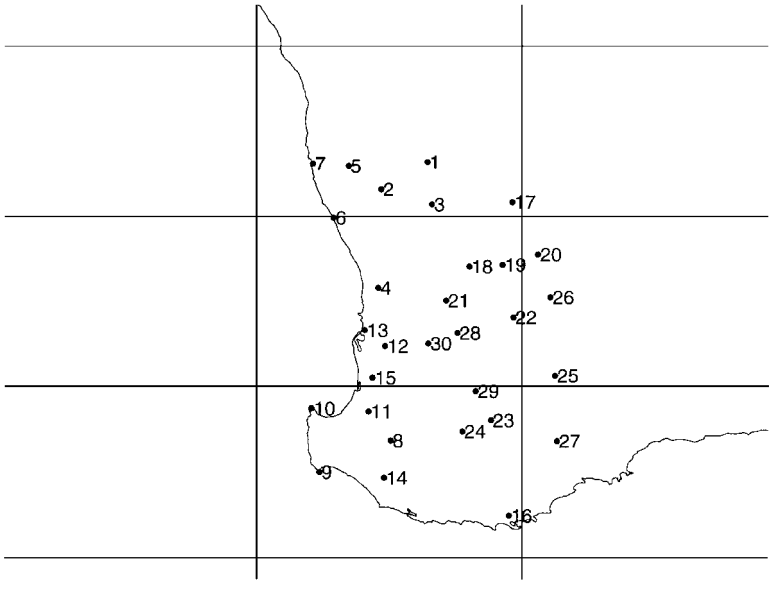

Fig. 1. Map of the study area showing the locations of the atmospheric data grid and rain-gauge stations in south-western Australia: atmospheric data are interpolated to the vertices of the grid as described in the text

and the wet season is relatively long owing to the advection of moist air by strong westerly winds, the presence of a warm offshore current and the orographic uplift provided by the Darling Range (an escarpment 300 m high running north–south for much of the region, approximately 25–50 km inland from the west coast).

For this study a 15-year record (1978–1992) of daily winter rainfall occurrences (2760 days, total) at 30 stations in south-western Australia was used. The locations of the stations are shown in Fig. 1. Each rainfall value represents the total rainfall over a 24-h period ending at 9.00 a.m. which was then dichotomized into the presence or absence of rainfall depending on whether the amount was greater or less than 0.3 mm respectively. Atmospheric data were obtained on a Lambert conformal grid and interpolated to a rectangular grid of similar scale — 2.25° latitude by 3.75° longitude (also shown in Fig. 1). Available atmospheric measures included sea-level pressure, geopotential height at 850 hPa (hectopascals) and 500 hPa, air temperature, dew point temperature and u (north–south) and v (east–west) wind speed components. The atmospheric measurements were taken at 7.00 p.m. on the preceding day (i.e. approximately the middle of the 24-h rainfall period). The first 10 years of data are used for model fitting and the last 5 years are reserved for model evaluation.

3. Methods

3.1. Model

Hughes and Guttorp (1994) described a class of models (NHMMs) to relate broad scale atmospheric circulation patterns to local rainfall. Effectively, they postulated an unobserved discrete-valued process — the 'weather state' — which acts as a link between the two disparate scales. Formally, let $\mathbf{R}_t = \{R_t^1, \ldots, R_t^n\}$ be a multivariate random vector giving rainfall occurrences at a network of n sites with observed value $r_t^i = 1$ if rain occurs on day t at station i and $r_t^i = 0$ otherwise. Let S_t be the weather state at time t and $\mathbf{X}_t$ be the vector of atmospheric

measures at time t for $1 \leqslant t \leqslant T$. The $\mathbf{X}_t$ will usually consist of one or more derived measures from the available atmospheric data (e.g. north–south gradient in sea-level pressure). The notation ${}_1\mathbf{X}_T$ will be used to indicate the sequence of atmospheric data from time 1 to T and similarly for ${}_1\mathbf{R}_T$ and ${}_1S_T$.

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

$$P(\mathbf{R}_t|{}_1S_T, {}_1\mathbf{R}_{t-1}, {}_1\mathbf{X}_T) = P(\mathbf{R}_t|S_t) \qquad \text{(assumption 1);}$$

$$P(S_t|{}_1S_{t-1}, {}_1\mathbf{X}_T) = P(S_t|S_{t-1}, \mathbf{X}_t) \qquad \text{(assumption 2)}$$

and $P(S_1|{}_1\mathbf{X}_T) = P(S_1|\mathbf{X}_1)$. Specific models are defined by parameterizing $P(\mathbf{R}_t|S_t)$ and $P(S_t|S_{t-1}, \mathbf{X}_t)$ as discussed below.

The first assumption states that the rainfall occurrence process $\mathbf{R}_t$ is conditionally independent given the current weather state. In other words, all the temporal persistence in precipitation is captured by the persistence in the weather state described in assumption (2). Assumption 2 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. In the absence of the atmospheric data this is simply the Markov assumption applied to the weather state process. The resulting model is an example of a hidden Markov model (see Juang and Rabiner (1991) for a review of hidden Markov models). The atmospheric data, when included, are used to modify the transition probabilities of the Markov process — hence the term 'non-homogeneous'. Most weather state models described in the literature define the weather states as deterministic functions of the atmospheric variables. These models can be written as special cases of the NHMM by forcing $P(S_t|S_{t-1}, \mathbf{X}_t)$ to be degenerate (see Hughes and Guttorp (1994), Table 1).

To parameterize $P(\mathbf{R}_t|S_t)$ we use the autologistic model for multivariate binary data:

$$P(\mathbf{R}_t = \mathbf{r}|S_t = s) \propto \exp\left(\sum_{i=1}^{n} \alpha_{si} r^i + \sum_{j<i} \beta_{sij} r^i r^j\right) \tag{1}$$

where both α_{si} and β_{sij} must be finite. β_{sij} is the 'conditional log-odds ratio' of rain at station i to rain at station j (in state s) based on the probability distribution $P(r^i, r^j|\mathbf{r}^{-i,-j}, S_t = s)$ where $\mathbf{r}^{-i,-j}$ is the vector of rainfall occurrences at all sites other than i and j. When β_{sij} is positive, stations i and j are positively associated (within weather state s) whereas a negative value for β_{sij} implies a negative association. To reduce the number of parameters in this model it will often be reasonable to model β_{sij} as a function of the distance and direction between stations i and j (an example of this is described in Section 4).

An important special case of model (1) arises when $\beta_{sij} = 0$ for all i, j and s. Then

$$P(\mathbf{R}_t = \mathbf{r}|S_t = s) = \prod_{i=1}^{n} p_{si}^{r^i}(1 - p_{si})^{1-r^i} \tag{2}$$

where $p_{si} = \exp(\alpha_{si})/\{1 + \exp(\alpha_{si})\}$. The p_{si} give the probability of rain at station i in weather state s. This will be referred to as the 'conditional independence model' for $P(\mathbf{R}_t|S_t = s)$ since the rainfall occurrences R_t^i are assumed to be spatially independent, conditional on the weather state. Unconditionally, however, the R_t^i will be correlated owing to the influence of the common weather state. Hughes and Guttorp (1994) investigated this model and presented an example of a spatially dispersed network of rain-gauge stations for which the conditional independence model worked well.

To parameterize $P(S_t|S_{t-1}, \mathbf{X}_t)$ we first rewrite this probability as a base-line transition matrix and a multiplicative function of the covariates. As noted above, the relevant atmospheric measures $\mathbf{X}_t$ are usually derived measures (typically, linear combinations) of high dimensional atmospheric fields. Thus, it is often reasonable to assume that the $\mathbf{X}_t$ are multivariate normal. This leads to the following model for $P(S_t|S_{t-1}, \mathbf{X}_t)$:

$$\begin{aligned} P(S_t = j|S_{t-1} = i, \mathbf{X}_t) &\propto P(S_t = j|S_{t-1} = i)\, P(\mathbf{X}_t|S_{t-1} = i, S_t = j) \\ &= \gamma_{ij} \exp\{-\tfrac{1}{2}(\mathbf{X}_t - \mu_{ij})\mathbf{V}^{-1}(\mathbf{X}_t - \mu_{ij})'\} \end{aligned} \tag{3}$$

where μ_{ij} is the mean of $\mathbf{X}_t$, conditional on S_{t-1} and S_t, and $\mathbf{V}$ is the corresponding covariance matrix. This model shows clearly how the NHMM is a general version of the simpler homogeneous Markov model. The γ_{ij} may be thought of as elements of the base-line transition matrix of the weather state process which corresponds to the transition matrix of a homogeneous Markov model. The exponential term quantifies the effect of the atmospheric data on the base-line transition matrix. To ensure identifiability of the parameters, the constraints $\Sigma_j \gamma_{ij} = 1$ and $\Sigma_j \mu_{ij} = \mathbf{0}$ are imposed. Also, the estimates of γ and μ are only defined up to a scale factor which depends on $\mathbf{V}$. Although $\mathbf{V}$ is arbitrary we have found that setting $\mathbf{V}$ equal to the raw covariance matrix of $\mathbf{X}_t$ improves the numerical stability of the model.

3.2. Parameter estimation

Letting θ denote the model parameters, the likelihood can be written as

$$\begin{aligned} L(\theta) &= P({}_1\mathbf{R}_T|{}_1\mathbf{X}_T, \theta) \\ &= \sum_{S_1,\ldots,S_T} P({}_1\mathbf{R}_T, {}_1S_T|{}_1\mathbf{X}_T, \theta) \\ &= \sum_{S_1,\ldots,S_T} P(S_1|\mathbf{X}_1) \prod_2^T P(S_t|S_{t-1}, \mathbf{X}_t)\, P(\mathbf{R}_t|S_t). \end{aligned} \tag{4}$$

Computation of the likelihood is made tractable by successively passing each of the multiple summations in equation (4) as far to the right as possible. For example, the summation over S_T may be passed through all terms in the product except the Tth term.

Baum *et al.* (1970) developed an algorithm (later shown to be equivalent to the EM algorithm of Dempster *et al.* (1977)) to obtain maximum likelihood estimates for hidden Markov models by considering the hidden states ${}_1S_T$ to be 'missing' data. This same approach may be used with the NHMM. Let $\theta = (\theta_R, \theta_S)$, where θ_R and θ_S are the parameters of the observed and hidden processes respectively. Then, the EM algorithm for the NHMM may be written as

(a) *E-step* — compute

$$\begin{aligned} \upsilon_t(s) &= P(S_t = s|{}_1\mathbf{R}_T, {}_1\mathbf{X}_T, \theta), \\ w_t(s_1, s_2) &= P(S_{t-1} = s_1, S_t = s_2|{}_1\mathbf{R}_T, {}_1\mathbf{X}_T, \theta), \end{aligned} \tag{5}$$

(b) *M-step* — maximize

$$\begin{aligned} \Psi(\theta_R'|\theta) &= \sum_{t,s} \upsilon_t(s)\, \ln\{P(\mathbf{R}_t|s, \theta_R')\}, \\ \Psi(\theta_S'|\theta) &= \sum_{t,s_1,s_2} w_t(s_1 s_2)\, \ln\{P(S_t = s_2|S_{t-1} = s_1, \mathbf{X}_t, \theta_S')\} \end{aligned} \tag{6}$$

as functions of θ'.

$P(S_t|_1\mathbf{R}_T, {}_1\mathbf{X}_T, \theta)$ and $P(S_{t-1}, S_t|_1\mathbf{R}_T, {}_1\mathbf{X}_T, \theta)$ may be computed by using a recursive procedure known as the forward–backward algorithm (see Juang and Rabiner (1991) for details).

In the non-homogeneous case, the maximization of $\Psi(\theta_S'|\theta)$ always requires numerical optimization. Maximizing $\Psi(\theta_R'|\theta)$ has a simple closed form solution when model (2) is used for $P(\mathbf{R}_t|S_t)$, namely

$$\hat{p}_{si} = \sum_t \upsilon_t(s) r_t^i / \sum_t \upsilon_t(s).$$

When the more general formulation (1) is used, numerical optimization is required for $\Psi(\theta_R'|\theta)$ also.

When the rain-gauge stations are widely separated the conditional independence model (2) for $P(\mathbf{R}_t|S_t)$ will generally provide a good fit to the data. However, in a spatially dense network of stations, as shown in Fig. 1, it is likely that the autologistic model for $P(\mathbf{R}_t|S_t)$ will be required to capture local spatial dependences successfully. Then both the E-step and the M-step become computationally intractable as the number of stations, n, increases since the computation of the normalizing constant for the distribution $P(\mathbf{R}_t|S_t)$ requires summing over 2^n terms. To avoid the direct computation of this normalizing constant we used the method of Monte Carlo maximum likelihood (MCML) (Geyer and Thompson, 1992) as described below.

The autologistic model (1) may be written as

$$P(\mathbf{R}_t = \mathbf{r}|S_t = s; \eta) = \frac{1}{c(\eta)} \exp\{w(\mathbf{r})\eta^{\mathrm{T}}\}$$

where $\eta = (\alpha_{s1}, \ldots, \beta_{s12}, \ldots)$, $w(\mathbf{r}) = (r^1, r^2, \ldots, r^1 r^2, \ldots)$ and $c(\eta)$ is the normalizing constant of the distribution. Geyer and Thompson (1992) showed that for any η_0 in the parameter space

$$c(\eta) \approx \frac{c(\eta_0)}{N} \sum_{i=1}^{N} \exp\{w(\mathbf{r}_i)(\eta - \eta_0)^{\mathrm{T}}\}, \tag{7}$$

where $\mathbf{r}_1, \ldots, \mathbf{r}_N$ are samples from $P(\mathbf{R}_t|S_t; \eta_0)$, which can be generated by using the Gibbs sampler (Gelfand and Smith, 1990). Thus, if there is at least one η_0 in the parameter space for which the $c(\eta_0)$ can be computed, then expression (7) can be used to approximate the normalizing constant anywhere in the parameter space. For the autologistic model, this can be achieved by setting $\beta_{sij} = 0$ for all i and j so that $c(\eta_0) = \Pi_i \{1 + \exp(\alpha_{si})\}$.

The first and second moments of $\mathbf{R}_t$ (which are used to provide first derivatives for the numerical optimization routine in the M-step) may be approximated in a similar manner:

$$\begin{aligned}\mathbf{E}_\eta(R_t^k) &= \frac{\mathbf{E}_{\eta_0} r^k \exp\{w(\mathbf{r})(\eta - \eta_0)^{\mathrm{T}}\}}{\mathbf{E}_{\eta_0} \exp\{w(\mathbf{r})(\eta - \eta_0)^{\mathrm{T}}\}} \\ &\approx \sum_{i=1}^{N} r_i^k \exp\{w(\mathbf{r}_i)(\eta - \eta_0)^{\mathrm{T}}\} \Big/ \sum_{i=1}^{N} \exp\{w(\mathbf{r}_i)(\eta - \eta_0)^{\mathrm{T}}\}\end{aligned} \tag{8}$$

and

$$\mathbf{E}_\eta(R_t^k R_t^h) \approx \sum_{i=1}^{N} r_i^k r_i^h \exp\{w(\mathbf{r}_i)(\eta - \eta_0)^{\mathrm{T}}\} \Big/ \sum_{i=1}^{N} \exp\{w(\mathbf{r}_i)(\eta - \eta_0)^{\mathrm{T}}\}. \tag{9}$$

On the basis of these results we define a modified EM algorithm (which will be referred to as the EM–MCML algorithm) for estimation when the autologistic model is used for $P(\mathbf{R}_t|S_t)$:

(a) *E-step* — compute

$$\begin{aligned} \hat{v}_t(s) &= \hat{P}(S_t = s|_1\mathbf{R}_T, {}_1\mathbf{X}_T, \theta), \\ \hat{w}_t(s_1, s_2) &= \hat{P}(S_{t-1} = s_1, S_t = s_2|_1\mathbf{R}_T, {}_1\mathbf{X}_T, \theta) \end{aligned} \tag{10}$$

(b) *M-step* — maximize

$$\begin{aligned} \Psi(\theta_R'|\theta) &= \sum_{t,s} \hat{v}_t(s) \ln\{\hat{P}(\mathbf{R}_t|s, \theta_R')\}, \\ \Psi(\theta_S'|\theta) &= \sum_{t,s_1,s_2} \hat{w}_t(s_1 s_2) \ln\{P(S_t = s_2|S_{t-1} = s_1, X_t, \theta_S')\} \end{aligned} \tag{11}$$

as functions of θ'.

Here $\hat{P}$ indicates that the probability uses the estimated normalizing constant computed in approximation (7). Equations (8) and (9) are used in the M-step to compute first derivatives of $\Psi(\theta_R'|\theta)$.

To improve the efficiency of this approach η_0 and $c(\eta_0)$ in approximation (7) are updated at the beginning of each EM iteration with the values from the previous iteration. In addition, it is often advantageous to limit the parameter change in each EM iteration by limiting the number of Newton–Raphson iterates in the M-step. Such an algorithm remains self-consistent (Rai and Matthews (1993)) but will reduce the number of samples needed to update the normalizing constant and moments via MCML. Other computational issues and strategies were discussed in Geyer and Thompson (1992).

3.3. Model selection

In the present context, model selection means determining the number of weather states (the order) and atmospheric measures to include in the model. Order selection in a homogeneous Markov model is similiar to selecting the number of components in a mixture model, a problem which is best approached by using Bayes factors (Kass and Raftery, 1995). However, in the present application, a computation of the exact Bayes factor for each possible model is prohibitive. Therefore, we have investigated various approximations to the Bayes factor such as the Bayes information criterion

$$\text{BIC} = 2l - k \log(T) \tag{12}$$

where l is the log-likelihood, k is the number of model parameters and T is the number of days of data. Although this measure cannot be rigorously defended (since certain assumptions underlying the approximation are violated — see Kass and Raftery (1995) and Titterington (1990) for details), our experience has been that BIC is useful in that it identifies relatively parsimonious models which fit the data well.

We also consider physical interpretability of the identified weather states to be an important factor in model identification. The weather states can be visualized by first classifying each day into its most likely weather state according to the mode of the posterior distribution $P({}_1S_T|_1\mathbf{R}_T, {}_1\mathbf{X}_T)$ (a procedure known as the Viterbi algorithm — see Juang and Rabiner (1991)). It is then possible to identify the predominant patterns in precipitation and the atmospheric variables that are associated with each weather state by averaging these fields over all the days in a particular weather state (see Section 4 for an example). We prefer models in which these patterns are distinct and interpretable. Finally, since models which include more atmospheric information are likely to be more responsive to climate variability

Table 1. Comparison of several NHMMs using the conditional spatial independence model for $P(\mathbf{R}_t|S_t)$†

Number of states	*Covariates*	*Number of hidden parameters* (γ, μ)	*Number of output parameters* (α)	*BIC*
6	1, 4	90	180	36458
6	1, 4, 8	120	180	36475
7	1, 4	126	210	35751
7	1, 4, 8	168	210	36336

†The covariates are as follows: 1, mean sea-level pressure; 4, north–south gradient in sea-level pressure; 8, east–west gradient in geopotential height at 850 hPa. Hidden parameters refer to the parameters in equation (3); output parameters refer to the parameters in equation (1).

and change, we prefer models with more atmospheric variables and fewer weather states. This issue is discussed in greater detail in Section 5.

4. Results

Consultation with atmospheric scientists produced a list of 24 summary measures of the atmospheric data that might influence rainfall in this area (Charles *et al.*, 1996). These included measures such as the mean sea-level pressure and geopotential height over the region of interest and north–south and east–west gradients. Some preliminary analyses were conducted to assess the ability of each of these summary measures to predict rainfall occurrence. These analyses included simple procedures such as correlating each summary atmospheric measure with rainfall at each station, as well as more complex multivariate procedures such as using tree-based classification (Breiman *et al.*, 1984) to determine which measures best predicted rainfall occurrence patterns at a subset (stations 7, 9, 16 and 17) of the stations. Using the results of these preliminary analyses to provide a tentative ranking of the 24 atmospheric measures, a series of NHMMs were fitted varying the number of weather states and atmospheric measures in the model. The conditional spatial independence model (2) was used during this preliminary model fitting stage. Since the EM–MCML procedure is computationally demanding, the need for the general autologistic model was not evaluated until after the number of weather states and the atmospheric measures had been selected.

In the present application the best models by objective criteria (BIC) had either six or seven weather states and two or three atmospheric measures (Table 1). For each candidate model we classified each day into its most likely weather state and compared the precipitation occurrence patterns associated with each weather state with the corresponding sea-level pressure and geopotential height fields. This comparison suggested that the six-state NHMM had a high degree of physical realism. In addition, the patterns associated with the six-state models were distinct whereas the patterns for two of the states in the seven-state models were almost indistinguishable. For this reason, and because the seven-state model did not noticeably improve the fit to the data by the measures examined below, a six-state model was chosen. Since the BIC values for the two (six-state) models with and without the geopotential height covariate were similar, we chose to include this covariate. Our final model, therefore, included six weather states and three atmospheric measures (mean sea-level pressure MSLP, north–south gradient in sea-level pressure and the east–west gradient in 850 hPa geopotential height, GPH850) (Table 1).

Figs 2(a) and 3 illustrate the fit of this model to important observed rainfall statistics from

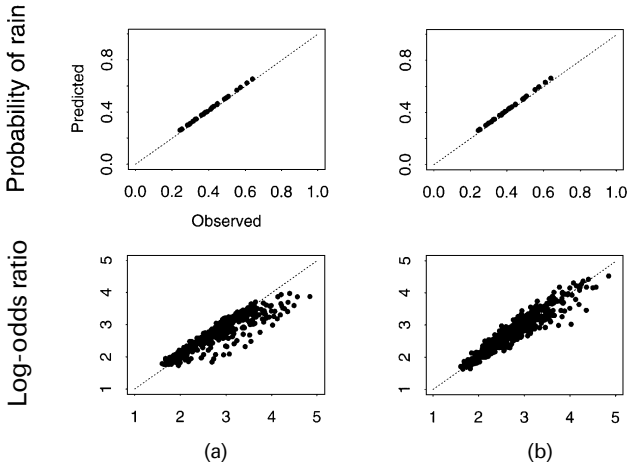

Fig. 2. Comparison of observed and model-predicted rainfall statistics for the 10 years of data used for model fitting (the model-predicted statistics are generated by simulation from the fitted model using the observed atmospheric data; the observed statistics are on the *x*-axis and the model-predicted statistics are on the *y*-axis) (a) conditional independence model; (b) spatial (EM–MCML) model

the 10-year period used for model fitting. Fig. 2(a) shows first and second moments from the conditional independence model whereas Fig. 3 shows the distribution of 'storm durations' at selected stations (storm duration is defined as the number of consecutive days of rain; the model-based statistics were computed by generating multiple simulations from the model, conditionally on the observed atmospheric data, and then averaging over the simulations so that variability in the predicted quantities is negligible). Although we expect the model to provide a good fit to the average precipitation frequency, the distributions of wet and dry spell lengths have proven to be the most difficult characteristic of rainfall to reproduce using weather state models (see, for example, Wilby *et al.* (1994)). However, these are key features in hydrologic models as they affect the likelihood of floods and droughts. It is clear that the fitted conditional independence model does well in reproducing the observed probability of rainfall at each station and the distribution of storm durations. The distributions of interarrival times (not shown) are also modelled well. The model also captures the spatial correlation between stations that is induced by the common weather states, which represents a substantial proportion of the observed dependence. However, additional correlation between nearby stations is created by local orographic and other 'sub-weather-state' scale effects and this source of correlation is not captured in the independence model for $P(\mathbf{R}_t|S_t)$.

Since correct modelling of the local distribution of rainfall may be important in many hydrologic models (e.g. flood simulation) we sought further insight into these local effects. Empirical estimates of the pairwise (unconditional) log-odds ratios were plotted against the distance and direction between the stations for each weather state (based on all days classified in a given weather state). These plots suggested that the within-state spatial correlation varied elliptically with direction and declined linearly with the log-distance between stations.

To include these local effects in the NHMM the EM–MCML procedure was used to fit a model in which the conditional log-odds ratios β_{sij} (see equation (1)) were modelled as a function of the distance and direction between the stations. The following functional form was found to give a good fit to the empirical log-odds ratios and was, therefore, adopted as a model for the conditional log-odds ratios:

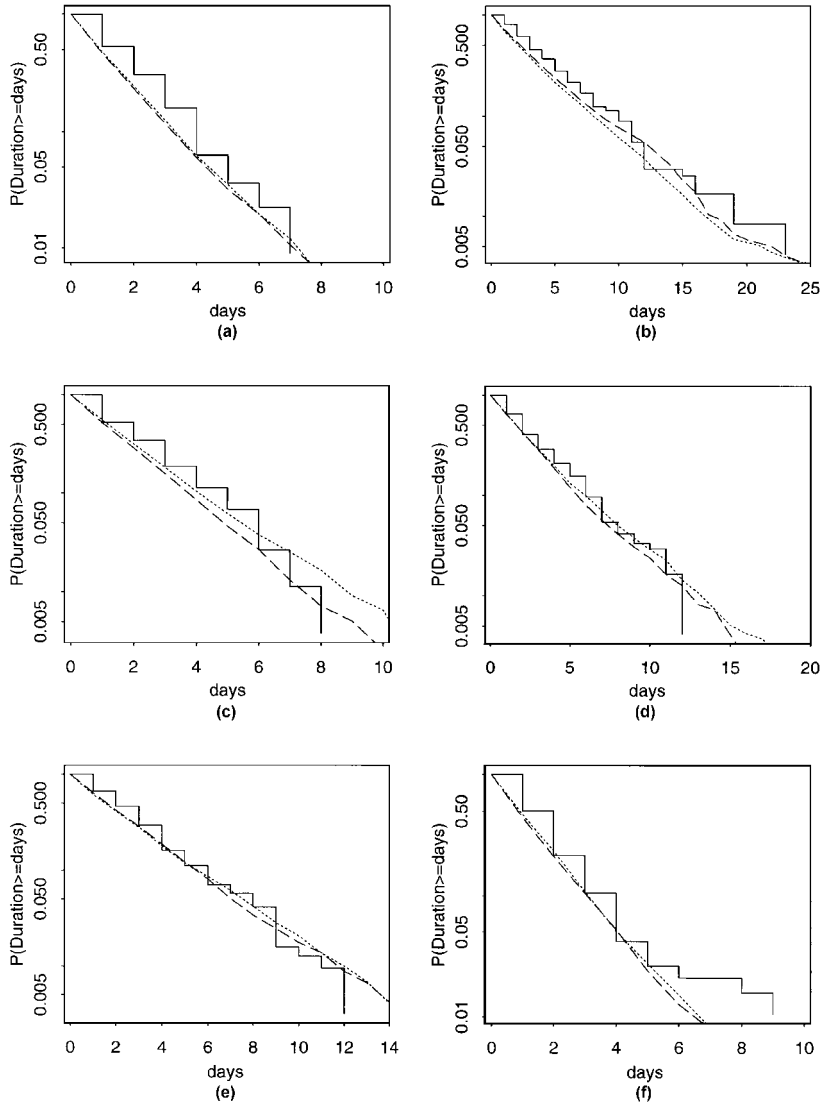

Fig. 3. Comparison of observed (———) and model-predicted (·····, independence; — — —, EM–MCML) rainfall statistics based on the 10 years of data used for model fitting—duration distribution: (a) station 2; (b) station 9; (c) station 29; (d) station 13; (e) station 16; (f) station 19

$$\beta_{sij} = b_{0s} + b_{1s} \log[d_{ij}\sqrt{\{\cos(\phi_s + h_{ij})^2 + \sin(\phi_s + h_{ij})^2/e_s\}}] \tag{13}$$

where d_{ij} and h_{ij} are respectively the distance and direction between stations i and j. For each state s, there are four parameters in this model. Although, theoretically, all four parameters could be estimated by the methods outlined in Section 3.2, an estimation of the non-linear parameters ϕ_s and e_s slows the computations substantially. Therefore, these parameters were fixed at the values obtained from a non-linear least squares regression analysis of the empirical log-odds ratios. The bs were then estimated by using the procedure described in Section 3.2. The resulting model will be referred to as the 'spatial model'.

This approach significantly improved the fit of the model to the empirical log-odds ratios, as seen in Fig. 2(b). The EM–MCML algorithm converged to a model with maximized log-likelihood 1472 units greater than under the conditional independence model at the expense of 24 additional β-parameters.

The ability of the NHMM to reproduce key precipitation statistics conditionally on the observed atmospheric data suggests that this model could be useful for generating conditional rainfall simulations for the period 1978–1987. However, if the model is to be used to generate precipitation simulations for other periods or alternative atmospheric data sets (e.g. to investigate the effects of climate variability or climate change) then it is important to evaluate the model on reserved data. Fig. 4 compares various observed rainfall statistics with those predicted by the model for the 5 years of reserved data (1988–1992). Results for the spatial model are shown. Fig. 4 shows increased variability when the model is applied to reserved data (as expected) but no systematic biases. This latter point is important since a small but measurable shift in the mean atmospheric data fields occurred during the 5-year period of reserved data compared with the 10-year period used for model fitting (-0.81 hPa in MSLP, 0.47 hPa in north–south sea-level pressure gradient, 0.45 m in east–west GPH850). If this shift is deliberately removed from the atmospheric data (e.g. by subtracting the mean shift between the 5- and 10-year periods from the atmospheric measures in the 5-year period) then a small but noticeable downward bias is observed in the predicted rainfall probabilities (averaging 3.1 percentage points over the 30 stations). However, when the atmospheric data are correctly included in the model the rainfall bias is essentially eliminated. In other words, the lack of bias seen in Fig. 4 indicates that the model could adjust the rainfall probabilities to account for the (slight) non-stationary shift in the atmospheric data.

Although the weather states are abstract constructs of the model, they can be examined by first classifying each day into its most likely state and then averaging, at each grid node, the values of sea-level pressure, geopotential height or other atmospheric measures over all days in a given state. The resulting 'composite' field can be contoured to give a visual representation of the average field in that state and can provide a means of assessing the physical realism of the hidden states as they are analogous to traditional synoptic classifications used by meteorologists and climatologists (e.g. Yarnal (1993)).

Fig. 5 shows the rainfall probabilities and composite mean sea-level pressure and 850 hPa geopotential height fields associated with three of the six hidden states. The synoptic pattern associated with state 2 (high rainfall probabilities at all stations) is a typical winter pattern, indicating a strong cold front traversing the study region. The strong and widespread rains associated with this pattern produce the majority of the run-off within the water-supply catchments of the region (Gentilli, 1972). This state occurred 20% of the time, with little change from the 10-year to the 5-year period.

In contrast, the synoptic pattern associated with state 5 (low probability of rain at all stations) shows a dominant high pressure system centred in the Great Australian Bight. The intensity and rate of movement of such systems controls the weather experienced across the entire study region (Sturman and Tapper, 1996). Their progression eastwards typically follows a well-defined period of 5–7 days. However, high pressure systems can remain stationary in the Bight for longer periods, blocking the general circulation and leading to prolonged dry periods (Southern, 1979). This pattern is the single most common weather state, occurring 29% of the time, and is also the most persistent state with a mean duration of 2.7 days.

The remaining four states are characterized by rainfall in particular regions of the study area. For example, state 4 (Fig. 5(b)) exhibits a high probability of rain at the south-west stations but a low probability in the north and western stations. The frontal systems

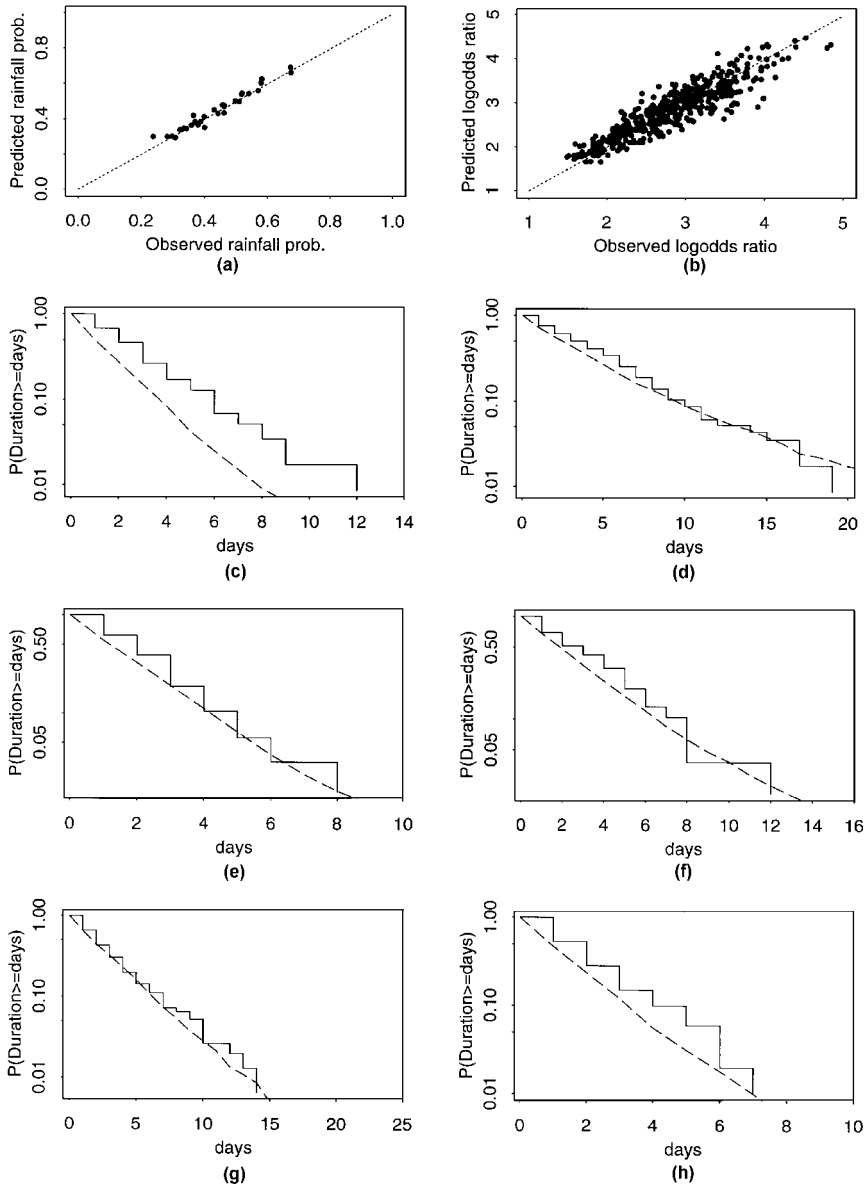

Fig. 4. Comparison of observed (———) and model-predicted (— — —, EM–MCML) rainfall statistics on the 5 years of reserved data (the model-predicted statistics are generated by simulation from the fitted model using the observed atmospheric data; duration distributions are shown at a representative subset of stations; station 2 represents the poorest fit seen at any station): (a) rainfall probability; (b) log-odds ratio; (c) station 2; (d) station 9; (e) station 29; (f) station 13; (g) station 16; (h) station 19

associated with this synoptic pattern are weaker than those associated with state 2. They do not penetrate as far north or inland, with little rainfall falling east of the Darling Range.

The interpretability of the mean sea-level pressure patterns suggests that the NHMM weather states have a high degree of physical realism. Also, as the patterns seen in the mean sea-level pressure plots are complemented by those seen in the 850 hPa geopotential height

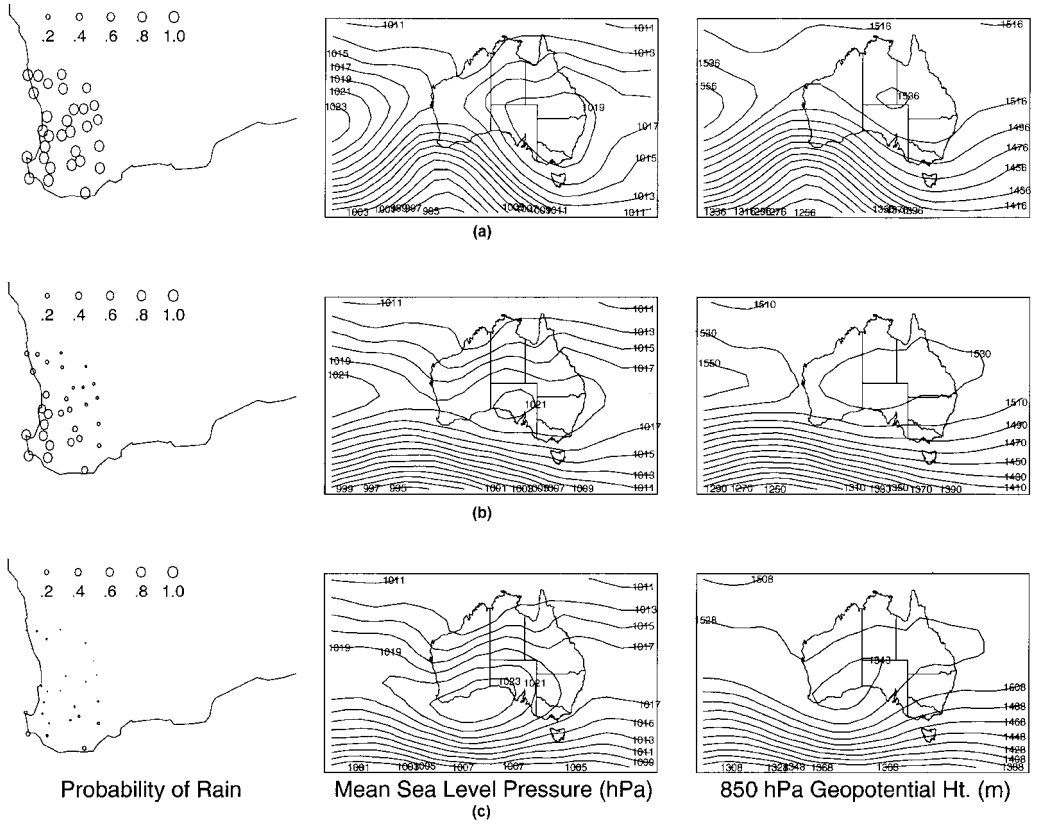

Fig. 5. Probability of rain, composite sea-level pressure and 850 hPa geopotential height fields for three states from the six-state spatial model estimated by using the EM–MCML algorithm (each day is first classified into its most likely state by using the Viterbi algorithm; all days in a particular state are then averaged at each station (for rainfall) or grid node (for the atmospheric variables) to obtain the composite fields): (a) state 2; (b) state 4; (c) state 5

field plots, we believe that the model has identified the dominant synoptic scale features of precipitation in this region. This supports our hypothesis that the model successfully downscales atmospheric circulation to the multisite precipitation occurrence process.

5. Discussion

NHMMs can provide hydrologists and atmospheric scientists with a useful tool for generating realistic simulations of precipitation and understanding the relationships between atmospheric circulation patterns and rainfall. This approach to precipitation modelling will be most successful in areas and/or seasons where precipitation is driven by synoptic scale systems. It is unlikely that these models will be successful in areas or seasons in which rainfall is driven primarily by convective activity (e.g. thunder-storms) since these processes evolve on relatively small scales and may not be predictable from synoptic scale circulation data.

NHMMs generalize the concept of a weather state model as described by Hay *et al.* (1991), Bardossy and Plate (1992), Kidson (1994) and others. In these models, however, the investigators explicitly defined the weather states. The resulting states, although reflecting

meteorological intuition, may not be optimal for modelling rainfall. An important advantage of the NHMM approach is that it allows us to combine meteorological intuition (through selection of the atmospheric measures) with modelling to define weather states that are 'optimal' in the sense of separating precipitation occurrence patterns. Plots such as Fig. 5 can provide insight into the interpretation of the weather states and the relationship between atmospheric circulation patterns and precipitation.

Another important distinction of the NHMM approach is the use of the Markov assumption in the definition of the weather states. This is both a strength and a weakness of the model. Although it is conceptually appealing to assume that the current weather state (and, therefore, the current rainfall pattern) should depend only on current atmospheric conditions, practical aspects of the data collection may compromise such an assumption. The atmospheric data are typically measured at a point in time whereas the rainfall measurements represent an accumulation over a 24-h period. We believe that conditioning on the previous day's weather state helps to recover some of the (unmeasured) atmospheric information that is relevant to the 24-h precipitation period. In all the examples that we have studied, conditioning on the previous day's weather state significantly improves the fit of the model to observed rainfall statistics, particularly the observed duration distribution. The danger of this assumption, however, as noted by one reviewer, is that the previous day's weather state can serve as an 'omitted covariate' which will not respond to non-stationary shifts in climate (e.g. if the model is used under an altered climate). Indeed, we have observed some trade-off between the number of weather states identified and the number of atmospheric variables included in the model — models with fewer weather states achieve a minimum BIC with more atmospheric variables whereas models with more weather states achieve a minimum BIC with fewer atmospheric variables. Since we would expect that a model with more atmospheric information will produce precipitation simulations which are more responsive to shifts in atmospheric conditions, we favour models with fewer weather states and more atmospheric variables.

Several methodological issues remain. Variances of the estimated parameters can be approximated by using the method devised by Hughes (1997) although this may be numerically difficult for large models. Also, estimation using the EM–MCML method is computationally intensive and simpler approaches to fitting a spatial model, even if only approximate, would be desirable. We have investigated the use of maximum pseudolikelihood (Besag, 1975) in the M-step (equation (11)) but have found that this can lead to nonsensical weather states (Hughes *et al.*, 1996).

Although the model described here is based on rainfall occurrences, the NHMM framework is theoretically sufficiently general to model amounts also. Indeed, for many applications a model for rainfall amounts is necessary. If the amounts could be considered conditionally spatially independent, given the occurrences, then we could include a separate amounts model at each station. However, if there is significant spatial correlation in the amounts even after the spatial correlation in the occurrences is accounted for, then the amounts must be modelled jointly. This requires the specification of a multivariate mixed discrete–continuous model for $P(\mathbf{R}_t|S_t)$. Research in this area is ongoing.

Another methodological issue is to develop methods to simulate rainfall occurrence at locations that have not been explicitly included in the model. In the context of the autologistic model this can be accomplished by spatially interpolating the α_{si}- and the β_{sij}-parameters to a new location i'. For the example presented in Section 4 we observed that the α_{si} from the best-fitting autologistic model were small and showed little within-weather-state variation in the interior of the network. Therefore, $\alpha_{si'}$ could be estimated as the mean value of α_{si} from other stations and $\beta_{sii'}$ could be computed using equation (13).

Software (written in Fortran with calls to Numerical Algorithms Group library routines) to implement the methods discussed in this paper is available from

`http://www.blackwellpublishers.co.uk/rss/`

Acknowledgements

Although the research described in this paper has been funded in part by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. In addition, partial support was provided by National Science Foundation grant DMS-9524770 and by a University of Washington Royalty Research Fund award to the first author. Portions of this work were performed while the first author was Visiting Scientist at the Commonwealth Scientific and Industrial Research Organisation, Perth, Australia. The Australian Bureau of Meteorology kindly provided the atmospheric and rainfall data for this study. The authors are grateful to Bryson Bates, Mick Fleming and Tom Lyons for helpful discussions.

References

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

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**, 164–171.

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

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Belmont: Wadsworth.

Charles, S. P., Hughes, J. P., Bates, B. C. and Lyons, T. J. (1996) Assessing downscaling models for atmospheric circulation—local precipitation linkage. In *Proc. Int. Conf. Water Resources and Environmental Research: Towards the 21st Century, Kyoto*, vol. 1, pp. 269–276. Kyoto: Kyoto University.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Gabriel, K. R. and Neumann, J. (1962) A Markov chain model for daily rainfall occurrences at Tel-Aviv. *Q. J. R. Meteorol. Soc.*, **88**, 85–90.

Gelfand, A. E. and Smith, A. F. M (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gentilli, J. (1972) *Australian Climate Patterns*. Melbourne: Nelson.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B, **54**, 657–699.

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

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

Hughes, J. P. (1997) Computing the observed information in the hidden Markov model using the EM algorithm. *Probab. Statist. Lett.*, **32**, 107–114.

Hughes, J. P. and Guttorp, P. (1994) A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Wat. Resour. Res.*, **30**, 1535–1546.

Hughes, J. P., Guttorp, P. and Charles, S. P. (1996) A nonhomogeneous hidden Markov model for precipitation. *Technical Report 316*. Department of Statistics, University of Washington, Seattle.

Intergovernmental Panel on Climate Change (1995) *Climate Change 1995, the Science of Climate Change* (eds J. T. Houghton, L. G. Meira Filho, B. A. Callander, N. Harris, A. Kattenberg and K. Maskell). Cambridge: Cambridge University Press.

Juang, B. H. and Rabiner, L. R. (1991) Hidden Markov models for speech recognition. *Technometrics*, **33**, 251–272.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.

Kavvas, M. L. and Delleur, J. W. (1981) A stochastic cluster model for daily rainfall sequences. *Wat. Resour. Res.*, **17**, 1151–1160.

Kidson, J. W. (1994) The relation of New Zealand daily and monthly weather patterns to synoptic weather types. *Int. J. Climatol.*, **14**, 723–737.

LeCam, L. (1961) A stochastic theory of precipitation. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*. Berkeley: University of California Press.

Rai, S. N. and Matthews, D. E. (1993) Improving the EM algorithm. *Biometrics*, **49**, 587–591.

Southern, R. L. (1979) The atmosphere. In *Environment and Science* (ed. B. J. O'Brien), pp. 183–226. Nedlands: University of Western Australia Press.

Stern, R. D. and Coe, R. (1984) A model fitting analysis of daily rainfall data (with discussion). *J. R. Statist. Soc.* A, **147**, 1–34.

Sturman, A. and Tapper, N. (1996) *The Weather and Climate of Australia and New Zealand*. Oxford: Oxford University Press.

Titterington, D. M. (1990) Some recent research in the analysis of mixture distributions. *Statistics*, **21**, 619–641.

Waymire, E. and Gupta, V. K. (1981) The mathematical structure of rainfall representations: 2, A review of the theory of point processes. *Wat. Resour. Res.*, **17**, 1273–1285.

Wilby, R. L., Greenfield, B. and Glenny, C. (1994) A coupled synoptic-hydrological model for climate change impact assessment. *J. Hydrol.*, **153**, 265–290.

Yarnal, B. (1993) Synoptic climatology in environmental analysis. London: Blehaven.