A Nonhomogeneous Hidden Markov Model for Precipitation

S. P. Charles

James P. Hughes

Peter Guttorp

NRCSE

Technical Report Series

NRCSE-TRS No. 004

A Nonhomogeneous Hidden Markov Model for Precipitation

James P. Hughes[1]
Peter Guttorp[2]
Stephen P. Charles[3]

September 22, 1997

[1]Dept. of Biostatistics 357232, University of Washington, Seattle, WA 98195 USA

[2]Dept. of Statistics 354322, University of Washington, Seattle, WA 98195 USA

[3]CSIRO Division of Water Resources, Private Bag, P.O. Wembley, Western Australia 6014

Address correspondence to James P. Hughes (email: hughes@biostat.washington.edu)

Abstract

A stochastic model for relating precipitation occurrences at multiple rain gauge stations to broad-scale atmospheric circulation patterns (the so-called "downscaling problem") is proposed. The model is an example of a nonhomogeneous hidden Markov model and generalizes existing downscaling models in the literature. The model assumes that atmospheric circulation can be classified into a small number of (unobserved) discrete patterns (called "weather states"). The weather states are assumed to follow a Markov chain in which the transition probabilities depend on observable characteristics of the atmosphere (e.g. mean sea-level pressure). Precipitation is assumed to be conditionally temporally independent given the weather state. An autologistic model for multivariate binary data is used to model rainfall occurrences and capture local spatial dependencies. A modified EM algorithm based on Markov chain maximum likelihood procedures is developed for estimation.

This approach is used to model a 15 year sequence of winter data from 30 rain stations in southwestern 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 is able to accurately reproduce the observed rainfall statistics in the reserved data, even in the face of a non-stationary 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. We discuss the role that models such as this might play in assessing the impact of climate change.

Keywords: hidden Markov model, climate change, precipitation model, Monte Carlo maximum likelihood, EM algorithm

1 Introduction

Stochastic models for precipitation have long been used to aid in understanding the probabilistic structure of rainfall and for simulation studies. In particular, precipitation simulations are often used as input into hydrologic models of flooding, runoff, water supply, agricultural models of crop growth, and other applications. In the past these models considered only the rainfall process, without reference to the atmospheric processes that drive precipitation. In part, this reflected the absence of good, long-term records of atmospheric circulation. Thus, Gabriel and Neuman (1962) used a Markov chain with homogeneous transition matrix to model daily wet/dry occurrences at a single rain gauge station in Israel. Stern and Coe (1984) extended this model by making the (logits of) transition probabilities a Fourier series to represent seasonal variations. Others developed more mechanistic models. For example, LeCam (1961) described rainfall using a cluster point process whereby cyclonic storms were assumed to contain "bands" (areas of high rainfall intensity) and the bands contained rain cells where precipitation activity occurs. Waymire and Gupta (1981), Kavvas and Delleur (1975, 1981) and others expanded on the point process approach.

These models have several limitations, however. In developing hydrologic models researchers 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. In addition, 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 nonstationary, 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 would not be useful in studies of climate variability or climate change.

Over the past few decades advances in data gathering and our understanding of atmospheric circulation have lead 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 circulations — general circulation models (GCMs) — has lead to the creation of realistic simulations of atmospheric circulation of essentially unlimited duration (some background on GCMs can be found in IPCC (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" 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 the sense that any nonstationarity 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 general circulation models 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 GCM's is constrained by both computational considerations as well as our understanding of atmospheric dynamics to scales of approximately 2° to 5° of longitude and latitude. Precipitation, however, varies on much more local scales. For this reason, GCMs have been unable to generate realistic simulations of rainfall (Giorgi and Mearns, 1991). Weather state models provide one solution to this so-called downscaling problem. Using the GCM atmospheric simulations as input, a weather state model can be used to generate realistic simulations of local precipitation.

A final, more speculative, application of weather state models is to investigate the effect of hypothesized climate changes on precipitation. One such effect of particular interest is the theory (popularly termed the "greenhouse effect") that observed increases (along with predicted future increases) in atmospheric CO_2 will lead to a global rise in temperature. These predictions are based on experiments with GCMs in which the model is run with an increased (typically doubled) atmospheric concentration of CO_2. Under the strong assumption that the historical relationship between precipitation and large-scale circulation would still apply, a weather state model could be used to access the impact of the altered climate on precipitation.

Hughes and Guttorp (1994) describe a class of models, which they term non-homogeneous hidden Markov models (NHMM), that can be used to model the relationship between atmospheric circulation and precipitation and to generate conditional simulations of precipitation. In a basic hidden Markov model (HMM), one assumes the existence of two processes — an observed process and a hidden process. The observed process (such as rain occurrence at a fixed set of stations) is assumed to be conditionally temporally independent given the hidden process; the hidden process is assumed to evolve according to a first order Markov chain (see Juang and Rabiner (1991) for a review of hidden Markov models). A nonhomogeneous hidden Markov model (NHMM) extends this idea by allowing the transition matrix of the hidden states to depend on a set of observed covariates. In the present application the covariates are derived from the atmospheric data. This approach provides a general framework for the development of weather state models, since Hughes and Guttorp (1994) show that most existing weather state models can be written as special cases of the NHMM.

In this article we develop a model for precipitation at 30 rain gauge stations in southwestern Australia. The data are described in section 2. In section 3, we describe the model and an estimation procedure (a modification of the EM algorithm) that is computationally feasible for spatially dense networks of stations. Section 4 provides the results of our data analysis. In particular, we show that the proposed model is able to detect shifts in precipitation frequency that result from changes in circulation patterns. In section 5 we discuss future directions for research on downscaling models and the NHMM as tools for precipitation modelling and climate change impact assessment.

2 Data

A fifteen year record (1978–1992) of daily winter (May–October) rainfall occurrences (2760 days, total) at 30 stations in southwestern Australia was made available by the Australian Bureau of Meteorology. The locations of the stations are shown in figure 1. Each rainfall value represents the total rainfall over a 24 hour period ending at 0900 (local standard time). Atmospheric data were obtained from the Australian Bureau of Meteorology on a Lambert conformal grid and interpolated to a rectangular grid of similar scale—2.25^o latitude by 3.75^o longitude (also shown in figure 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 1900 (local standard time) on the preceeding day. 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

Our goal is to develop a model which will identify and quantify relationships between the observed synoptic (large-scale) atmospheric measures and local precipitation patterns. We postulate the existence of an unobserved discrete valued process — the "weather state" — which acts as a link between the two disparate scales. Formally, let $\mathbf{R}_t$ be a multivariate vector giving rainfall amounts or occurrences at a network of sites at time t, S_t be the weather state at time t, and $\mathbf{X}_t$ the vector of atmospheric measures at time t for $1 \leq t \leq T$. 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 $\mathbf{X}_1^T$ will be used to indicate the sequence of atmospheric data from time 1 to T and similarly for $\mathbf{R}_1^T$ and S_1^T. Lower case will be used to indicate realized values of random variables (i.e. $P(\mathbf{R}_t = \mathbf{r})$). All vectors are row vectors and all vectors and matrices will be written in bold type.

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

$$\begin{aligned} &(M1) \qquad P(\mathbf{R}_t \mid S_1^T, \mathbf{R}_1^{t-1}, \mathbf{X}_1^T) = P(\mathbf{R}_t \mid S_t) \\ &(M2) \qquad P(S_t \mid S_1^{t-1}, \mathbf{X}_1^T) = P(S_t \mid S_{t-1}, \mathbf{X}_t) \end{aligned}$$

and $P(S_1 \mid \mathbf{X}_1^T) = P(S_1 \mid \mathbf{X}_1)$. Specific NHMM's are defined by parameterizing $P(\mathbf{R}_t \mid S_t)$ and $P(S_t \mid S_{t-1}, \mathbf{X}_t)$ as discussed below.

The first assumption (M1) states that the rainfall 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 (M2). 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. In the absence of the atmospheric data this is simply the Markov assumption applied to the hidden process. The atmospheric data, when included, are used to modify the transition probabilities of the Markov process — hence the term "nonhomogeneous". Most weather state models 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 $P(S_t \mid S_{t-1}, \mathbf{X}_t)$ to be degenerate.

Various parameterizations for $P(\mathbf{R}_t \mid S_t)$ are possible. In the present application a model for rainfall occurrence (rainfall below/above 0.3 mm) is developed; approaches for modelling amounts are discussed in section 5.

For an n-station network, let $\mathbf{R}_t = \{R_t^1, \ldots, R_t^n\}$ with observed value of $\mathbf{r}_t = \{r_t^1 \ldots, r_t^n\}$. Let $r_t^i = 1$ if rain occurs on day t at station i and 0 otherwise. The autologistic model for multivariate binary data is defined as

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

where both α_{si} and β_{sij} must be finite and $\beta_{sii} = 0$. 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 $P(r^i, r^j \mid \mathbf{r}^{-\mathbf{i},-\mathbf{j}}, S_t = s)$. When β_{sij} is positive, stations i and j are positively associated (within weather state s) while 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 important special case of (1) arises when $\beta_{sij} = 0$ for all i, j, and s. Then

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

where $p_{si} = \exp(\alpha_{si})/(1+\exp(\alpha_{si}))$. This will be referred to as the "conditional independence model" for $P(\mathbf{R}_t \mid S_t = s)$. The p_{si} give the probability of rain at station i in weather state s. 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 due to the influence of the common weather state). Hughes and Guttorp (1994) present an example of a spatially dispersed network of rain gauge stations for which the conditional independence model works well.

The parameterization for $P(S_t \mid S_{t-1}, \mathbf{X}_t)$ is motivated by Bayes formula and uses the normal kernel for the joint distribution of the atmospheric data:

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

where μ_{ij} is the mean of $\mathbf{X}_t$ and $\mathbf{V}$ is the corresponding covariance matrix. This model shows clearly how the NHMM is a general version of the simpler HMM. The γ_{ij} may be thought of as the baseline transition matrix of the weather state process and corresponds to the transition matrix of an HMM. The exponential term quantifies the effect of the atmospheric data on the baseline transition matrix. To ensure identifiability of the parameters, the constraints $\sum_j \gamma_{ij} = 1$ and $\sum_j \mu_{ij} = \mathbf{0}$ are imposed. In this formulation $\mathbf{V}$ is used for scaling and numerical stability (typically, $\mathbf{V}$ is set equal to the raw covariance matrix of the atmospheric variables). It is not estimated as part of the model.

3.2 Parameter Estimation

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

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

which appears to be computationally intractable, even for a short sequence of data. However, the forward-backward procedure, a recursive algorithm developed to solve the standard hidden Markov model (e.g. Juang and Rabiner, 1991) can be extended to the NHMM and makes the calculation possible. The basic idea is to successively pass each of the multiple summations in the likelihood 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 T'th term. If one has several independent sequences of data (for instance, multiple years of data) then the likelihoods for each sequence are multiplied together to form the overall likelihood.

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, S_1^T, to be “missing” data. This same approach may be used with the NHMM. Let $\theta = (\theta_R, \theta_S)$, the parameters of the observed and hidden processes, respectively. Then, the EM algorithm for the NHMM may be written as (see Juang and Rabiner, 1991, for details)

- E step: compute

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

- M step: maximize

  $$\Psi(\theta'_R \mid \theta) = \sum_{t,s} v_t(s) \ln P(\mathbf{R}_t \mid s, \theta'_R)$$

  and

  $$\Psi(\theta'_S \mid \theta) = \sum_{t,s_1,s_2} w_t(s_1 s_2) \ln P(S_t = s_2 \mid S_{t-1} = s_1, \mathbf{X}_t, \theta'_S) \tag{6}$$

  as functions of θ'.

In the nonhomogeneous case, maximization of $\Psi(\theta'_S \mid \theta)$ always requires numerical optimization. Maximizing $\Psi(\theta'_R \mid \theta)$ has a simple closed form solution when model (2) is used for $P(\mathbf{R}_t \mid S_t)$, namely, $\hat{p}_{si} = \sum_t v_t(s) r_t^i / \sum_t v_t(s)$. When the more general formulation (1) is used, however, numerical optimization is required for $\Psi(\theta'_R \mid \theta)$ also and both the E-step and the M-step become computationally intractable as the number of stations, n, increases. In the E-step $P(\mathbf{R}_t \mid S_t)$ is used to compute the weights $v(s)$ and $w(s_1, s_2)$ so the normalizing constant of this distribution (which requires summing over 2^n terms) is needed; in the M-step both the the normalizing constant as well as the first and second moments of $\mathbf{R}_t$ are needed (the moments are used to compute the derivatives of $\Psi(\theta'_R \mid \theta)$ and $\Psi(\theta'_S \mid \theta)$; numerical optimization techniques are more efficient if first derivatives are provided). To address these difficulties, a modified EM algorithm was developed using the method of Monte Carlo maximum likelihood (MCML) (Geyer and Thompson, 1992). This modified EM algorithm is described below.

The autologistic model (1) may be written as

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

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) show that if some η_0 is in the parameter space then

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

where $\mathbf{r}_1, \ldots, \mathbf{r}_N$ are samples from $P(\mathbf{R}_t \mid S_t; \eta_0)$. Thus, if there is at least one η in the parameter space, say η_0, for which the normalizing constant $c(\eta_0)$ can be computed, then (7) can be used to approximate the normalizing constant anywhere in the parameter space. For the autologistic model, this can be achieved by setting $\beta_{sij} = 0$ where $c(\eta) = \prod_i (1 + \exp(\alpha_{si}))$.

The first and second moments of $\mathbf{R}_t$ may be approximated in a similar manner:

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

and

$$\mathbf{E}_\eta(\mathbf{R}_t^k \mathbf{R}_t^h) \approx \frac{\sum_{i=1}^N r_i^k r_i^h \exp(w(\mathbf{r}_i)(\eta - \eta_0)^T)}{\sum_{i=1}^N \exp(w(\mathbf{r}_i)(\eta - \eta_0)^T)} \tag{9}$$

We used these results to develop a modified EM algorithm (which will be referred to as EM/MCML) which may be summarized as follows:

- E step: compute

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

- M step: maximize (or partly maximize)

$$\Psi(\theta'_R \mid \theta) = \sum_{t,s} \hat{v}_t(s) \ln \hat{P}(\mathbf{R}_t \mid s, \theta'_R)$$

and

$$\Psi(\theta'_S \mid \theta) = \sum_{t,s_1,s_2} \hat{w}_t(s_1 s_2) \ln \hat{P}(S_t = s_2 \mid S_{t-1} = s_1, X_t, \theta'_S) \tag{11}$$

as functions of θ'.

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

To improve the efficiency of this approach we update η_0 and $c(\eta_0)$ in 7 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 are dicussed in Geyer and Thompson (1992).

4 Results

Consultation with atmospheric scientists produced a list of 24 summary measures of the atmospheric data that might influence rainfall in this area. These included measures such as mean sea-level pressure and geopotential height over the region of interest, north–south and east–west gradients, etc. Some preliminary analyses were conducted to get a rough idea of the ability of each of these summary measures to predict rainfall. 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 summary atmospheric 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 NHMM's were fit to the first 10 (of 15) years of data. Both the number of weather states and the atmospheric measures included in the model were varied. For computational considerations, the conditional spatial independence model (equation 2) was used during this preliminary model fitting stage; the EM/MCML procedure was used to fit the general autologistic model after the number of weather states and atmospheric measures had been selected.

Selection of a final model was somewhat subjective, in part because existing procedures for model selection cannot be justified theoretically for hidden Markov models (for instance, order selection in HMM's inherits many of the difficulties associated with selecting the number of components in a mixture model - see Titterington xxxx for a review). As a rough guide, we have found the Bayes information criterion (BIC) (see, for example, Kass and Raftery, 1995) useful for identifying the best fitting models but we do not rely solely on this measure. Interpretability of the weather states is also an important consideration. In the present example the "best" models (i.e. lowest BIC) had either 6 or 7 weather states and 2 or 3 atmospheric measures (table 1). Comparison of the precipitation occurrence patterns associated with each weather state with their corresponding composite MSLP and GPH850 fields (see figure 5 for an example) suggested that the six state NHMM had a high degree of physical realism. In addition, the patterns associated with the six states were distinct. In contrast, 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 noticably improve the fit of the the model by the measures we examine below, a six state model was chosen. Finally, for reasons discussed further in section 5 we prefer models with fewer weather states and more atmospheric variables. Since the BIC's for the two 6 weather state models with and without the geopotential height covariate were similar, we selected the model with more atmospheric measures. 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)) and had a log-likelihood of -17110 (table 1).

Table 1: Comparison of several nonhomogeneous hidden Markov models using the conditional spatial independence model for $P(\mathbf{R}_t \mid S_t)$. Covariates are 1 = mean sea-level pressure; 2 = Mean geopotential height at 500mb; 4 = N-S gradient in sea level pressure; 8 = E-W gradient in geopotential height at 850mb.

no. states	covariates	log-likelihood	df	BIC
6	1,4	17214	270	36458
6	1,4,8	17110	300	36475
7	1,4	16876	336	35751
7	1,4,8	16747	378	36336

Figures 2 and 3 illustrate the fit of this model to important observed rainfall statistics, including first and second moments, and the distribution of storm lengths (defined as the number of consecutive days of rain; the model-based statistics are computed by generating multiple simulations from the model, conditional on the observed atmospheric data, and then averaging over the simulations so that variability in the predicted quantities is negligible). The distribution of storm lengths is of particular interest to hydrologists because storm duration strongly influences flood magnitude and frequency. This distribution has also proven to be the most difficult characteristic of rainfall to reproduce using weather state models.

From these figures it is clear that the fitted model (which assumes conditional spatial independence) does well in reproducing the observed probability of rainfall at each station and the distribution of "storm durations" (number of consecutive days with rain). However, this model does less well at reproducing the observed patterns of spatial correlation between stations, particularly for stations that are highly correlated. This makes sense: most of the spatial correlation between stations is induced by the common weather state and this source of correlation is captured by the model. 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 $P(\mathbf{R}_t \mid S_t)$. We note that none of the other models shown in table 1 fit this aspect of the data either.

To include these local effects, an NHMM was fit using the general autologistic model (eq. 1) for $P(\mathbf{R}_t \mid S_t)$. The conditional log-odds ratios, β_{sij}, were modelled as a function of the distance and direction between the stations to reduce the number of parameters. To determine an appropriate functional form for the β_{sij}, each day was first classified into its most likely weather state using the 6 state, 3 atmospheric variables, conditional spatial independence model described above (a procedure known as the Viterbi algorithm can be used to classify each day into a weather state; see, for example, Juang and Rabiner, 1991). Then, for each state, empirical estimates of the pairwise (unconditional) log-odds ratios were generated and plotted against the distance and direction between the stations. These plots suggested that the within-state spatial correlation declined as the distance between stations increased and varied elliptically with direction. Using a nonlinear least-squares regression analysis, 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:

$$\beta_{sij} = b_{0s} + b_{1s} \log(d_{ij}\sqrt{\cos(\phi_s + h_{ij})^2 + \sin(\phi_s + h_{ij})^2/e_s}) \qquad (12)$$

where d_{ij} and h_{ij} are, respectively, the distance and direction between stations i and j. For each state, s, there are 4 parameters in this model. Although, theoretically, all four parameters could be estimated by the methods outlined in section 3.2, estimation of the nonlinear parameters, ϕ_s and e_s, slows down the computations substantially. Therefore, these parameters were fixed at the values obtained from the nonlinear regression analyses of the empirical log-odds ratios. The b's were then estimated using the procedure described in section 3.2.

This approach significantly improved the fit of the model to the empirical log-odds ratios, as seen in figure 2. The EM/MCML algorithm converged to a model with estimated log-likelihood equal to -15688. This may be compared to the log-likelihood of the final model obtained under the assumption of conditional independence of -17110 with a difference of 24 parameters.

The ability of the NHMM to reproduce key precipitation statistics conditional 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 datasets (e.g. to investigate the effects of climate change) then it is important to test the model on reserved data. Figure 4 compares various observed rainfall statistics to those predicted by the model for the 5 years of reserved data. Results for the spatial model fit using the EM/MCML algorithm are shown. Figure 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 measureable shift in the mean atmospheric data fields occurred during the 5 year period of reserved data (-0.81hPa in MSLP, +0.47hPa in N-S SLP gradient, +0.45m in E-W GPH850). If this shift is deliberately removed from the atmospheric data (e.g. by recentering the atmospheric measures in the 5 year period around the same means as were observed in the 10 year period) then a small but noticable 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 figure 4 indicates that the model was able to adjust the rainfall probabilities to account for the (slight) nonstationary shift in the atmospheric data. This is clearly a necessary condition if the model is to be able to make useful predictions about rainfall under altered climates.

Although the weather states are abstract constructs of the model, they can be examined by first classifying each day into its most likely state (using the Viterbi algorithm) and then averaging the values of sea-level pressure, geopotential height or other atmospheric measures, over all days in a given state, at each node of the atmospheric data grid. The resulting "composite" field can be contoured to give a visual representation of the average field in that state. The resulting fields 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).

Figure 5 shows the rainfall probabilities and composite sea-level pressure and 850 hPa geopotential height fields associated with three out 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 runoff within the water supply catchments of the region (Ref to add!). This state occurred 20% of the time during the 1978 - 1987 period and then decreased to 18% of the time during the 1988 - 1992 period (table 2) which partly explains the reduced probability of precipitation during the latter period.

In contrast, the synoptic pattern associated with state 5 (low rainfall probability of rain at all stations) shows a dominant high pressure system centered in the Great Australian Bight. Such systems dominate the atmospheric circulation of the entire continent. Their intensity and rate of movement control the weather changes experienced across the study region (Sturman and Tapper, 1996). Their progression eastwards typically follows a well defined period of 5 to 7 days. However, highs can remain stationary in the Bight for longer periods, blocking the general circulation and leading to prolonged dry periods of continental easterlies throughout the study region (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 (table 2).

The remaining four states are characterized by rainfall in particular regions of the study area. For example, state 4 (shown in the middle row of plots in figure 5) exhibits high probability of rain at the southwest stations but low probability in the north and western stations. The frontal systems 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 (a 500 m escarpment running north-south for much of the region, approximately 25-50 km inland from the west coast).

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 MSLP plots are complemented by those seen in the 850 hPa GPH 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 the synoptic-scale atmospheric circulation to the point-scale multisite precipitation process.

Other aspects of the weather state process such as the percent of time spent in each weather state or the pattern of transitions between weather states can be derived from the model. Table 2 contrasts the relative frequency and mean duration for each of the six weather states for the 10 year and 5 year period. It is interesting that only the small changes in the frequency and persistance of the weather states seen in table 2 were required to produce the change in precipitation between the two periods that we have noted previously.

Table 2: Model predicted relative frequency and mean duration for the weather states.

	relative frequency					
	1	2	3	4	5	6
10yr	0.06	0.2	0.13	0.20	0.29	0.12
5yr	0.06	0.2	0.14	0.18	0.29	0.13

	mean duration (days)					
	1	2	3	4	5	6
10yr	2.4	1.8	1.4	1.8	2.7	1.4
5yr	2.3	1.8	1.4	1.7	2.6	1.5

5 Discussion

Nonhomogeneous hidden Markov models 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. thunderstorms) since these processes evolve on relatively small scales and may not be predictable from synoptic circulation patterns.

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, while reflecting meteorological intuition, may not be optimal for modelling rainfall. An important advantage of the NHMM approach is that it allows one to combine meteorological intuition (through selection of the atmospheric measures) with data analysis to define weather states that are "optimal" in the sense of separating precipitation patterns. Plots such as figure 5 can provide insight into the interpretation of the weather states and the relationship between atmospheric circulation patterns and precipitation. A comparison of the total variance in the atmospheric measures to the within-weather state variance can be used to assess the homogeneity of the weather states.

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 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 invalidate such an assumption. The atmospheric data are typically measured at a point in time while the rainfall measurements represent an accumulation over a 24 hour period. We believe that conditioning on the previous day's weather state helps recover some of the atmospheric information that is relevant to the 24 hour precipitation period. From a strictly practical point of view, inclusion of the Markov assumption typically 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 a climate change signal when the model is used to generate simulations 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 while models with more weather states achieve a minimum BIC with fewer atmospheric variables. Since one would expect that a model with more atmospheric information will produce precipitation simulations which are more responsive to shifts in atmospheric conditions, we favor models with fewer weather states and more atmospheric variables.

NHMMs represent a completely stochastic approach to the downscaling problem. Thus far, more mechanistic approaches, such as GCM-based simulations of precipitation, have proved to be deficient at the spatial and temporal scales of relevance to regional and local hydrology (Grotch and MacCraken, 1991). Although it is, at present, computationally impossible to implement an entire GCM at local scales, some progress has been made in developing "nested" GCMs which implement phenomenologic models for rainfall on a finer grid over a restricted area and use the coarse scale GCM data as boundary conditions. These limited area models are able to achieve grid spacings on the order of 1^o by 1^o. Even at this scale, however, deficiencies in the precipitation simulations have been noted (Mearns et al., 1995). Additional studies to compare the NHMM approach with the nested GCM approach in terms of ability to accurately reproduce current climate precipitation patterns are ongoing (Charles et al., 1996). However, even if GCMs are, at some point, able to accurately characterize local precipitation patterns, downscaling models will still be valuable for modelling phenomena that are not explicitly included in the GCMs (e.g. air pollution patterns).

We believe that future research in this area should focus on both conceptual and methodological issues. The outstanding conceptual issue in research on downscaling is making predictions under altered climate regimes. Predictions of the effects of hypothesized changes in climate (e.g. global warming) are based on GCM simulations and are, therefore, restricted to large scale effects. As described in IPCC (1995, sec. 6.6), there are considerable discrepancies between predictions of different GCMs in terms of changes in precipitation that would occur on a sub-continental scale under a doubled CO_2 climate. In addition, there are substantial biases in precipitation between GCM control runs and observations. At present, therefore, assessment of the local hydrologic effects of climate change necessitates the use of models to downscale the (altered climate) GCM circulation patterns. However, this means that the downscaling models must be used under different conditions than they were fit under. Clearly, in the absence of observations from the altered climate, it is impossible to completely ascertain the validity of a downscaling model under an altered climate regime; indeed, model validity may vary depending on the nature of the climate change. However, certain aspects of the model and the intended application may increase our confidence in the ability of the model to predict precipitation under an altered climate. These include

- The range of the atmospheric variables under the altered climate should be similar to the range of these measures under the current climate

- The model should include all atmospheric measures showing distributional changes under the altered climate (unless it can be demonstrated that the omitted covariates have no effect on precipitation)

- The model should respond to changes in observed precipitation caused by climate variability (as in the example presented here) and other natural changes in the current climate regime (e.g. the eruption of Mt. Pinatubo in 1991, which caused measureable changes in global climate)

Of course, the validity of downscaling models for impact assessment also depends on the validity of the GCM model which provides the atmospheric information that drives precipitation. Assessment of GCM models is an active area of research — the interested reader is referred to the IPCC report (1995, chapter 5) for a summary of the current state of knowledge.

Several methodological issues remain, also. The model developed here deals with rainfall occurrences only. For some applications it is also necessary to simulate amounts. One approach is to first fit an NHMM to the occurrence data and then fit a model to the amounts, conditional on occurrence (and, possibly, weather state), a posteriori. The simplest approach to simulating amounts is to independently fit a model at each station. However, we have observed that there is considerable spatial correlation in the amounts even after the spatial correlation in the occurrences is accounted for using the model proposed here. Correct simulation of the spatial distribution of amounts is particularly important for runoff and flood models. When only a few stations are modelled, simulated amounts can be obtained by resampling an entire vector of amounts from the historical data, conditional on the simulated occurrence pattern (i.e. sampling from the joint amounts distribution). In applications with many stations this approach is problematic since it is possible to simulate occurrence patterns that never occur in the historical data. In that case, it is possible to incorporate spatial dependencies in the simulated amounts by resampling separately from the historical data at each station, conditional on the simulated occurrence pattern at nearby stations (i.e. sampling from the univariate conditional distributions).

In either of the approaches outlined above, however, the amounts do not influence the definitions of the weather states. To fully integrate an amounts model into the NHMM would require specification of a multivariate mixed discrete-continuous model for $P(\mathbf{R}_t \mid S_t)$.

While there is a large literature on rainfall models at single stations, multi-station models are less common. In the context of a weather state model, Bardossy and Plate (1992) used a truncated multivariate normal distribution to model amounts at multiple stations. We are currently investigating this and other approaches to this problem and hope to report results at a future date. To extend this idea further, models based on multivariate observations (e.g. precipitation and temperature) could be developed and would be useful for input into hydrologic models.

To further extend the utility of the weather state approach, methods could be developed to simulate rainfall occurrence at locations that have not been explicitly included in the model. In the context of the autologistic model this could be accomplished by spatially interpolating the α_{si} (note that a spatially smooth model for β_{sij} has already been included in the present analysis). For the example presented in section 4 we observed that the α_{si} from the best-fitting autologistic model were small and showed little variation within weather state in the interior of the network (e.g. -2.0 to -6.0, depending on the weather state; note that $\exp(\alpha_{si})/1 + \exp(\alpha_{si})$ is the probability of rain at station i given no rain at all other stations). Thus, to generate rainfall probabilities at a new location, i', in the interior of the network one could set $\alpha_{si'}$ equal to the mean value of α_{si} from other stations in the interior and compute $\beta_{si'j}$ from (12).

Acknowledgements: Although 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, it has as 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 CSIRO, 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. Software (written in Fortran with calls to NAG library routines) to implement the methods discussed in this paper is available from the first author.

References

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

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

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

[4] Besag, J. (1977) Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika, 64*, 616-618.

[5] Breiman L., Friedman J.H., Olshen R.A., and Stone, C.J. (1984) *Classification and Regression Trees*, Wadsworth International Group, Belmont CA.

[6] Charles, S.P., J.P. Hughes, B.C. Bates and T.J. Lyons (1996) Assessing downscaling models for atmospheric circulation — local precipitation linkage. *Proceedings of the International Conference on Water Resources and Environmental Research: Towards the 21st Century*, Tokyo, Japan.

[7] Cleveland, W.S., and Devlin, S.J., (1988) Locally-weighted regression: An approach to regression analysis by local fitting. *J. Am. Statist. Assoc., 83*, 596-610.

[8] Cressie, N.A.C. (1993) *Statistics for Spatial Data.* John Wiley and Sons, Inc., New York, New York, 900pp.

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

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

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

[12] Geman, S and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *I.E.E.E. Trans. Pattern Anal. Machine Intell., 6*, 712-741.

[13] Giorgi, F. and L. O. Mearns (1991) Approaches to the simulation of regional climate change: A review. *Reviews of Geophysics, 29*, 191- 216.

[14] Grenander, U. (1989) Advances in pattern theory. *Annals of Statistics, 17*, 1-30.

[15] Grotch, S.L. and MacCracken, MC (1991) The use of general circultion models to predict climate change. *J. Climate, 4*, 286-303.

[16] 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.

[17] Hughes, J. P. and P. Guttorp (1994) A Class of Stochastic Models for Relating Synoptic Atmospheric Patterns to Regional Hydrologic Phenomena. *Water Resour. Res., 30*, 1535-1546.

[18] Intergovermental Panel on Climate Change (1995) *Climate Change 1995, The Science of Climate Change*, (ed. J.T. Houghton, L.G. Meira Filho, B.A. Callander, N. Harris, A. Kattenberg, and K. Maskell) Cambridge University Press, Cambridge.

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

[20] Kass, R.E. and A.E. Raftery (1995) Bayes factors. J. Am. Statist. Assoc., 90, 773-795.

[21] Kavvas, M. L. and J. W. Delleur (1975) The stochastic and chronological structure of rainfall sequences: Application to Indiana, *Water Resour. Res. Center Rep. 57,* Purdue Univ., West Lafayette, Ind.

[22] Kavvas, M. L. and J. W. Delleur (1981) A stochastic cluster model for daily rainfall sequences. *Water Resour. Res., 17*, 1151-1150.

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

[24] LeCam, L. (1961) A stochastic theory of precipitation. Fourth Berkeley Symposium on Mathematics, Statistics, and Probability, Univ. of Calif., Berkeley, Calif.

[25] Mearns, L.O., Giorgi, F., McDaniel, L. and Shields, C. (1995) Analysis of daily variability of precipitation in a nested regional climate model: comparison with observations and doubled CO2 results. *Glob. Planet. Change, 10*, 55-78.

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

[27] Southern, R.L. (1979) The Atmosphere in *Environment and Science,* O'Brien, B.J. (ed.), University of Western Australia Press, Nedlands, Western Australia, pp183-226.

[28] Sturman, A. and Tapper, N. (1996) The Weather and Climate of Australia and New Zealand. Oxford University Press, Oxford. 476pp.

[29] Stern, R. D. and R. Coe (1984) A model fitting analysis of daily rainfall data. *J. R. Statist. Soc. A, 147*, 1-34.

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

[31] Waymire, E. and V. K. Gupta (1981) The mathematical structure of rainfall representations 2. A review of the theory of point processes. *Water Resour Res., 17*, 1273-1285.

[32] Yarnal, B. (1993) Synoptic climatology in environmental analysis. Blehaven Press, London. 195pp.

List of Figures

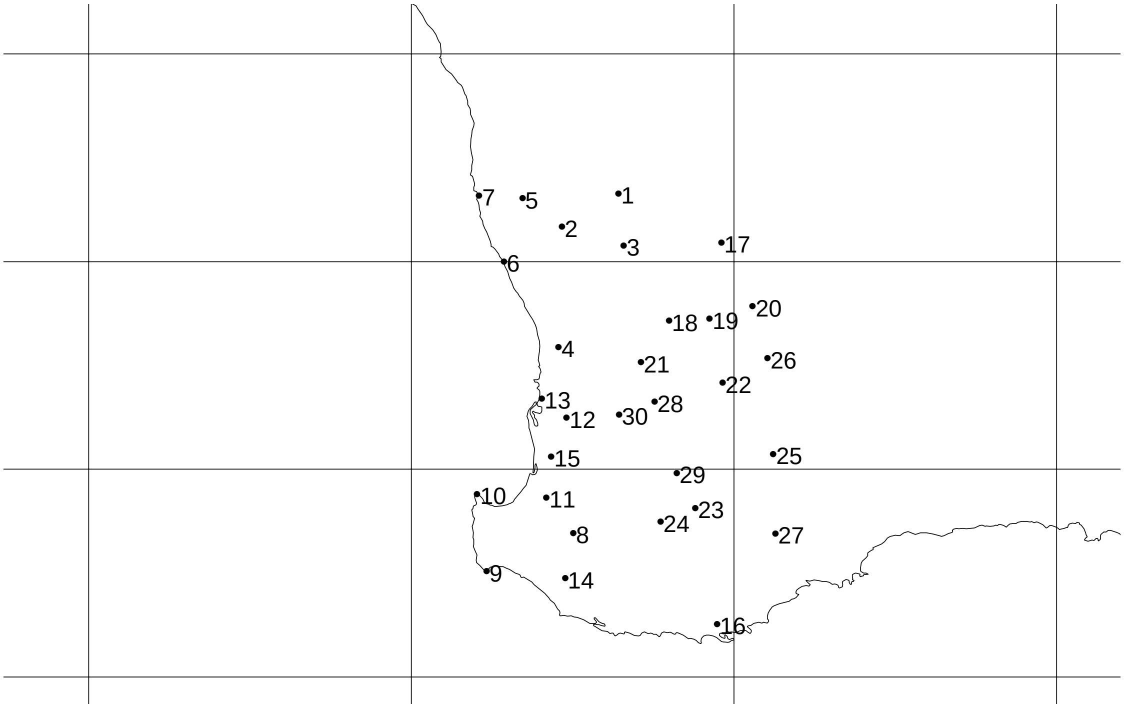

Figure 1: Map of study area showing the locations of the atmospheric data grid and rain gauge stations in southwestern Australia. Atmospheric data are interpolated to the verticies of the grid as described in the text.

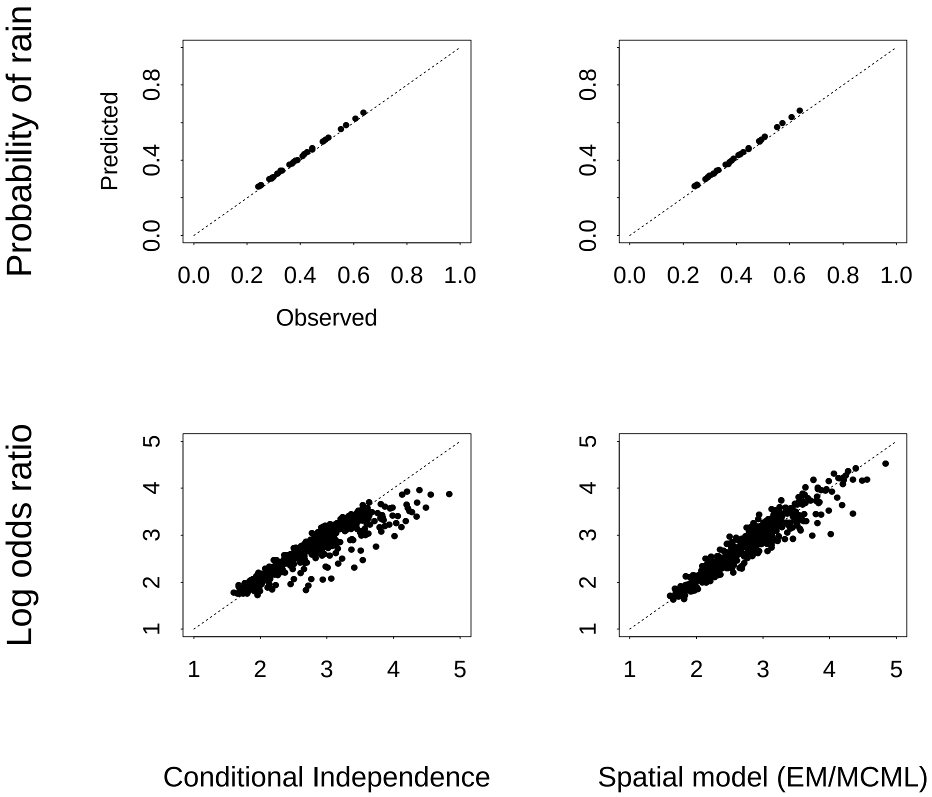

Figure 2: Comparison of observed and model-predicted rainfall statistics based on the 10 years of data used for model fitting. Model-predicted statistics are generated by simulation from the fitted model using the observed atmospheric data. Observed statistics are on the x-axis; model-predicted statistics are on the y-axis.

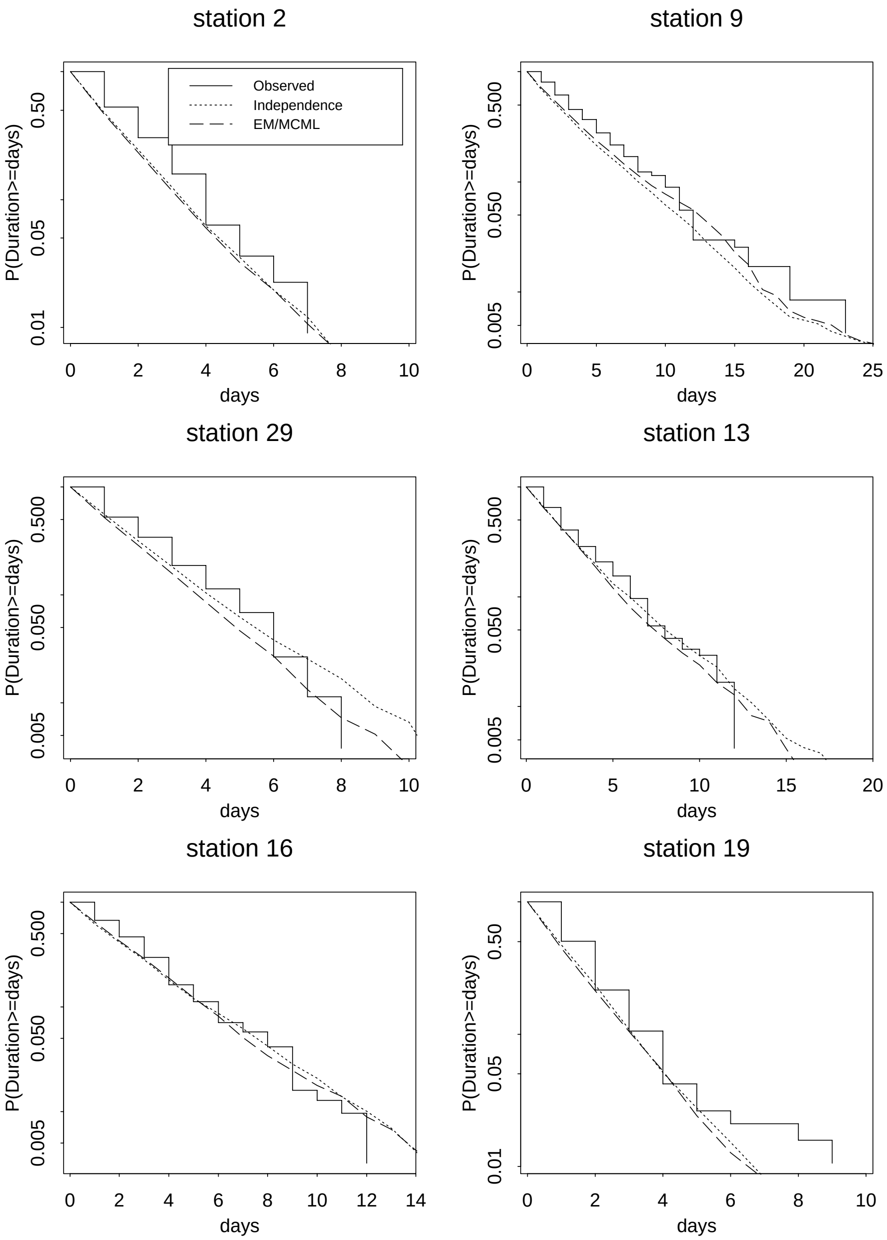

Figure 3: Comparison of observed and model-predicted rainfall statistics: duration distribution. Results are presented for 6 representative stations (see figure 1). Observed and model-predicted statistics are based on the 10 years of data used for model fitting.

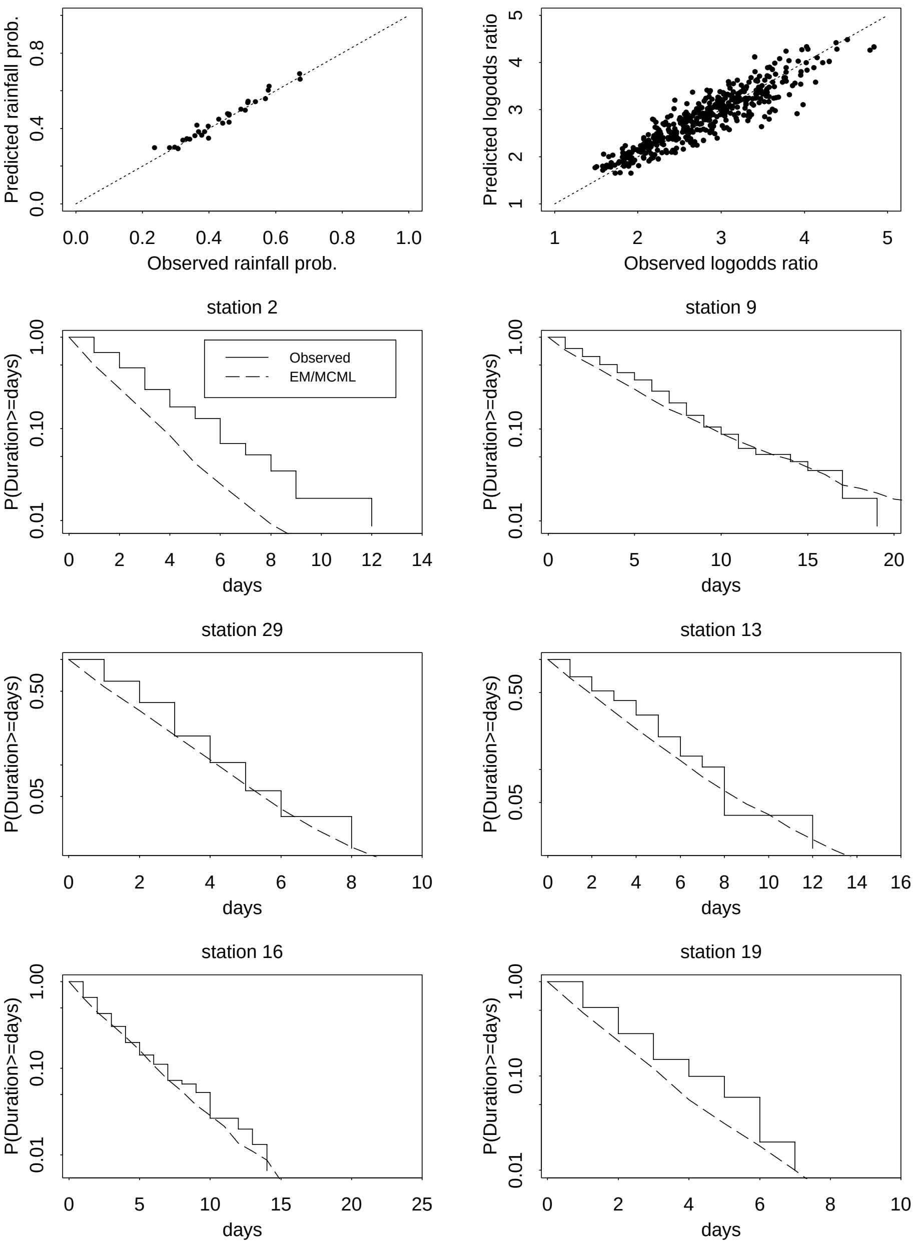

Figure 4: Comparison of observed and model-predicted rainfall statistics on the 5 years of reserved data. Model-predicted statistics are generated by simulation from the fitted model using the observed atmospheric data. Duration distributions are shown at a representative subset of stations. Station 2 represents the poorest fit seen at any station.

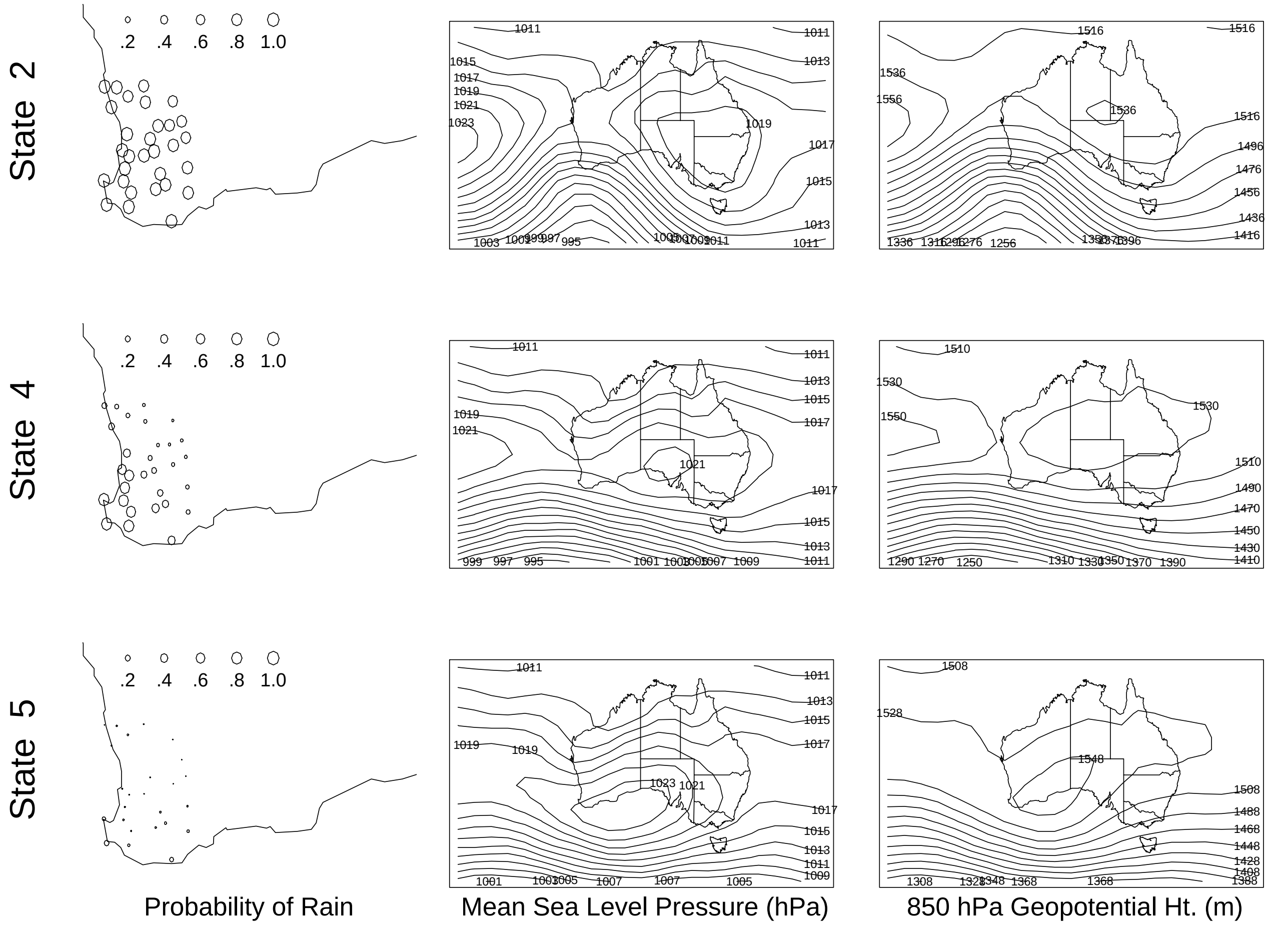

Figure 5: Probability of rain, composite sea-level pressure (hPa) and 850 hPa geopotential height (m) fields for three states from the six state spatial model estimated using EM/MCML. Each day is first classified into its most likely state using the Viterbi algorithm. All days in a particular state are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