J. Appl. Probab. **38A**, 142–157 (2001)
Printed in England
© *Applied Probability Trust* 2001

FITTING HIDDEN SEMI-MARKOV MODELS TO BREAKPOINT RAINFALL DATA

JOHN SANSOM,[1] *National Institute of Water and Atmospheric Research, New Zealand*
PETER THOMSON,[2] *Statistics Research Associates Ltd*

Abstract

The paper proposes a hidden semi-Markov model for breakpoint rainfall data that consist of both the times at which rain-rate changes and the steady rates between such changes. The model builds on and extends the seminal work of Ferguson (1980) on variable duration models for speech. For the rainfall data the observations are modelled as mixtures of log-normal distributions within unobserved states where the states evolve in time according to a semi-Markov process. For the latter, parametric forms need to be specified for the state transition probabilities and dwell-time distributions.

Recursions for constructing the likelihood are developed and the EM algorithm used to fit the parameters of the model. The choice of dwell-time distribution is discussed with a mixture of distributions over disjoint domains providing a flexible alternative. The methods are also extended to deal with censored data. An application of the model to a large-scale bivariate dataset of breakpoint rainfall measurements at Wellington, New Zealand, is discussed.

Keywords: Hidden semi-Markov models; high-resolution rainfall data; dwell-time distributions; EM algorithm.

AMS 2000 Subject Classification: Primary 62M05
Secondary 86A10

1. Introduction

Many meteorological datasets are generated by mixtures of components that correspond to particular physical phenomena. In particular, rainfall is generated by at least two processes, one caused by small-scale thermal processes leading to convection and the other by large-scale air movements leading to frontal systems, each of which is characterized by its own distribution of rain rates and durations. A primary objective is to model and identify these phenomena accurately: this is important from a meteorological standpoint, both in terms of understanding the nature and dynamics of the underlying rainfall process and for forecasting.

There are a number of time scales over which rainfall can be investigated ranging from coarse annual time scales to more traditional daily data. Finer time scales are also available, in principle down to counts of raindrops. The focus of this research

Received November 2000.

[1] Postal address: PO Box 14-901, National Institute of Water and Atmospheric Research, Wellington, New Zealand.

[2] Postal address: PO Box 12-649, Statistics Research Associates Ltd, Wellington, New Zealand. Email: `peter@statsresearch.co.nz`

is on *breakpoint rainfall data*: these record the timings of rain-rate changes and the steady rates between changes. Such high-resolution datasets have only recently become available and promise to capture better the information needed to model rainfall dynamics. Breakpoint data are typically digitized from pluviographs (see Sansom, 1999, for further discussion) and are able to resolve steady rain that lasts for periods of 1–20 minutes during which amounts of 0.01–0.50 mm can accumulate. As a consequence, the range of time scales present in breakpoint rainfall data is extremely diverse.

To cope with this diversity of time scales, we have chosen to model the breakpoint data as a hidden semi-Markov model (HSMM) whose unobserved meteorological states form a suitable hierarchy of rainfall events evolving over time, each with its own dynamic time scale. Within the unobserved states, the breakpoint observations are modelled as bivariate and univariate mixtures of log-normal distributions while the states follow a semi-Markov model with suitably specified parametric state transition probabilities and *dwell-time* distributions (known as sojourn time distributions in the stochastic process literature). We discuss the choice of these distributions and show that a mixture of distributions over disjoint domains provides a flexible alternative. The methods are also extended to deal with censored data where the values of observations below a threshold (univariate data) or a line (bivariate data) are ignored.

This model builds on and extends the seminal work of Ferguson (1980) on variable duration models for speech and includes the more common hidden Markov model (HMM) and standard mixture models. More general references to HSMM and HMM modelling include Baum and Petrie (1966) who first proposed the HMM model, Levinson *et al.* (1983), Rabiner (1989), Elliott *et al.* (1995) and MacDonald and Zucchini (1997). These and related methods have been used widely in meteorological contexts (see e.g. Zucchini and Guttorp (1991), Hughes *et al.* (1999), Sansom (1995, 1998, 1999) and Sansom and Thomson (1992, 1998)).

The EM algorithm (Dempster *et al.* 1977) is used to fit the parameters of the model using maximum likelihood with model orders selected by minimizing the BIC criterion (Akaike, 1977) where

$$\text{BIC} = -2(\text{maximized log likelihood}) + 2(\text{number of parameters}) \log T$$

and T denotes the number of observations. As is the case with other HSMM and HMM applications, the recursions involved must be appropriately scaled in order to preserve numerical precision. Monte Carlo methods are used to obtain variance estimates for the parameters. These issues are discussed in relation to an application of the model to a large-scale bivariate dataset of breakpoint rainfall measurements taken over a 5 year period at Wellington, New Zealand.

2. The model

To accommodate the various time scales present in the breakpoint data the model assumes a hierarchy of states. These include large-scale precipitation events and concomitant inter-event dry periods; medium-scale rain and shower episodes within precipitation events where the precipitation generating mechanism remains constant; and fine-scale periods of steady rain or dry periods within episodes which are recorded by the breakpoint data. This hierarchy is depicted in Figure 1.

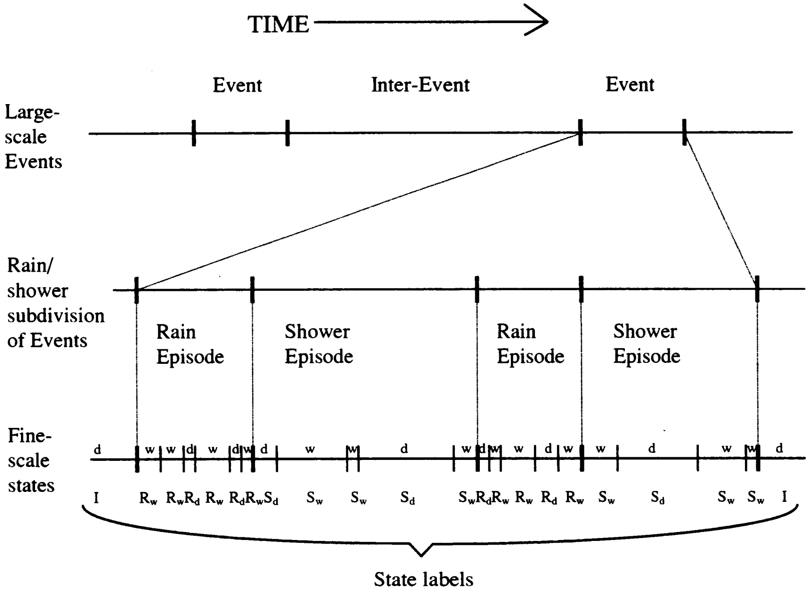
TIME
Event
Inter-Event
Event
Large-scale Events
Rain/shower subdivision of Events
Rain Episode
Shower Episode
Rain Episode
Shower Episode
Fine-scale states
d w w d w d w d w w d w d w w d w w d w w d
I R_w $R_w$$R_d$ R_w $R_d$$R_w$$S_d$ S_w S_w S_d $S_w$$R_d$$R_w$ R_w R_d R_w S_w S_d S_w S_w I
State labels

FIGURE 1: *Hierarchical division of time into its component time scales.* Large-scale precipitation events and dry inter-events depicted in the top line are subdivided into medium-scale rain or shower episodes (middle line), and then episodes are further subdivided into fine-scale individual wet and dry breakpoint durations (bottom line). Here 'd' and 'w' indicate dry and wet, which is known from the breakpoint data, and 'R', 'S' and 'I' indicate rain or shower episodes and dry inter-events which are not known from the breakpoint data.

In this way a variety of different rainfall mechanisms are incorporated at the various time scales in a way that is both relatively simple and transparent yet accords with the general division of precipitation types in the meteorological literature (see Rogers, 1979, for example). Note, however, that this structure is not directly or completely observed and must be inferred from the breakpoint data itself which represents the last level of the hierarchy.

Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_T$ denote the sequence of suitably transformed breakpoint vector observations and $S_1, \dots, S_T$ the corresponding sequence of hidden states. In our case the bivariate observations $\boldsymbol{X}_t$ are the logarithms of the rainfall rates and durations since this choice of transformation made the $\boldsymbol{X}_t$ approximately Gaussian conditional on the state information. The states S_t are assumed to take values in $\{1, \dots, N\}$ with $S_t = i$ if S_t is in the ith state: N is typically unknown and must be estimated from the data.

Given the state information $\boldsymbol{S} = (S_1, \dots, S_T)$, the observations $\{\boldsymbol{X}_t : t = 1, \dots, T\}$ are assumed to be independent with distributions

$$\Pr\{\boldsymbol{x} \le \boldsymbol{X}_t < \boldsymbol{x} + \mathrm{d}\boldsymbol{x} \mid \boldsymbol{S} = \boldsymbol{s}\} = \Pr\{\boldsymbol{x} \le \boldsymbol{X}_t < \boldsymbol{x} + \mathrm{d}\boldsymbol{x} \mid S_t = s_t\} = f_{s_t}(\boldsymbol{x} \mid \Theta_{s_t})\,\mathrm{d}\boldsymbol{x}$$

where Θ_{s_t} denotes the parameters of the distribution of the observations $\boldsymbol{X}_t$ when $S_t = s_t$ and the inequalities involving the vectors $\boldsymbol{X}_t$ are interpreted componentwise. $\{S_t\}$ is assumed to be a semi-Markov process where the transitions between different states are generated by a finite Markov chain and the dwell times within the same state i by the state-dependent duration distributions $p_i(d)$ $(i = 1, \dots, N; d = 1, 2, \dots)$. The embedded Markov chain is assumed to be irreducible and ergodic with transition probability matrix $\boldsymbol{A} = \{a_{ij}\}$, where

$$a_{ij} = \Pr\{S_{t+1} = j \mid S_t = i, S_{t+1} \neq i\} \qquad (i, j = 1, \dots, N)$$

and $\sum_{j=1}^{N} a_{ij} = 1$ $(i = 1, \dots, N)$. Note that self-transitions are forbidden.

This is an example of a *hidden semi-Markov model* where, as might be expected, the observed process is assumed to be in a stationary equilibrium.

The S_t can be regarded as a function of the bivariate stochastic process (I_k, D_k) $(k = 1, 2, \dots)$ where k indexes the state visits, I_k denotes the state visited and D_k the duration of that visit. Here the durations D_k are conditionally independent given the state information $\boldsymbol{I} = (I_1, I_2, \dots)$ with

$$\Pr\{D_k = d \mid \boldsymbol{I} = \boldsymbol{i}\} = \Pr\{D_k = d \mid I_k = i_k\} = p_{i_k}(d) \qquad (i_k = 1, \dots, N;\ d = 1, 2, \dots),$$

and the I_k constitute an irreducible ergodic finite Markov chain with transition probability matrix $\boldsymbol{A}$. Thus the probability of an observed sequence of states is now

$$\Pr\{I_1 = i_1, D_1 = d_1, \dots, I_r = i_r, D_r = d_r\} = \pi_{i_1} p_{i_1}(d_1) \prod_{k=2}^{r} a_{i_{k-1} i_k} p_{i_k}(d_k),$$

where π_i $(i = 1, \dots, N)$ denotes the initial probability distribution for the Markov chain.

We are now in a position to formulate the likelihood of the observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_T$ which, in keeping with EM terminology, we shall refer to as the *incomplete likelihood*. The *complete likelihood* is given by the likelihood of the $(\boldsymbol{X}_t, S_t)$ $(t = 1, \dots, T)$. However, before doing so, we first make some additional assumptions and observations concerning the data and the model.

In keeping with the nature of the breakpoint data, it is assumed that the time points $t = 1$ and $t = T$ are known to be state boundaries where $S_0 \neq S_1$ and $S_T \neq S_{T+1}$ respectively. This is a natural assumption in our context since a state change must occur whenever the process moves from a dry period to a wet period or vice versa. In fact more can frequently be assumed, but this additional information has not been taken into account. Given that $t = 1$ and $t = T$ are state boundaries

$$T = D_1 + \cdots + D_R\,,$$

where R denotes the number of state visits in the observation sequence. Depending on how the data collection process is viewed, either or both of T and R are random variables. Here we shall assume that R is a fixed, but unknown, constant so that it is R rather than T that indexes the information collection process and estimation accuracy. In particular the states S_t in the observation sequence can now be regarded as being generated by $\{(I_k, D_k) \colon k = 1, \dots, R\}$.

Given these assumptions and observations the log-likelihood of the complete data is given by

$$\log L_c(\phi) = \sum_{t=1}^{T} \log f_{S_t}(\boldsymbol{X}_t \mid \Theta_{S_t}) + \log P(\boldsymbol{I}, \boldsymbol{D}), \tag{1}$$

where

$$\log P(\boldsymbol{I},\boldsymbol{D}) = \log \pi_{I_1} + \sum_{k=2}^{R} \log a_{I_{k-1}I_k} + \sum_{k=1}^{R} \log p_{I_k}(D_k) \tag{2}$$

and the second term on the right-hand side is interpreted as zero when $R = 1$. Here the vector $\boldsymbol{\phi}$ in (1) denotes the totality of parameters in the model including the parameters Θ_i specifying the conditional distributions of the observations $\boldsymbol{X}_t$, the parameters $\boldsymbol{A}$ and π_i specifying the embedded Markov chain, and the parameters specifying the dwell-time distributions $p_i(d)$. If the conditional distribution of $\boldsymbol{X}_t$ given $S_t = i$ is Gaussian with mean $\boldsymbol{\mu}_i$ and covariance matrix Σ_i then

$$\log f_i(\boldsymbol{X}_t \mid \Theta_i) = -\tfrac{1}{2}\log|\Sigma_i| - \tfrac{1}{2}(\boldsymbol{X}_t - \boldsymbol{\mu}_i)^{\top}\Sigma_i^{-1}(\boldsymbol{X}_t - \boldsymbol{\mu}_i) \qquad (i = 1,\dots,N) \tag{3}$$

and $\Theta_i = (\boldsymbol{\mu}_i, \Sigma_i)$.

Finally the likelihood of the incomplete data $\boldsymbol{X}_1,\dots,\boldsymbol{X}_T$ is given by

$$L(\boldsymbol{\phi}) = \sum_{\boldsymbol{I}}\sum_{\boldsymbol{D}} P(\boldsymbol{I},\boldsymbol{D}) \prod_{t=1}^{T} f_{S_t}(\boldsymbol{X}_t \mid \Theta_{S_t}), \tag{4}$$

and it is this which should ideally be optimized to determine maximum likelihood estimators $\widehat{\boldsymbol{\phi}}$ of the model parameters $\boldsymbol{\phi}$. However its structure makes it more difficult to optimize than $\log L_c(\boldsymbol{\phi})$, and for this reason we turn to the EM algorithm.

3. The EM algorithm

If the states S_t were known, then it is the relatively simple complete log-likelihood $\log L_c(\boldsymbol{\phi})$ that would be optimized to determine estimators of $\boldsymbol{\phi}$. Given only the observations $\boldsymbol{X} = (\boldsymbol{X}_1,\dots,\boldsymbol{X}_T)$, the best (quadratic loss) predictor of $\log L_c(\boldsymbol{\phi})$ is

$$\log \widetilde{L}(\boldsymbol{\phi} \mid \boldsymbol{\phi}_0) = \mathrm{E}(\log L_c(\boldsymbol{\phi}) \mid \boldsymbol{X}),$$

where the expectation operator E is with respect to the true distribution indexed by $\boldsymbol{\phi}_0$. Given an initial estimate of $\boldsymbol{\phi}_0$ a new estimate can be found by maximizing $\log \widetilde{L}(\boldsymbol{\phi} \mid \boldsymbol{\phi}_0)$ with respect to the parameters $\boldsymbol{\phi}$. This forms the basis of the celebrated EM algorithm (Dempster *et al.*, 1977) where the determination of the conditional expectation $\log \widetilde{L}(\boldsymbol{\phi} \mid \boldsymbol{\phi}_0)$ is referred to as the E-step and its maximization with respect to $\boldsymbol{\phi}$ the M-step. Under certain general conditions it can be shown that the sequence of estimates constructed in this way yields monotonically increasing values of $L(\boldsymbol{\phi})$ and converges to the maximum likelihood estimator $\widehat{\boldsymbol{\phi}}$ for the incomplete data. The computational efficiency of the EM algorithm is greatly enhanced if the E-step and M-step are readily evaluated, particularly the M-step where simple closed form solutions are desired. These design objectives underpin the development that follows.

To develop a convenient computational form for $\log \widetilde{L}(\boldsymbol{\phi}) \mid \boldsymbol{\phi}_0)$ we first define the indicator random variables $N_t(i,j)$ and $M_t(i,d)$. Set $N_t(i,j) = 1$ if $(S_t, S_{t+1}) = (i,j)$ and $N_t(i,j) = 0$ otherwise, and $M_t(i,d) = 1$ if $S_t = i$ is in a visit of duration d and $M_t(i,d) = 0$ otherwise. Then, within the observation sequence, the number of transitions from state i to state j equals $\sum_{t=1}^{T-1} N_t(i,j)$ and the number of visits to state i that have duration d is $d^{-1}\sum_{t=1}^{T} M_t(i,d)$. Thus (2) can be written as

$$\log P(\boldsymbol{I},\boldsymbol{D}) = \log \pi_{I_1} + \sum_{i=1}^{N}\sum_{j\neq i} \log a_{ij} \sum_{t=1}^{T-1} N_t(i,j) + \sum_{i=1}^{N}\sum_{d=1}^{T} \frac{1}{d}\sum_{t=1}^{T} M_t(i,d).$$

Note that the last summation over d is terminated at T since there is a state boundary at $t = T$. Using this representation $\log \widetilde{L}(\boldsymbol{\phi} \mid \boldsymbol{\phi}_0)$ is now given by

$$\begin{aligned}\log \widetilde{L}(\boldsymbol{\phi} \mid \boldsymbol{\phi}_0) &= \sum_{i=1}^{N}\sum_{t=1}^{T} \gamma_t(i) f_i(\boldsymbol{X}_t \mid \Theta_i) + \sum_{i=1}^{N} \gamma_1(i)\log \pi_i \\ &\quad + \sum_{i=1}^{N}\sum_{j\neq i}\left(\sum_{t=1}^{T-1}\gamma_t(i,j)\right)\log a_{ij} + \sum_{i=1}^{N}\sum_{d=1}^{T}\left(\frac{1}{d}\sum_{t=1}^{T}\delta_t(i,d)\right)\log p_i(d), \quad (5)\end{aligned}$$

where $f_i(\boldsymbol{X}_t \mid \Theta_i)$ is given by (3), and as functions of the observations $\boldsymbol{X}$ and $\boldsymbol{\phi}_0$ we have

$$\begin{aligned}\gamma_t(i,j) &= \Pr\{S_t = i, S_{t+1} = j \mid \boldsymbol{X}, \boldsymbol{\phi}_0\}, \\ \delta_t(i,d) &= \Pr\{S_t = i \text{ and } S_t \text{ is in a visit of duration } d \mid \boldsymbol{X}, \boldsymbol{\phi}_0\}, \\ \gamma_t(i) &= \Pr\{S_t = i \mid \boldsymbol{X}, \boldsymbol{\phi}_0\}.\end{aligned}$$

Evidently,

$$\gamma_t(i) = \sum_{j=1}^{N} \gamma_t(i,j) = \sum_{d=1}^{T} \delta_t(i,d),$$

so that, in principle, only the $\gamma_t(i,j)$, $\delta_t(i,d)$ need to be determined. Computationally efficient, recursive procedures for calculating these probabilities were first given by Ferguson (1980) in the context of speech recognition (see also Rabiner, 1989; Guédon and Cocozza-Thivent, 1990, for example) and are not given here. However these procedures, although straightforward, involve careful scaling to preserve numerical accuracy. Sansom and Thomson (2000) give further details on the derivation of the recursions and the use of scaling.

Since the components that make up the summation in (5) all involve disjoint parameter sets, the maximization of $\log \widetilde{L}(\boldsymbol{\phi} \mid \boldsymbol{\phi}_0)$ involves separate maximizations for each component. For the first three of these, simple standard closed form solutions exist for the updated estimates of the parameters. In particular the values of the $\boldsymbol{\mu}_i$, Σ_i, π_i and a_{ij} that maximize (5) are given by

$$\tilde{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T}\gamma_t(i)\boldsymbol{X}_t}{\sum_{t=1}^{T}\gamma_t(i)}, \qquad \tilde{\Sigma}_i = \frac{\sum_{t=1}^{T}\gamma_t(i)(\boldsymbol{X}_t - \tilde{\boldsymbol{\mu}}_i)(\boldsymbol{X}_t - \tilde{\boldsymbol{\mu}}_i)^{\top}}{\sum_{t=1}^{T}\gamma_t(i)},$$

$$\tilde{\pi}_i = \gamma_1(i), \qquad \tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1}\gamma_t(i,j)}{\sum_{k\neq i}\sum_{t=1}^{T-1}\gamma_t(i,k)},$$

where i, $j = 1,\ldots,N$ and $i \neq j$. To maintain the computational efficiency of the algorithm it is therefore important to parameterize the dwell-time distributions so that the maximization of the fourth component also yields closed form solutions. This is addressed in the next section.

Subject to certain general conditions including identifiability conditions (Dempster *et al.*, 1977), the EM algorithm converges to the relevant maximum likelihood estimator $\widehat{\phi}$. Moreover $\widehat{\phi}$ is a consistent estimator of the vector of true parameters ϕ_0 and, as R increases, $\sqrt{R}(\widehat{\phi}-\phi_0)$ has a limiting multivariate Gaussian distribution with zero mean and a covariance matrix that is the inverse of the Fisher information matrix derived from the incomplete likelihood. The latter can in principle be obtained by numerical evaluation of the incomplete likelihood and its derivatives or, more efficiently, by using the EM algorithm and the technique of Meilijson (1989) (see also Hughes (1997)). Alternatively and as done here, the expectations involved in the determination of the information matrix can be estimated by Monte Carlo simulation of the model about $\widehat{\phi}$.

Although R is a fixed parameter it is in general unknown and must be estimated from the data. From the definition of $M_t(i,d)$ observe that

$$R = \sum_{i=1}^{N} (\text{number of visits to state } i) = \sum_{i=1}^{N}\sum_{d=1}^{T}\frac{1}{d}\sum_{t=1}^{T} M_t(i,d).$$

Taking expectations of both sides of this identity, conditioned on the observations, yields

$$R = \sum_{i=1}^{N}\sum_{d=1}^{T}\frac{1}{d}\sum_{t=1}^{T}\delta_t(i,d), \tag{6}$$

which is a function of the observations and ϕ_0. A natural estimator of R is now obtained by replacing ϕ_0 by $\widehat{\phi}$ in (6).

4. Dwell-time distributions

Maximizing (5) with respect to the parameters of the duration distributions requires the maximization of

$$\sum_{d=1}^{T}\Delta_i(d)\log p_i(d) \qquad (i=1,\dots,N), \tag{7}$$

where

$$\Delta_i(d) = \frac{1}{d}\sum_{t=1}^{T}\delta_t(i,d) \qquad (d=1,\dots,T)$$

are functions of the observations and $p_i(d)$ is suitably parameterized. Some examples follow.

Non-parametric distribution. Here the dwell times are assumed to have an arbitrary discrete distribution over the range $1,\dots,D_i$ for suitably chosen D_i, and maximizing (7) subject to $\sum_{d=1}^{D_i} p_i(d) = 1$ yields

$$\tilde{p}_i(d) = \frac{\Delta_i(d)}{\sum_{d=1}^{D_i}\Delta_i(d)} \qquad (d=1,\dots,D_i;\ i=1,\dots,N).$$

This is just the empirical distribution based on the observations; it is particularly useful for ascertaining the shape of the dwell-time probability functions prior to fitting a more parsimonious parametric model.

Geometric distribution. For the geometric,

$$p_i(d) = (1-p_i)^{d-1} p_i \qquad (d = 1, 2, \ldots;\ i = 1, \ldots, N),$$

and maximizing (7) yields

$$\tilde{p}_i = \frac{\sum_{d=1}^{T} \Delta_i(d)}{\sum_{d=1}^{T} d\Delta_i(d)} \qquad (i = 1, \ldots, N).$$

This case reduces the HSMM model to the more conventional HMM model.

Translated Poisson distribution. Here

$$p_i(d) = \mathrm{e}^{-\lambda_i} \frac{\lambda_i^{d-1}}{(d-1)!} \qquad (d = 1, 2, \ldots),$$

and maximizing (7) yields

$$\tilde{\lambda}_i = \frac{\sum_{d=1}^{T} d\Delta_i(d)}{\sum_{d=1}^{T} \Delta_i(d)} - 1 \qquad (i = 1, \ldots, N).$$

Other discrete distributions can be used. However, to maintain the computational simplicity and efficiency of the EM algorithm, we seek distributions that admit closed form maximum likelihood solutions. The need in practice for flexible parametric or semi-parametric families of discrete distributions leads us to consider a particular class of mixture distributions.

4.1. Mixtures of distributions over disjoint domains

In general, a mixture of discrete distributions will typically not lead to closed form maximum likelihood solutions even if the component distributions do. The case of a mixture of geometrics is one such example. However, if we restrict attention to mixtures of discrete distributions that do admit closed form solutions and which are defined over disjoint domains, then closed form solutions are obtained.

Consider a mixture distribution of the form

$$p_i(d) = \alpha_{ij} p_{ij}(d) \qquad (d \in D_{ij};\ j = 1, \ldots, m_i;\ i = 1, \ldots, N), \tag{8}$$

where the D_{ij} are disjoint sets of positive integers, $\bigcup_{j=1}^{m_i} D_{ij}$ is the set of all positive integers, the $p_{ij}(d)$ are discrete probability functions on D_{ij} and $0 \le \alpha_{ij} \le 1$, $\sum_{j=1}^{m_i} \alpha_{ij} = 1$. Thus $p_i(d)$ sews together discrete distributions with disjoint domains to make one flexible distribution. It has parameters α_{ij} in addition to those making up the $p_{ij}(d)$. In this case (7) becomes

$$\sum_{j=1}^{m_i} \sum_{d \in D_{ij}} \Delta_i(d)[\log \alpha_{ij} + \log p_{ij}(d)] \qquad (i = 1, \ldots, N), \tag{9}$$

where $\Delta_i(d)$ is defined to be zero for $d > T$. Optimizing with respect to the α_{ij} subject to $\sum_{j=1}^{m_i} \alpha_{ij} = 1$ yields

$$\tilde{\alpha}_{ij} = \frac{\sum_{d \in D_{ij}} \Delta_i(d)}{\sum_{d=1}^{T} \Delta_i(d)} \qquad (j = 1, \ldots, m_i);$$

closed form solutions for the remaining parameters can now be obtained provided each $p_{ij}(d)$ also admits closed form maximum likelihood solution. An example follows.

Modified geometric distribution. This is a distribution on the positive integers with an arbitrary head and a geometric tail. It is given by

$$p_i(d) = \begin{cases} \alpha_{id} & (d = 1, \dots, m_i - 1), \\ \alpha_{im_i}(1 - p_i)^{d-m_i} p_i & (d = m_i, m_i + 1, \dots), \end{cases}$$

where $\sum_{j=1}^{m_i} \alpha_{ij} = 1$. Maximizing (9) subject to this constraint yields

$$\tilde{\alpha}_{ij} = \frac{\Delta_i(j)}{\sum_{d=1}^{T} \Delta_i(d)}, \qquad \tilde{p}_i = \frac{\sum_{d=m_i}^{T} \Delta_i(d)}{\sum_{d=m_i}^{T} (d - m_i + 1)\Delta_i(d)},$$

where $j = 1, \dots, m_i - 1$, $i = 1, \dots, N$ and $\tilde{\alpha}_{im_i} = 1 - \sum_{j=1}^{m_i - 1} \tilde{\alpha}_{ij}$. This particular distribution has proved useful in practice.

Flexible dwell-time distributions such as these add further versatility to the model. However more flexibility is necessary if the model is to cope with the data imperfections often met in practice, such as in the next section.

5. Data screening using censoring

For the rainfall applications that we have in mind, the quality of the measurements of certain observations can sometimes be suspect, particularly for low rates and short durations. Although such situations are hopefully infrequent there is, nevertheless, a need for a flexible screening procedure where poor quality measurements can be ignored by the analysis while retaining the fact that an observation has taken place. This latter information is important for recovering the hidden semi-Markov structure.

A suitable and flexible screening procedure can be constructed by censoring the observations so that a vector observation $\boldsymbol{X}_t$ is recorded if $\boldsymbol{a}^\top \boldsymbol{X}_t > c$ where $\boldsymbol{a}$, c are suitably chosen by the analyst, and $\boldsymbol{X}_t$ is recorded as a missing observation otherwise. Adopting this strategy the relevant likelihoods can be derived and the EM recursions developed as before. The resulting procedures have much the same form as those given in Section 3, but typically with missing values replaced by appropriate predictors among other minor modifications. Full details are given in Sansom and Thomson (2000).

It is noted that the incorporation of censoring generalizes the original model while retaining its computational advantages and, most importantly, provides a very useful and versatile practical tool for analysts to explore the structure of large and complex datasets such as the rainfall data studied here.

6. Application

Breakpoints are the times at which the ambient rain rate changes and the breakpoint data comprise the observed sequence of steady rates (zero in the case of no precipitation) and their durations. Data from a rain gauge at Wellington (41°17′S, 174°46′E) for the 5 year period from January 1988 to December 1992 were used. This dataset is displayed in Figure 2 where a histogram of the *drys* (durations when no precipitation occurred) is shown at the top and a scatterplot of the *wets* (durations and rates when precipitation occurred) is shown below. All rates and durations have been logarithmically transformed in an effort to make the data more Gaussian—this is consistent

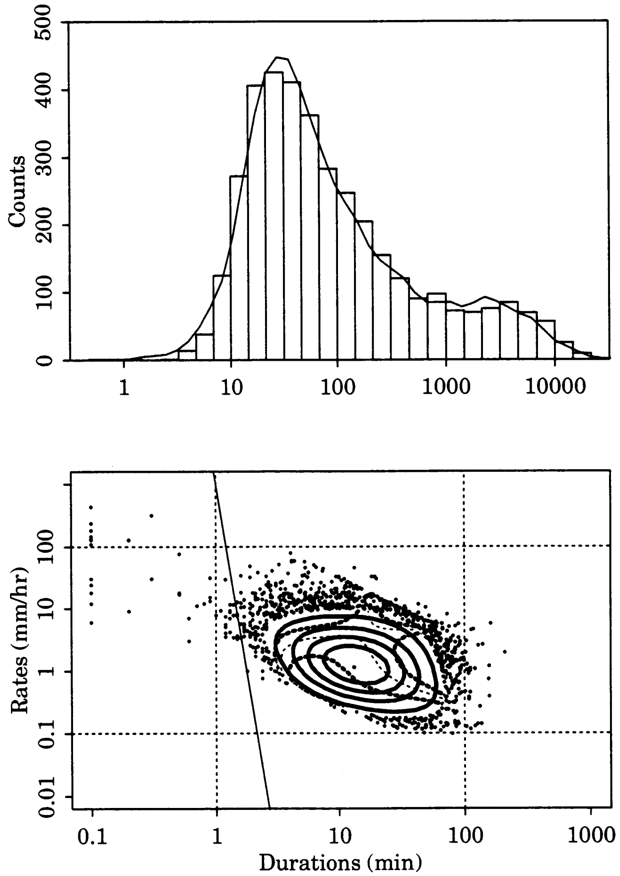

FIGURE 2: *Breakpoint rainfall data and its fitted HSMM.* The top panel shows a histogram of the 3713 dry durations and the bottom panel a scatterplot of the 9390 bivariate wet observations. Both the time and the rain-rate axes are logarithmic, and the central portion of the bivariate wet data has been contoured. The goodness of fit of the model to the dry data is shown by the curve superimposed on the histogram, while that of the wet datais shown by the dashed lines within the contoured portion of the scatterplot. The darker of these is the ±5% difference contour, and the lighter shows where there is no difference.

with the suggestions of many authors (see e.g. Biondini, 1976; Kedem *et al.*, 1994) that many types of rainfall measurements are log-normally distributed. It can be seen that the transformation did not completely normalize the data; the drys have a heavy upper tail and the contours in the wet scatterplot are distorted from being elliptical. It was assumed that the logarithms of the data could be well-represented by a mixture of normals or, equivalently, that the original data could be well-represented as a mixture of log-normals.

With these assumptions the HSMM model can be fitted to the data using the EM algorithm and the procedures described in the previous sections. However, although the EM algorithm can identify various states and their characteristics, the allocation of physically meaningful labels (rain, showers, etc.) to those states is the preserve of the experienced meteorologist. We have taken this subjective approach (the art of applied science) and ascribed specific physical causes to each state in the HSMM.

The fitting procedure is iterative so that initial values had to be chosen. Following Rabiner (1989), uniform probabilities are sufficient to initialize the transition probability matrix and, from Sansom (1999), if the dwell-time probabilities are non-parametric (as assumed initially) it is also sufficient to initialize these to be uniform. On the other hand (Rabiner, 1989), the parameters of the state distributions need to be initialized with values near to those that give the global maximum in the likelihood otherwise convergence may only reach a local maximum. To minimize this risk, initial values were randomly selected many times and the fitting procedure followed to convergence. Then the fit with the greatest likelihood was taken as the global maximum. Each random selection was restricted to feasible values with all the location parameters confined to the range of the data and the scale/correlation parameters kept to the same order as those of the data. While this method does not guarantee a global maximum, it is a sensible and practical strategy which, if many such fits are made, provides some degree of confidence that the global maximum has been reached.

One of the reasons for fitting an HSMM to the data, rather than an HMM, is that the dwell-time distributions are implicitly geometric for the HMM and it is far from certain that rainfall behaves that way. Indeed, because it is uncertain how the dwell times are distributed for rain states, a non-parametric dwell-time distribution over durations $1, \ldots, D$ was initially fitted. Here D must be large enough to capture the longest likely dwell periods, but short compared to the amount of data. For the dataset concerned, dwell periods of at most 30 breakpoints were expected and about 13 000 data were available, and so we chose $D = 50$. Once the non-parametric distribution has been fitted it can be used to suggest a more parsimonious parametric model.

A series of models, each with a different number of states is fitted in Sansom (1999) to a similar dataset from Invercargill, New Zealand. In general, the greater the number of states the smaller the BIC and hence the better the statistical fit. However, the physical cause for each state was also assessed, largely on the basis of the locations of the components, but the inter-relationships found within the transition probability matrix were also considered. Labels for the states were chosen: R, rain; S, showers; I, inter-event dry; E, error (so named since Sansom and Thomson (1992) showed that such a component arises through imperfections in the manual digitizing process); R1, R2, S1, S2, subdivisions of R or S. The same label was sometimes suitable for two states because they were co-located with similar links to other states and/or one represented only a small percentage of the data. In such a case, even if the model was statistically justifiable according to the BIC criterion, it was considered that the decomposition had progressed too far because it was unlikely that two physically distinct states had been found. A nine-state model with four drys and five wets was adopted by Sansom (1999) since all the more complicated models had at least two states with the same physical interpretation. This model was chosen as the starting point for the analysis of the Wellington data.

In the nine-state fit to the Wellington dataset, E was located such that it was covering more data than could be attributed to poor digitizing. Therefore one of the wet states was omitted and an eight-state model fitted to data in which the wet data to the left of the line in Figure 2 were censored. The BIC for the nine-state model was 69 888.6 and that of the eight-state was 69 095.4 which is much smaller, indicating

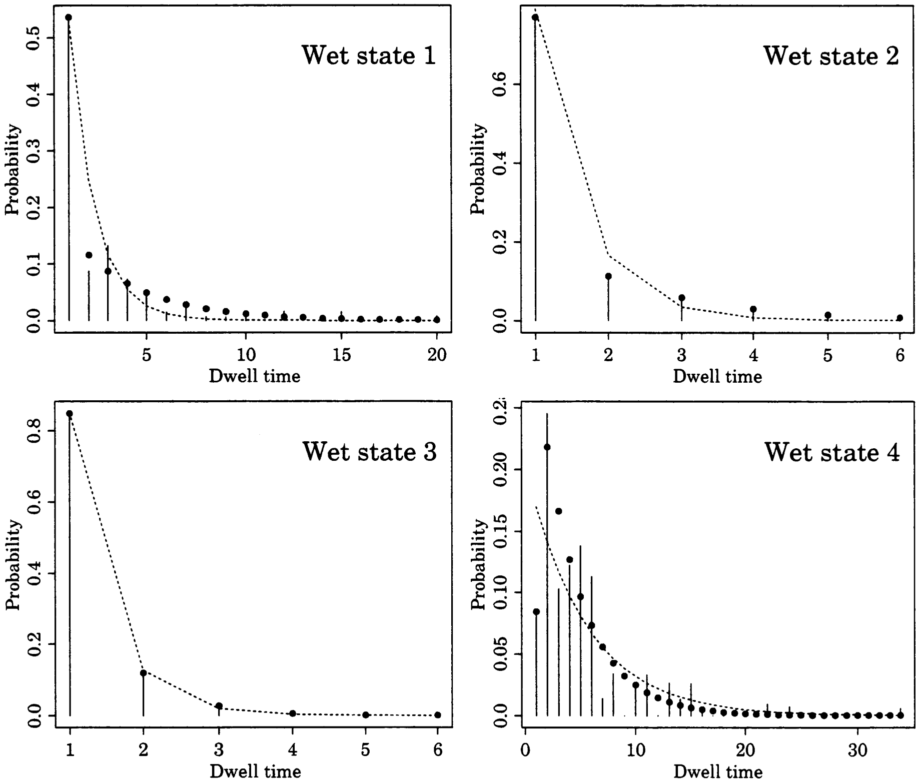

FIGURE 3: *Fitted dwell-time distributions.* Empirical or non-parametric distributions (solid lines), geometric distributions (dashed lines) and modified geometric distributions (points) superimposed.

a significantly better model. The necessity for an eight-state HSMM rather than an eight-state HMM can be assessed from Figure 3 which shows the fitted non-parametric dwell-time distribution with the fitted geometric superimposed. Clearly the geometric is a poor fit for three of the four states whereas the modified geometric (described in Section 4.1) with a free parameter for the probability of a dwell of 1 and a geometric tail for dwells greater than 1, fits the non-parametric probabilities adequately.

The Wellington data were refitted as an eight-state HSMM with modified geometric dwell times. The BIC improved significantly and, to assess estimation errors, fits were made to simulated data based on the parameters estimated by the fit to the actual data. Fifty simulated datasets and 50 sets of parameter estimates were created in this way using the same R as that estimated from the actual data using (6). The standard error of each parameter estimate from the actual data was then estimated by its standard deviation s over the 50 simulations. Any bias in the estimation procedure can also be checked by comparing the value of the parameter estimated from the actual dataset, θ say, to the mean of that parameter's estimates from the 50 simulations, $\overline{X}$. The statistic of interest is $(\overline{X}-\theta)/(s/\sqrt{50})$, shown in Table 1 for the state locations.

TABLE 1: Summary statistics for models fitted to the Wellington dataset

Model	I	II	III	III	
BIC	67402.9	67566.3	67520.0	67520.0	
	$(\overline{X}-\theta)/(s/\sqrt{50})$			θ	s
R1 Wet Duration	−1.55	−2.62	−1.34	3.14	0.018
R1 Wet Rate	−5.04	−1.14	0.42	0.43	0.019
R2 Wet Duration	−7.16	−0.05	−0.68	2.56	0.027
R2 Wet Rate	−2.57	−0.55	1.37	1.48	0.037
S1 Wet Duration	−14.89	0.20	1.34	2.92	0.018
S1 Wet Rate	0.75	0.05	−1.04	−0.47	0.019
S2 Wet Duration	−26.51	−0.88	1.65	2.13	0.014
S2 Wet Rate	13.85	0.18	−1.28	0.52	0.017
R Dry Duration	−1.16	−0.75	1.49	3.28	0.038
S1 Dry Duration	−5.49	−0.14	1.44	3.20	0.039
S2 Dry Duration	−1.70	−0.21	2.12	5.14	0.198
I Dry Duration	−3.90	−3.05	0.98	4.77	0.093
M Dry Duration	—	—	0.72	8.08	0.066

Notes: The BIC values for various models are given with bias diagnostics for some of the model's location parameters. Model I resulted when about 1.5% of the data were censored and Model II when the censoring was reduced to under 0.5%. In Model III the same amount of censoring was used, but instead of a single log-normal to represent the inter-event drys a mixture of two log-normals was used. The estimated parameter values (θ) and the standard error of these estimates (s) are given for Model III.

Table 1 indicates that some of the parameter estimates from the initial fit, in which about 1.5% of the data had been censored (Model I), had large biases with the largest being associated with the states which had been most heavily censored. When less severe censoring (under 0.5%) was applied (Model II) most of the biases decreased to acceptable levels so that, overall, a better model was found. However, a few parameters still remained biased including, in particular, the location of state I. This state had previously (Sansom and Thomson, 1992) been represented by two components and so a fifth dry state (called M since it may represent multiple occurrences of state I) was introduced. Using four dry states, the lower level of censoring, and with inter-event drys now represented as a mixture of I and M (Model III), the lowest BIC value was achieved and there were no significant biases in the estimates of any of the parameters.

The components of this HSMM fit to the data are shown in Figure 4 where the univariate and bivariate normal component distributions are shown appropriately normalized. The scaling factors for the latter are determined from the estimated transition probability matrix and the dwell-time distributions. A comparison between the fit and the data is given in Figure 2. For the drys, the overall fitted distribution is superimposed on the histogram and shows a good fit between the data and the model. For the wets, the percentage difference between the empirical distribution of the wets and the overall fitted distribution is shown by dashed lines. It can be seen in Figure 2 that the bulk of the wets is modelled well, and even around the extremes the absolute difference is not often more than 5%.

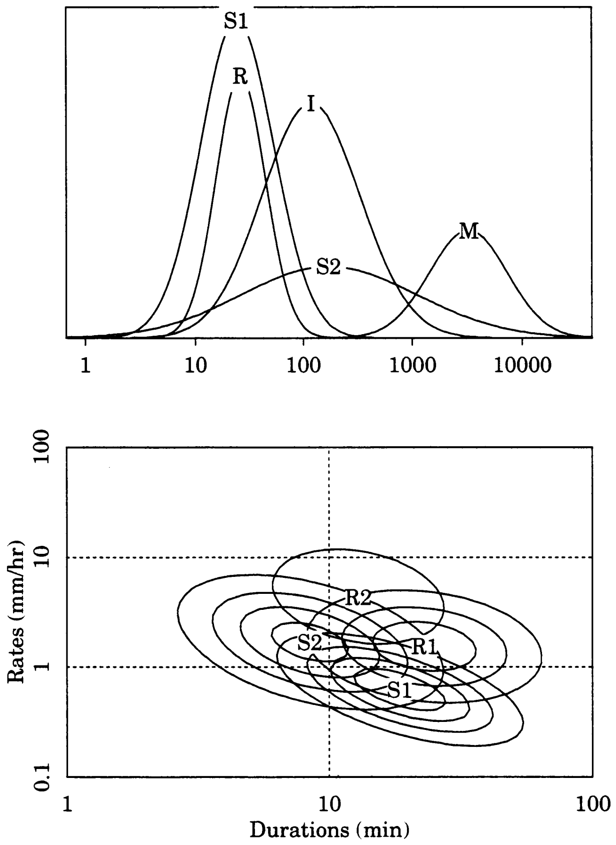

FIGURE 4: *Fitted state distributions, appropriately normalized.* The top graph shows the fitted univariate log-normal distributions for the dry durations and the bottom graph shows the fitted bivariate log-normal distributions for the wet breakpoint observations. The labels are the same as those used in Table 1, and from state to state the same valued contours are used for the bivariate distributions.

Figure 5 shows those transitions which collectively account for 94% of all the transitions that occurred in the data. The states are placed in the same relative positions as in Figure 4 except that the I and M states are shown together (i.e. the solid dot on the right) at a position which is the mean of their combined distribution. It can be seen from Figure 5 that eight transitions, in four pairs, are indicated as each contributing over 4% to the total number and collectively account for 70% of the total number. These four connections partitioned six of the eight states into two groups. The first group linked S1 wet with R and I, M (to which R1 wet and R2 wet are also weakly linked), and the second group linked S2 wet with S1 dry and S2 dry. The first group appears to represent the sequence of showers and rain associated with a frontal system, the second group periods of convective activity. The two groups are weakly linked from S2 wet to S1 wet, but are largely independent since otherwise they are separated by inter-event drys. It is also interesting to note that heavy and persistent rain (R2 wet) always occurs after an inter-event dry and decays to lighter and less persistent rain (R1 wet).

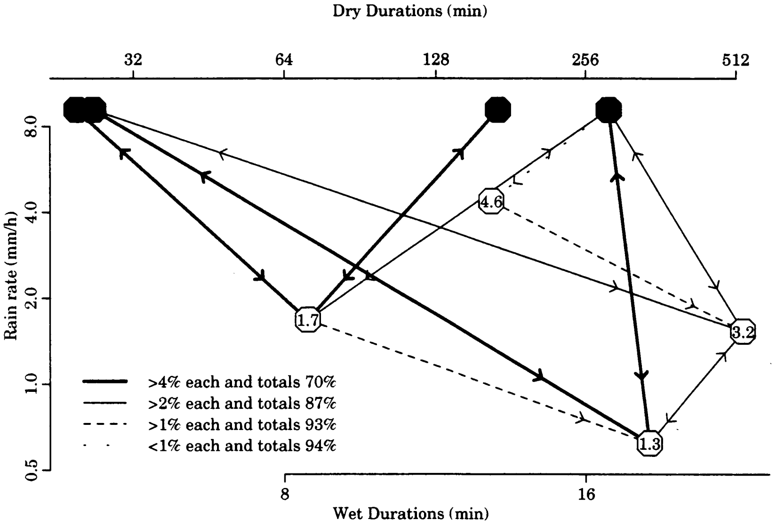

FIGURE 5: *Transitions between and dwells within the states of the HSMM.* The states are placed schematically with the dry ones as large solid dots along a time line indicated by the upper axis and the wet ones as large open dots in the time–rain-rate plane indicated by the lower axes. Transitions are shown by lines with attached arrows indicating the direction of transition; most, but not all, such lines have arrows pointing both ways. The solidity of the lines indicate the importance of the transitions, ranging from the heavy solid lines denoting transitions each covering at least 4% of the total number of transitions, to dotted lines for those covering less than 1%. At each wet state's location is shown the mean number of breakpoints that occurs before a shift to another state takes place; for the drys this is always just one before a transition to a wet state takes place.

7. Conclusions

The HSMM model provides a flexible, practical and computationally efficient tool for the exploration of large, high-resolution, bivariate, breakpoint rainfall datasets. The diversity of time scales inherent in the data are well-modelled by a hierarchy of unobserved states whose dynamics follow a semi-Markov process and which result in observations modelled by mixtures of log-normals. The iterative model fitting can be made relatively robust to initialization, even with moderately censored data. Moreover, censoring procedures can be used for screening poor quality data at little computational cost.

Most importantly, the fitted models provide a reasonable physical explanation of rainfall mechanisms. The next challenge is to extend the model to space as well as time, and to incorporate other variables such as seasonality.

Acknowledgements

The authors are pleased to take this opportunity to acknowledge their debt of gratitude to David Vere-Jones for his support over the past twenty years. Our research programmes, this and others, have benefitted enormously from his wise counsel, breadth of scholarship, unflagging enthusiasm and encouragement. Above all we acknowledge his contributions as mentor and friend.

References

Akaike, H. (1977). An entropy maximisation principle. In *Applications of Statistics*, ed. P. R. Krishnaiah, North Holland, Amsterdam, 27–41.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37**, 1554–1563.

Biondini, R. (1976). Cloud motion and rainfall statistics. *J. Appl. Meteorology* **15**, 205–224.

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

Elliott, R. J., Aggoun, L. and Moore, J. B. (1995). *Hidden Markov Models.* Springer, New York.

Ferguson, J. D. (1980). Variable duration models for speech. In *Symposium on the Application of Hidden Markov Models to Text and Speech*, ed. J. D. Ferguson, Institute for Defense Analyses, Princeton, NJ, 143–179.

Guédon, Y. and Cocozza-Thivent, C. (1990). Explicit state occupancy modelling by hidden semi-Markov models: application of Derin's scheme. *Comput. Speech Language* **4**, 167–192.

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

Hughes, J. P., Guttorp, P. and Charles, S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *J. Roy. Statist. Soc. Ser. C* **48**, 15–30.

Kedem, B. H., Pavlopoulos, H., Guan, X. and Short, D. A. (1994). A probability distribution model for rain rate. *J. Appl. Meteorology* **33**, 1486–1493.

Levinson, S. E., Rabiner, L. R. and Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Systems Tech. J.* **62**, 1035–1074.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and other Models for Discrete-valued Time Series.* Chapman and Hall, London.

Meilijson, I. (1989). A fast improvement to the EM algorithm in its own terms. *J. Roy. Statist. Soc. Ser. B* **51**, 127–138.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–285.

Rogers, R. R. (1979). *A Short Course in Cloud Physics.* Pergamon, New York.

Sansom, J. (1995). Rainfall discrimination and spatial variation using breakpoint data. *J. Climate* **8**, 624–636.

Sansom, J. (1998). A hidden Markov model for rainfall using breakpoint data. *J. Climate* **11**, 42–53.

Sansom, J. (1999). Large-scale spatial variability of rainfall through hidden semi-Markov models of breakpoint data. *J. Geophysical Res.* **104 (D24)**, 31631–31643.

Sansom, J. and Thomson, P. J. (1992). Rainfall classification using breakpoint pluviograph data. *J. Climate* **5**, 755–764.

Sansom, J. and Thomson, P. J. (1998). Detecting components in censored and truncated meteorological data. *Environmetrics* **9**, 673–688.

Sansom, J. and Thomson, P. J. (2000). Fitting hidden semi-Markov models. *Tech. Rep.* 77, National Institute of Water and Atmospheric Research, New Zealand.

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